International Journal of
Molecular Sciences

Special Issue Reprint

Plant Genomics 2019 Volume I

Edited by
Frank M. You

www.mdpi.com/journal/ijms

MDPI

Plant Genomics 2019—Volume I

Plant Genomics 2019—Volume I

Editor

Frank M. You

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin

Editor
Frank M. You
Ottawa Research and
Developmental Centre
Agriculture and Agri-Food Canada
Ottawa
Canada

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *International Journal of Molecular Sciences* (ISSN 1422-0067) (available at: https://www.mdpi.com/journal/ijms/special_issues/plant_genomics_2019).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

Volume I
ISBN 978-3-0365-8088-3 (Hbk)
ISBN 978-3-0365-8089-0 (PDF)

Volume I-II
ISBN 978-3-0365-8086-9 (Hbk)
ISBN 978-3-0365-8087-6 (PDF)

Contents

About the Editor

Frank M. You

Dr. Frank M. You is a highly accomplished senior research scientist in bioinformatics and genomics at the Agriculture and Agri-Food Canada (AAFC) Ottawa Research and Development Centre. He is also an Adjunct Professor in the Department of Plant Science at the University of Manitoba and a Guest Professor at Nanjing Agricultural University in China. Dr. You received his Ph.D. in plant genetics and breeding, with a specialization in statistical genetics, in 1989, and holds two Bachelor's degrees, one in agronomy from 1982 and another in computer science from 1999.

Dr. You is an expert in computational biology and bioinformatics, statistical genetics, and plant genetics and breeding. He has a wealth of experience in plant comparative and statistical genomics, quantitative genetics, genome assembly and annotation of complex genomes, gene expression and microarray data analysis, physical mapping and data analysis, high-throughput molecular marker design and development, and bioinformatics software development.

Dr. You's recent research projects focus on genome sequencing and annotation, QTL mapping, and the identification and characterization of genes associated with seed yield and disease resistance in flax and cereal crops. With his impressive expertise and extensive research accomplishments, Dr. You continues to contribute to the advancement of the field of plant genomics and genetics.

Preface to "Plant Genomics 2019—Volume I"

In recent years, researchers have uncovered the genes and genomic regions responsible for plants' growth, development, and stress responses. This reprint the 'Plant Genomics 2019' Special Issue comprises 57 papers exploring various aspects of plant genomics. These papers delve into gene discovery, genomic prediction, genome editing, plant chloroplast genome sequencing and comparative analysis, microRNA analysis, and comparative genomics.

The studies featured in this Special Issue employ a comprehensive research approach that combines bioinformatics and transcriptome analyses. With this approach, researchers have identified the genes associated with biotic and abiotic stress responses. Studies on the genome-wide identification of gene families, gene characteristics and distributions analysis, and gene expression profiles have shed light on various traits across multiple species.

In addition to gene discovery, the Special Issue also considers microRNAs (miRNAs) and their regulatory roles in gene expression. The roles of miRNAs in plant species have been explored, including the development of an artificial miRNA precursor system for gene silencing and the identification of miRNAs involved in seed development. These findings contribute to our understanding of how miRNA function in plant growth and development, offering potential avenues for crop improvement.

The Special Issue places a significant emphasis on the application of genomic tools for crop improvement. Papers on Chinese winter wheat and flax highlight the effectiveness of genomic prediction and marker-assisted selection in enhancing their yield, salt tolerance, and fruit ripening. Furthermore, the issue explores the revolutionary impact of the CRISPR/Cas9-mediated genome editing of plants, including targeted mutagenesis and gene replacement.

Comparative genomics is another key theme within this Special Issue, providing insights into the evolution of plant species. By comparing genomes, researchers can identify conserved gene families and regulatory elements, shedding light on plants' development and their adaptation to environmental stresses. Comparative genomics can also be used for phylogenetic analyses and resolving relationships between species.

Overall, the 57 papers featured in 'Plant Genomics 2019' exemplify the substantial progress made in understanding plant genetics and genomics. These studies offer valuable insights into the application of genomic tools for crop improvement, sustainable agriculture, and fundamental questions about the evolution and function of plant genes and genomes.

Frank M. You
Editor

International Journal of
Molecular Sciences

Article

Genomic Prediction for Grain Yield and Yield-Related Traits in Chinese Winter Wheat

Mohsin Ali [1], Yong Zhang [1], Awais Rasheed [2,3], Jiankang Wang [1] and Luyan Zhang [1,*]

[1] National Key Facility for Crop Gene Resources and Genetic Improvement, and Institute of Crop Sciences, Chinese Academy of Agricultural Sciences (CAAS), Beijing 100081, China; mali1990@yahoo.com (M.A.); zhangyong05@caas.cn (Y.Z.); wangjiankang@caas.cn (J.W.)

[2] International Maize and Wheat Improvement Center (CIMMYT) China Office, c/o CAAS, 12 Zhongguancun South Street, Beijing 100081, China; arasheed@qau.edu.pk

[3] Department of Plant Sciences, Quaid-i-Azam University, Islamabad 45320, Pakistan

[*] Correspondence: zhangluyan@caas.cn; Tel.: +86-10-82108572

Received: 11 January 2020; Accepted: 14 February 2020; Published: 17 February 2020

Abstract: Genomic selection (GS) is a strategy to predict the genetic merits of individuals using genome-wide markers. However, GS prediction accuracy is affected by many factors, including missing rate and minor allele frequency (MAF) of genotypic data, GS models, trait features, etc. In this study, we used one wheat population to investigate prediction accuracies of various GS models on yield and yield-related traits from various quality control (QC) scenarios, missing genotype imputation, and genome-wide association studies (GWAS)-derived markers. Missing rate and MAF of single nucleotide polymorphism (SNP) markers were two major factors in QC. Five missing rate levels (0%, 20%, 40%, 60%, and 80%) and three MAF levels (0%, 5%, and 10%) were considered and the five-fold cross validation was used to estimate the prediction accuracy. The results indicated that a moderate missing rate level (20% to 40%) and MAF (5%) threshold provided better prediction accuracy. Under this QC scenario, prediction accuracies were further calculated for imputed and GWAS-derived markers. It was observed that the accuracies of the six traits were related to their heritability and genetic architecture, as well as the GS prediction model. Moore–Penrose generalized inverse (GenInv), ridge regression (RidgeReg), and random forest (RForest) resulted in higher prediction accuracies than other GS models across traits. Imputation of missing genotypic data had marginal effect on prediction accuracy, while GWAS-derived markers improved the prediction accuracy in most cases. These results demonstrate that QC on missing rate and MAF had positive impact on the predictability of GS models. We failed to identify one single combination of QC scenarios that could outperform the others for all traits and GS models. However, the balance between marker number and marker quality is important for the deployment of GS in wheat breeding. GWAS is able to select markers which are mostly related to traits, and therefore can be used to improve the prediction accuracy of GS.

Keywords: wheat; genomic selection; missing data; minor allele frequency

1. Introduction

Wheat (*Triticum aestivum* L.) is one of the major cultivated crops that is growing on approximately 200 million hectares worldwide and delivers one fifth of the total caloric demands of the global population [1]. The increasing population and climate fluctuations impose new breeding challenges and require wheat breeders to use more efficient selection methods to develop high-yield cultivars with multiple resistances and wide adaptations [2]. Improvement of grain yield is still a considerable challenge to wheat breeding and production. Hence, modern wheat breeding approaches, such as the combination of accurate or suitable experimental designs, multiyear and multilocation trials, the application of concepts of quantitative and population genetics, and the integration of various disciplines such as computer science, statistics, and mathematics have been utilized widely in the last decade [3].

Recent advancements in high-throughput sequencing platforms have generated genome-wide dense molecular markers for genetic analysis in wheat [4]. Genomic selection (GS) is a special type of marker-assisted selection that incorporates genome-wide dense markers, as proposed by Meuwissen et al. [5]. GS could be a powerful tool in crop breeding to improve the prediction and selection accuracy for quantitative traits [6]. GS utilizes one or more training populations (TP) that have been genotyped and phenotyped to calibrate or train a statistical model. Then, the trained model is used to predict genomic estimated breeding values (GEBVs) in a validating population (VP), which is only genotyped. Superior parents for the next breeding cycle are selected based on the GEBV and consequently reduce the generation interval. Generally, the number of markers used for training the statistical model is far larger than the number of observations. Whole-genome regression methods based on ordinary least squares cannot estimate all marker effects simultaneously due to insufficient degrees of freedom. To address this issue, various classical statistical, Bayesian, and machine learning methods have been proposed for predicting the genetic merits of individuals [7]. These methods differ from each other mainly by a range of assumptions in the estimation of breeding values and variances in quantitative traits and computational complexity [2,7]. Among the parametric models, ridge regression (RidgeReg), ridge regression best linear unbiased predictions (RRBLUP), and genomic-BLUP (GBLUP) assume the normal distribution of marker effects with equal variance [7]. Least absolute shrinkage and selection operator (LASSO), Bayes A, and weighted Bayesian shrinkage regression or nonlinear regression assume the prior distribution of marker effects with a high probability and moderate to large effects, while Bayes B and Bayes Cπ assume some marker effects to be zero [7]. Nonparametric or semiparametric models, such as random forest (RForest), reproducing kernel Hilbert space (RKHS), and neural network approaches, have also been applied in GS [6,8,9]. Nonparametric models, such as RForest and RKHS, are capable of capturing non-additive effects and complex and nonexplicit interactions [2,9]. Previous efforts to compare the predictive ability of various GS models in wheat showed the good performances of RF and RKHS for traits of interest, but no single GS model outperformed the other models in all cases [9,10].

The efficiency of GS is always expressed by prediction accuracy, i.e., the correlation coefficient between observed phenotypic values and predicted GEBVs in VP. Previous studies have indicated that many factors are interrelated in a comprehensive manner [7,11], such as the genetic architecture of traits [11,12], heritability, population structure [13], type of statistical models, i.e., parametric and nonparametric models [9], cross-validation strategies [12], training population size and composition [12,13], marker density [6,13], and linkage disequilibrium (LD) between markers and QTL. Recent studies in animal and plant breeding have demonstrated that quality control (QC) on markers can improve the prediction accuracy of GS [14,15]. However, studies on the effect of missing rate and minor allele frequency (MAF) QC on the prediction accuracy of yield and yield-related traits are limited in wheat.

GS holds potential for the genetic improvement of qualitative and quantitative traits and has been widely used in wheat breeding to predict various traits, such as grain yield [12], test weight, heading time [10,12], disease resistance [16], end-use quality [17], iron and zinc contents [18], and physiological traits [19]. In addition, some studies have described the practical applications of GS in wheat breeding, such as cultivar development [20], cross prediction [21], and heterosis [22]. In this study, a wheat training population was developed from 166 elite wheat cultivars collected mainly from China. More than 80% of the cultivars (144) were collected from the Yellow and Huai River valley of China, which is one of the most important agricultural regions of wheat production in China and has an area of approximately 15 million hectares [23]. The main objectives of this study were (1) to evaluate the performance of seven GS models in predicting yield and yield-related traits in this wheat population, (2) to assess the effects of missing rate and MAF QC on the prediction accuracy of GS models, (3) to evaluate the effect of genotype imputation and genome-wide association studies (GWAS)-derived markers on prediction accuracy of GS.

2. Results

2.1. Phenotypic Evaluation

The descriptive statistics of grain yield (GY) and yield-related traits, i.e., spike number per square meter (SN), thousand-kernel weight (TKW), spike length (SL), heading days (HD), and plant height (PH) of the 166 wheat accessions in different environments (locations in cropping seasons) are shown in Table S1. The average values in each environment ranged from 6320.25 to 9318.19 kg per hectare (kg·ha^{-1}) for GY, 534 to 693 for SN, 39.38 to 49.82 g for TKW, 8.83 to 9.64 cm for SL, 184 to 199 days for HD, and 77.60 to 91.52 cm for PH (Table S1). Overall, the averages of BLUE values for GY, SN, TKW, SL, HD, and PH across all environments were 7268.81 kg·ha^{-1}, 605, 43.17 g, 9.15 cm, 187 days, and 83.36 cm, respectively (Table S1). High heritability was observed for all traits in all environments and ranged from 0.70 (for SL) to 1 (for HD). The difference in heritability among traits reflected the contribution of the environment to variations across locations and years (Table S1).

The Pearson's correlation coefficient (*r*) between traits ranged from −0.45 to 0.39 (Figure 1). Under the significance level of 0.001, GY had the highest positive correlation with TKW (*r* = 0.39) and lowest negative correlation with PH (*r* = −0.45, Figure 1). SN was negatively correlated with TKW (*r* = −0.38) but positively correlated with PH (*r* = 0.27). TKW was negatively correlated with HD (*r* = −0.25).

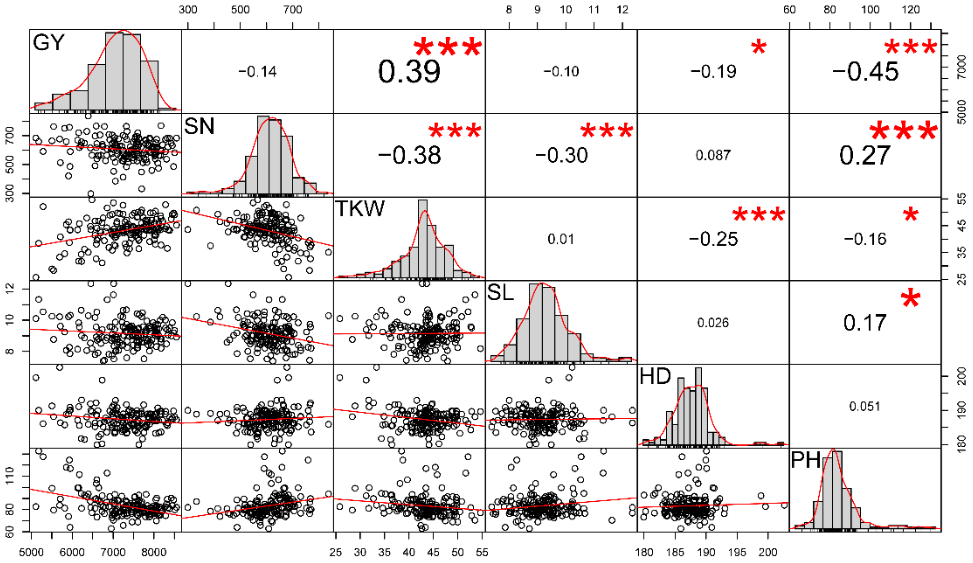


Figure 1. Pearson's correlation matrix among yield and yield-related traits based on their best linear unbiased estimates (BLUE). The upper corner represents the correlation coefficient, with the significance level indicated by asterisks. Three symbols ("*" and "***") correspond to three *p*-values (0.05 and 0.001, respectively). The lower corner contains bivariate scatter plots with fitted lines. The diagonally arranged plots show the phenotypic distribution of traits based on BLUE values. GY, indicates grain yield; SN, spike number per square meter; TKW, thousand-kernel weight; SL, spike length; HD, heading days; PH, plant height.

From ANOVA across environments, the genotype, block, environment, and genotype-by-environment interaction effects were all significant at a level of 0.001. For TKW, SL, and PH, the variance of environment and genotype-by-environment interaction was lower than the genotypic variance (Table 1). Environmental variance was the highest for HD, and the variance of the genotype-by-environment interaction was the largest for GY. Plot-level heritability was high for PH (0.85), HD (0.81), and TKW (0.77) but was relatively low for GY (0.42) (Table 1).

Table 1. Variance components and heritability of yield and yield-related traits in 166 wheat accessions.

Trait [1]	Variance Components (%)				Heritability [3]	
	Genotype	Environment	G by E Interaction [2]	Random Error	Plot Level	Genotypic Mean Level
GY	12.12	43.00	39.39	5.50	0.42	0.85
SN	34.32	24.11	36.19	5.39	0.69	0.92
TKW	41.38	27.68	23.76	7.18	0.77	0.97
SL	42.94	8.24	34.71	14.12	0.67	0.96
HD	12.64	79.29	7.26	0.81	0.81	0.97
PH	60.19	11.97	23.09	4.75	0.85	0.98

[1] GY, grain yield; SN, spike number per square meter; TKW, thousand-kernel weight; SL, spike length; HD, heading days; PH, plant height. [2] G by E; genotype-by-environment. [3] Heritability was estimated from analysis of variance across environments.

2.2. Marker Coverage, Genetic Diversity, and Linkage Disequilibrium Analysis

A total of 11,997 SNPs from 90 K of genotypic data were chosen to create two QC scenarios that were used for genomic prediction. In the first scenario, five subsets of markers were generated by removing markers with missing rate values above or equal to different thresholds (0%, 20%, 40%, 60%, and 80%). In the second scenario, three subsets of markers were generated by removing markers with MAF levels under or equal to different thresholds (0%, 5%, and 10%) for each missing level. The number of SNPs decreased significantly after the application of missing rate and MAF QC (Table 2). The distribution of these markers on the 21 wheat chromosomes is shown in Table S2. Markers were unevenly distributed along chromosomes. Generally, for the MAF level of 0%, the B genome had more markers than the A genome, and the A genome had more markers than the D genome; for the other MAF levels, the A genome had the most markers, and the D genome had the least markers (Table S2). The estimated polymorphic information content (PIC) values ranged from 0.005 to 0.702 across wheat accessions, with an average value of 0.13, whereas the genetic diversity (GD) ranged from 0.006 to 0.749, with a mean value of 0.149 (Figure S1).

Table 2. Number of markers used for genomic predictions under five missing rate levels (i.e., 0%, <20%, <40%, <60%, and <80%) and three minor allele frequency (MAF) levels (i.e., >0%, >5%, and >10%).

Missing Rate (%)	MAF (%)		
	0 [1]	5	10
0 [2]	1442	259	181
20	8674	5343	4368
40	9851	5513	4494
60	10818	5635	4596
80	11997	5725	4675

[1] MAF greater than zero. In other words, this QC scenario actually only removed non-polymorphism markers in the population, and therefore the remaining markers were polymorphic after this control. [2] Markers contained no missing values.

In total, 9851 SNPs with missing rate levels <40% were used to evaluate LD decay across the whole genome. The average r^2 was 0.065 in the whole genome, and the average LD decay distances for 10, 100, and 10,000 Mb were estimated to be 0.38, 0.28, and 0.14, respectively. The scatter plot between r^2 and physical distance (Mb) showed that LD decreased with increasing physical distance (Figure S2).

2.3. Prediction Accuracy of Different GS Models under Different Missing Rate and MAF Levels

Seven GS models were evaluated in this study. The prediction accuracies of the GS models ranged from 0.026 (PH) to 0.682 (TKW) and varied significantly among the six traits, five missing rate levels, and three MAF levels (Figure 2). Overall, QC for the missing rate and MAF improved the prediction accuracy, irrespective of traits and GS models (Figure 2). The prediction accuracy under a missing rate level of 0% was always the lowest for all traits as compared with other missing rate levels, and the prediction accuracy under the MAF level of 10% was the lowest for most traits as compared with other MAF levels. Steep slope was observed between prediction accuracies for missing rate levels of 0% and 20%. The major reason could be the significant difference of marker number between the two levels (Table 2). Stringent QC resulted in insufficient genome coverage and poor accuracy as well. However, prediction accuracy did not steadily increase with missing rate level (0% to 80%) and MAF level (0% to 10%). To find the best missing rate level, prediction accuracy under each missing rate was averaged across the seven GS models and three MAF levels for each trait. Considering the top three accuracies for each trait, missing rate levels 0%, 20% 40%, 60%, and 80% achieved the top accuracy for 0, 5, 5, 4, and 4 times, respectively, which indicated that, 20% to 40% was a suitable level for missing rate QC. Similarly, to find the best MAF level, prediction accuracy under each MAF level was averaged across the seven GS models and five missing rates for each trait (results not shown here). Considering the top two accuracies for each trait, MAF level 0%, 5%, and 10% achieved the top accuracy for 4, 5, and 3 times, respectively, which indicated that, 5% was a suitable level for MAF QC. In conclusion, missing rate levels of 20% to 40% and MAF level of 5% led to a suitable marker number and good or comparable prediction accuracies for all traits.

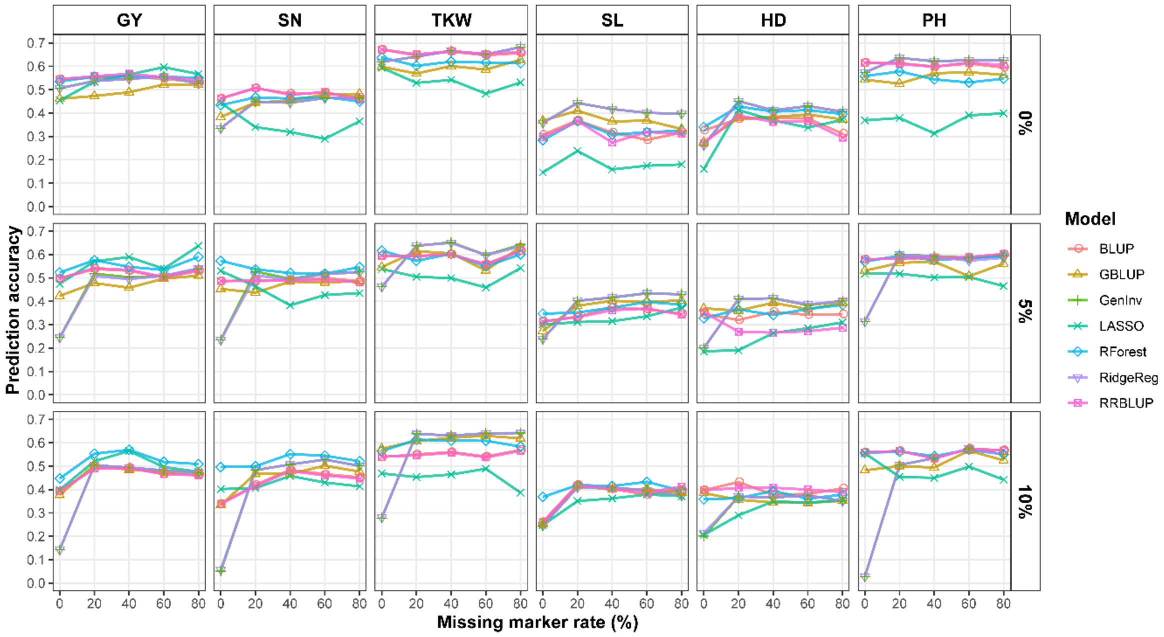


Figure 2. Genomic prediction accuracy of seven genomic selection (GS) models for yield and yield-related traits with five missing rates (columns) and three minor allele frequencies (MAFs) (rows). 0% MAF, represents markers with MAF greater than 0; GY, indicates grain yield; SN, spike number per square meter; TKW, thousand-kernel weight; SL, spike length; HD, heading days; PH, plant height.

Prediction accuracy for the six traits and seven GS models with non-QC or QC to keep the missing rate levels <40% and MAF values >5% are shown in Table 3. The prediction accuracies for SN, SL, HD, and PH with QC were consistently higher than those with non-QC for all GS models, except LASSO for SN, RRBLUP for HD, and BLUP and RRBLUP for PH. The improvement was not significant for GY and TKW, except that LASSO for GY (Table 3). In this QC scenario, independent of the GS models, moderate prediction accuracies were observed for all traits (Table 3). The average prediction accuracy of GY, SN, TKW, SL, HD, and PH across all GS models was 0.522, 0.480, 0.601, 0.380, 0.350, and 0.572, which was partially related to trait heritability. For example, TKW and PH had a high heritability and high prediction accuracy, and SN and SL had a low heritability and low accuracy.

However, GY had a low heritability but moderately high accuracy and HD had a high heritability but low accuracy. The possible reason could be the different genetic architectures of traits. For GY, variance of environment and genotype-by-environment interaction were similar, which were higher than genotypic variance, but for HD, environmental variance was much higher than the other two variances. The best models for GY, SN, TKW, SL, HD, and PH were LASSO, RForest, Moore–Penrose generalized inverse (GenInv), RidgeReg, GenInv (and also RidgeReg), and RidgeReg, respectively (Table 3). The average prediction accuracy of the seven GS models across the six traits was 0.489, 0.486, 0.512, 0.425, 0.494, 0.510, and 0.473, respectively. GenInv had the highest accuracy, followed by RidgeReg, RForest, BLUP, GBLUP, RRBLUP, and LASSO. Considering the top three accuracies for each trait, BLUP, GBLUP, GenInv, LASSO, RForest, RidgeReg, and RRBLUP achieved the top accuracy for 1, 2, 4, 1, 3, 5, and 2 times, respectively, which indicated that GenInv, RForest, and RidgeReg were relatively better among the seven models. In summary, independent of the GS models, moderate prediction accuracies were observed for all traits, and GenInv, RidgeReg, and RForest had a better performance than the other GS models for most traits (Table 3).

Table 3. Prediction accuracy with marker quality control (QC) or non-QC for six traits and seven genomic selection (GS) models. The QC to keep markers with missing rate levels <40% and minor allele frequency (MAF) >5% was used as an example.

Trait [1]	Scenario	Genomic Selection Model [2]							
		BLUP	**GBLUP**	**GenInv**	**LASSO**	**RForest**	**RidgeReg**	**RRBLUP**	**Mean**
GY	QC	0.531	0.458	0.503	**0.589** [5]	**0.547**	0.495	**0.534**	0.522
		(0.027) [3]	(0.033)	(0.029)	(0.022)	(0.028)	(0.030)	(0.027)	(0.016)
	Non-QC [4]	0.545	0.461	0.506	0.454	0.535	0.506	0.545	0.507
		(0.030)	(0.035)	(0.033)	(0.029)	(0.028)	(0.033)	(0.030)	(0.014)
	p-value	0.3687	0.3892	0.3926	0.0001	0.3688	0.4662	0.3994	0.248
SN	QC	0.491	0.484	**0.496**	0.383	**0.521**	**0.494**	0.488	0.480
		(0.026)	(0.025)	(0.025)	(0.030)	(0.027)	(0.025)	(0.026)	(0.017)
	Non-QC	0.462	0.383	0.335	0.444	0.434	0.335	0.463	0.408
		(0.030)	(0.034)	(0.035)	(0.033)	(0.031)	(0.035)	(0.031)	(0.021)
	p-value	0.2447	0.0096	0.0002	0.0816	0.018	0.0002	0.2671	0.011
TKW	QC	0.600	**0.605**	**0.652**	0.499	0.603	**0.650**	0.601	0.601
		(0.028)	(0.038)	(0.027)	(0.036)	(0.028)	(0.028)	(0.028)	(0.019)
	Non-QC	0.672	0.598	0.619	0.594	0.638	0.619	0.672	0.630
		(0.025)	(0.032)	(0.026)	(0.027)	(0.027)	(0.026)	(0.025)	(0.012)
	p-value	0.0282	0.4463	0.19	0.0189	0.1809	0.2122	0.0298	0.116
SL	QC	0.373	**0.402**	**0.416**	0.315	0.373	**0.417**	0.362	0.380
		(0.046)	(0.041)	(0.041)	(0.041)	(0.040)	(0.040)	(0.050)	(0.014)
	Non-QC	0.307	0.367	0.358	0.146	0.284	0.358	0.296	0.302
		(0.047)	(0.042)	(0.042)	(0.047)	(0.042)	(0.042)	(0.048)	(0.029)
	p-value	0.1595	0.2805	0.1629	0.0041	0.0645	0.1575	0.1721	0.020
HD	QC	0.355	0.394	**0.413**	0.264	**0.342**	**0.413**	0.266	0.350
		(0.038)	(0.034)	(0.033)	(0.033)	(0.033)	(0.033)	(0.040)	(0.024)
	Non-QC	0.326	0.276	0.262	0.161	0.340	0.262	0.272	0.271
		(0.036)	(0.036)	(0.040)	(0.047)	(0.027)	(0.040)	(0.033)	(0.022)
	p-value	0.2912	0.0265	0.0065	0.0412	0.471	0.0063	0.4991	0.017
PH	QC	**0.585**	0.570	0.593	0.502	0.576	**0.591**	**0.586**	0.572
		(0.019)	(0.020)	(0.019)	(0.032)	(0.027)	(0.020)	(0.019)	(0.012)
	Non-QC	0.616	0.543	0.574	0.369	0.558	0.574	0.615	0.550
		(0.022)	(0.030)	(0.023)	(0.039)	(0.027)	(0.040)	(0.033)	(0.032)
	p-value	0.1438	0.2283	0.2445	0.0057	0.3201	0.2807	0.1372	0.269
Mean	QC	0.489	0.486	0.512	0.425	0.494	0.510	0.473	
		(0.043)	(0.035)	(0.039)	(0.051)	(0.045)	(0.039)	(0.054)	
	Non-QC	0.488	0.438	0.442	0.461	0.465	0.442	0.477	
		(0.061)	(0.049)	(0.059)	(0.072)	(0.056)	(0.059)	(0.067)	

[1] GY, grain yield; SN, spike number per square meter; TKW, thousand-kernel weight; SL, spike length; HD, heading days; PH, plant height. [2] BLUP, best linear unbiased prediction; GBLUP, genomic-BLUP; GenInv, Moore–Penrose generalized inverse; LASSO, least absolute shrinkage and selection operator; RForest, random forest; RidgeReg, ridge regression; and RRBLUP, ridge regression-BLUP. [3] Values in parenthesis indicate standard errors of the estimated parameter. [4] Non-QC indicates that all polymorphic markers were used. [5] The models with the top three prediction accuracies under QC are bolded for each trait.

2.4. Effect of Imputation for Missing Genotypes on GS

The effect of genotype imputation on prediction accuracy was evaluated, using the QC to keep missing rate levels <40% and the MAF values >5% as an example. Prediction accuracies of yield and yield-related traits for the seven GS models using imputed markers are shown in Table 4. Compared with non-imputation (see Table 3 with QC), prediction accuracy using imputed markers sometimes increased slightly, but sometimes decreased slightly, irrespective of the traits and GS models (Table 4). Difference of prediction accuracy between imputation and non-imputation was minor. Averaged across the seven GS models, the prediction accuracy was 0.537, 0.496, 0.607, 0.423, 0.345, and 0.538 using imputed markers for GY, SN, TKW, SL, HD, and PH (last column in Table 4). Compared with results in Table 3, the accuracy improved by imputation for each trait was 0.015, 0.016, 0.006, 0.043, −0.005, and −0.034, respectively. Averaged across the six traits, the prediction accuracy using imputed markers was 0.472, 0.480, 0.521, 0.440, 0.509, 0.524, and 0.466 for the seven GS models, respectively (last row in Table 4). Compared with the results in Table 3, the accuracy improved by imputation for each model was −0.017, −0.006, 0.009, 0.015, 0.015, 0.014, and −0.007, respectively. The best models for the six traits were LASSO, RForest, GenInv (and also RidgeReg), RidgeReg, RidgeReg, and RForest, respectively (Table 4). Regarding the average performance across all traits, RidgeReg had the highest accuracy, followed by GenInv, RForest, GBLUP, BLUP, RRBLUP, and LASSO. Considering the top three accuracies for each trait, BLUP, GBLUP, GenInv, LASSO, RForest, RidgeReg, and RRBLUP achieved the top accuracy for 0, 2, 6, 1, 4, 6, and 0 times, respectively, which indicated that, GenInv, RidgeReg, and RForest were better among the seven models. In conclusion, imputation had marginal effect on GS, and it may not be a necessary step in the deployment of GS in wheat breeding. GenInv, RidgeReg, and RForest still had a better performance than the other GS models for most traits.

Table 4. Prediction accuracies of yield and yield-related traits for the seven GS models using imputed markers. The QC to keep markers with missing rate levels <40% and minor allele frequency (MAF) >5% was used as an example.

Trait [1]	Genomic selection Model [2]							
	BLUP	**GBLUP**	**GenInv**	**LASSO**	**RForest**	**RidgeReg**	**RRBLUP**	**Mean**
GY	0.517 (0.031) [3]	0.491 (0.031)	**0.531** [4] (0.024)	**0.593** (0.022)	**0.577** (0.024)	**0.531** (0.024)	0.520 (0.031)	0.537 (0.13)
SN	0.488 (0.025)	0.477 (0.036)	**0.520** (0.024)	0.418 (0.029)	**0.569** (0.026)	**0.518** (0.029)	0.481 (0.026)	0.496 (0.018)
TKW	0.600 (0.036)	0.560 (0.043)	**0.636** (0.024)	0.586 (0.031)	**0.629** (0.03)	**0.636** (0.031)	0.602 (0.035)	0.607 (0.011)
SL	0.370 (0.04)	**0.455** (0.041)	**0.489** (0.024)	0.370 (0.034)	0.394 (0.036)	**0.494** (0.036)	0.392 (0.044)	0.423 (0.021)
HD	0.305 (0.031)	**0.380** (0.030)	**0.381** (0.024)	0.377 (0.034)	0.312 (0.057)	**0.401** (0.029)	0.256 (0.035)	0.345 (0.02)
PH	0.549 (0.019)	0.517 (0.0241)	**0.566** (0.024)	0.450 (0.035)	**0.570** (0.025)	**0.566** (0.02)	0.547 (0.02)	0.538 (0.016)
Mean	0.472 (0.046)	0.480 (0.025)	0.521 (0.036)	0.440 (0.036)	0.509 (0.051)	0.524 (0.051)	0.466 (0.051)	

[1] GY, grain yield; SN, spike number per square meter; TKW, thousand-kernel weight; SL, spike length; HD, heading days; PH, plant height; [2] BLUP, best linear unbiased prediction; GBLUP, genomic-BLUP; GenInv, Moore–Penrose generalized inverse; LASSO, least absolute shrinkage and selection operator; RForest, random forest; RidgeReg, ridge regression; and RRBLUP, ridge regression-BLUP. [3] Values in parenthesis indicate standard errors of the estimated parameter. [4] The models with the top three prediction accuracies with and without markers imputation are bolded for each trait.

2.5. Effect of Significant Markers Detected by GWAS

Manhattan and quantile-quantile (Q-Q) plots from GWAS were given in Figure S3 for the imputed scenario, and in Figure S4 for the non-imputed scenario. The number of selected markers by GWAS is shown in Table 5. For both imputed and non-imputed scenarios, around 500 significant markers were detected by GWAS for each trait, which were, then, used for GS. For GWAS with imputed genotypic data, the number of selected markers was the highest for SN (537), and the lowest for PH (497). For GWAS with non-imputed genotypic data, the number of selected markers was also the highest for the SN (576), and the lowest for HD (506, Table 5). A comparison of Tables 3 and 4 shows that using GWAS-derived markers in GS increased the prediction accuracy in most cases, irrespective of the traits and GS models. Conducting imputation before GWAS made a small increase on the prediction accuracy for most traits and GS models (Table 6). Averaged across the seven GS models, the prediction accuracy for the six traits was 0.847, 0.850, 0.873, 0.843, 0.793, and 0.798 under the imputed scenario, and was 0.785, 0.833, 0.843, 0.785, 0.800, and 0.803 under the non-imputed scenario (last column in Table 6). Average across the six traits, the prediction accuracy for the seven models was 0.913, 0.836, 0.895, 0.673, 0.711, 0.895, and 0.913 under the imputed scenario, and was 0.886, 0.792, 0.842, 0.737, 0.679, 0.836, and 0.886 under the non-imputed scenario (last row in Table 6). These values were much higher than those from GS without marker selection by GWAS (Tables 3 and 4). The difference in accuracies between the imputed and non-imputed scenarios was minor. BLUP and RRBLUP had higher accuracy than the other models under both scenarios. Considering the top three accuracies for each trait, BLUP, GBLUP, GenInv, LASSO, RForest, RidgeReg, and RRBLUP achieved the top accuracy for 6, 1, 3, 0, 0, 4, and 6 times under the imputed scenario, and for 6, 0, 6, 0, 0, 2, and 6 times under the non-imputed scenario, respectively, which indicated that BLUP and RRBLUP were better among the seven models. The best models were different from those observed in Tables 3 and 4. The reason could be that BLUP and RRLUP were more suitable for datasets with a small number of markers. In conclusion, using GWAS to select markers is a useful step for GS. Effect of genotype imputation before GWAS was very small. BLUP and RRBLUP had a better performance than the other GS models for most traits when GWAS-selected markers were used for GS.

Table 5. The number of significant markers detected by genome-wide association studies (GWAS) under the imputed and non-imputed scenarios. Threshold of $-\log_{10} P$ was set at 1.

Trait [1]	GWAS QTLs	
	Imputed [2]	Non-imputed
GY	525	514
SN	537	576
TKW	519	553
SL	520	509
HD	508	506
PH	497	522
Total	3106	3080

[1] GY, grain yield; SN, spike number per square meter; TKW, thousand-grain weight; SL, spike length; HD, heading days; PH, plant height. [2] Markers were imputed before GWAS.

Table 6. Prediction accuracy for yield and yield-related traits using significant markers detected by genome-wide association studies (GWAS). Both imputed and non-imputed scenarios were considered. The QC to keep markers with missing rate levels <40% and minor allele frequency (MAF) >5% was used as an example.

Trait [1]	Imputation [2]	Genomic Selection Model [3]							
		BLUP	**GBLUP**	**GenInv**	**LASSO**	**RForest**	**RidgeReg**	**RRBLUP**	**Mean**
GY	Yes	**0.901** [4] (0.008) [5]	**0.901** (0.008)	0.887 (0.008)	0.721 (0.016)	0.730 (0.020)	0.887 (0.008)	**0.901** (0.008)	0.847 (0.031)
	No	**0.866** (0.01)	0.712 (0.038)	**0.788** (0.014)	0.746 (0.017)	0.727 (0.017)	**0.788** (0.014)	**0.866** (0.010)	0.785 (0.024)
SN	Yes	**0.923** (0.007)	0.859 (0.037)	**0.915** (0.007)	0.673 (0.025)	0.741 (0.018)	**0.915** (0.007)	**0.923** (0.005)	0.850 (0.039)
	No	**0.889** (0.011)	0.857 (0.018)	**0.872** (0.010)	0.767 (0.016)	0.685 (0.023)	**0.872** (0.010)	**0.889** (0.011)	0.833 (0.029)
TKW	Yes	**0.942** (0.006)	0.848 (0.034)	**0.938** (0.005)	0.737 (0.024)	0.763 (0.021)	0.938 (0.005)	**0.942** (0.007)	0.873 (0.034)
	No	**0.926** (0.005)	0.79 (0.036)	**0.881** (0.010)	0.763 (0.018)	0.733 (0.021)	0.880 (0.010)	**0.926** (0.005)	0.843 (0.03)
SL	Yes	**0.932** (0.005)	0.861 (0.034)	0.938 (0.005)	0.628 (0.026)	0.674 (0.029)	**0.938** (0.005)	**0.932** (0.005)	0.843 (0.051)
	No	**0.881** (0.012)	0.773 (0.044)	**0.841** (0.016)	0.669 (0.029)	0.613 (0.034)	0.840 (0.016)	**0.881** (0.012)	0.785 (0.04)
HD	Yes	**0.878** (0.011)	0.792 (0.030)	0.846 (0.012)	0.660 (0.026)	0.648 (0.024)	**0.846** (0.002)	**0.878** (0.011)	0.793 (0.037)
	No	**0.873** (0.010)	0.818 (0.021)	**0.827** (0.013)	0.728 (0.018)	0.653 (0.020)	0.826 (0.013)	**0.873** (0.010)	0.800 (0.031)
PH	Yes	**0.901** (0.008)	0.756 (0.038)	**0.846** (0.013)	0.621 (0.029)	0.712 (0.024)	**0.846** (0.013)	**0.901** (0.008)	0.798 (0.040)
	No	**0.880** (0.011)	0.800 (0.023)	**0.840** (0.017)	0.746 (0.023)	0.664 (0.020)	0.810 (0.017)	**0.880** (0.011)	0.803 (0.029)
Mean	Yes	0.913 (0.010)	0.836 (0.021)	0.895 (0.017)	0.673 (0.019)	0.711 (0.018)	0.895 (0.017)	0.913 (0.100)	
	No	**0.886** (0.009)	0.792 (0.020)	**0.842** (0.014)	0.737 (0.015)	0.679 (0.019)	0.836 (0.019)	**0.886** (0.009)	

[1] GY, grain yield; SN, spike number per square meter; TKW, thousand-kernel weight; SL, spike length; HD, heading days; PH, plant height. [2] "Yes" indicates that genotypic data was imputed for missing values and then used for GWAS and GS analysis. [3] BLUP, best linear unbiased prediction; GBLUP, genomic-BLUP; GenInv, Moore–Penrose generalized inverse; LASSO, least absolute shrinkage and selection operator; RForest, random forest; RidgeReg, ridge regression; and RRBLUP, ridge regression-BLUP. [4] The models with the top three prediction accuracies with and without markers imputation are bolded for each trait. [5] Values in parenthesis indicate standard errors of the estimated parameter.

3. Discussion

A better understanding of the factors that affect the prediction accuracy of GS is crucial to deploying GS within the conventional breeding scheme [12,19]. Missing rate and MAF are important factors that determine the quality of genotypic data. They have been extensively studied in animal breeding, but few such studies have been conducted in plant breeding [14,15,24]. According to the GS literatures, there is no consensus on marker QC thresholds for genomic prediction. Therefore, this study investigated the effect of missing rate and MAF QC on prediction accuracy for yield and yield-related traits in wheat. In addition, the effect of missing genotype imputation and GWAS-derived markers were also explored.

3.1. Marker Quality Control, Density, and LD

In most cases, QC on SNPs improved the prediction accuracy, irrespective of the traits and GS models. But the significance of improvement varied with trait features, GS models, as well as QC combinations. The QC levels for missing rate and MAF are important factors that affect prediction accuracy, in accordance with previous studies [15,24]. The increase in missing rate level resulted in an increased marker number, while the increase in MAF level resulted in a decreased marker number in the genome. The threshold of SNP QC can affect the quality of genotypic data, LD between markers and QTL, estimation of genetic relationship between individuals, and population structure [25]. Different genomic studies, including studies on QTL mapping, marker-assisted selection, and GWAS, have used different thresholds for marker QC [25]. In addition, different genotyping and sequencing platforms, such as SNP arrays and GBS (genotyping by sequencing), result in different qualities on genotypic data [26]. For instance, GBS is an effective genotyping technology that provides high marker density at a relatively low cost per sample, but it also generates a large proportion of missing data (up to 80%) when a low sequencing depth of genomic loci is employed [27]. However, markers with low MAF probably occur due to the design bias of the SNP array, because only a few cultivars and landraces are used to discover SNPs in the array [28].

Our results also revealed that QC for missing rate and MAF affected genome coverage (Table S2). The lowest number of markers was identified in the D genome, followed by the A and B genomes, which is in agreement with previous reports [23]. In addition, a stringent MAF threshold (e.g., >10% used in the present study) results in reduced allelic diversity in genomic datasets [29]. However, intrachromosomal LD decay declined rapidly with increasing distance (Figure S2). The LD decay rate is important because it determines the sufficient marker density for genome-wide coverage, i.e., at least one marker should be in LD with each segregating segment of the genome. In natural populations (non-inbred lines), faster LD decay requires higher marker density [30].

3.2. Effect of Missing Rate and MAF QC on Prediction Accuracy

Different missing rate thresholds have been adopted for QC on SNPs in previous studies [14,15,24,31]. However, it is difficult to determine which threshold is best for prediction accuracy. Therefore, QC with different missing rates and MAF thresholds was conducted to assess the predictability of the seven GS models in this study. Habier et al. [32] indicated that increasing marker density improves the genetic similarity of individuals in TP and VP, and thereby improves prediction accuracy. However, results from this study indicated that prediction accuracy was not improved consistently with an increase in missing rate level and MAF, irrespective of the traits (Figure 2). After QC for missing rate levels of 20% to 40%, all traits showed improved prediction accuracies, irrespective of the GS models (Figure 2). Including markers with a high missing rate level (e.g., 80%) added noise to the estimation of GEBVs. It is not necessary to use a small missing rate threshold (e.g., 0%), which undoubtedly reduces marker density and genome coverage, or even near to exclude some chromosomes (Table S2), reducing the prediction accuracy regardless of the GS models (Table 2 and Figure 2). The prediction accuracy was the highest when QC with a moderate missing data rate (20% to 40%) and MAF (5%) were used. Our conclusion was consistent with some previous studies. Roorkiwal et al. [33] evaluated prediction accuracy of six GS models for GS under nine combinations of missing rate and MAF QC for yield and yield-related traits in chickpea. Results indicated that QC on SNP before GS increased prediction accuracy, and missing rate levels ≤30% and MAF values ≥10% were the best QC combination. Jarquín et al. [15] compared different SNP QC (i.e., missing rate and MAF) scenarios in soybean and concluded that there was no unique strategy that outperformed the results of the others.

For most traits, prediction accuracies were negatively affected by a high MAF threshold (Figure 2). There were three possible reasons. First, the number of markers was smaller (insufficient genome coverage) and statistically less informative in the prediction analysis when a high MAF threshold was used (Table 2 and Table S2). Second, excluding markers by high MAF threshold could result in a bottleneck of allelic diversity [29] and biased accuracies for diverse germplasms. Third, excluded markers by QC can be linked with QTL, affecting some traits. In our diversity panel, it was possible that yield and yield-related traits were associated with relatively low-MAF SNPs which could have an advantage in the estimation of genetic relationship between TPs. The results from this study suggested that a moderate missing rate level (20% to 40%) and MAF (5%) threshold provided better prediction accuracy for yield and yield-related traits (Figure 2). There is no single combination of QC scenarios that outperforms the others for all traits and GS models. Further investigation is needed to determine how to find a balance between marker number and marker quality to achieve higher prediction accuracy. Further investigation is required to quantify the impact of other factors that were not included in this study such as TP size, population structure, and genotype-by-environment interactions, and imputation methods on prediction accuracy.

3.3. Effect of GS Models on Prediction Accuracy

The choice of GS models depends on the maximum prediction ability and computation efficiency of a model across a wide range of traits and datasets. In this study, the prediction accuracy varied substantially among traits and GS models (Table 3). There was no consistently best GS model for predicting various traits. This could be because the selected traits in the present study varied with respect to genetic architecture and heritability, whereas GS models differed from each other because of underlying assumptions for estimating marker effects [6,7]. In this study, GenInv, RidgeReg, and RForest had a higher prediction accuracy than other GS models (Table 3). The high accuracy of GenInv and RidgeReg could be explained by the overfitting of these two models, which was caused by the multi-collinearity between dense markers, overfitting results in biased estimation of marker effects, however, increasing the prediction accuracy in some cases. Another possible reason could be the trait features. Ornella et al. [34] reported the superiority of RidgeReg in some rust resistant traits because of the additive nature of rust resistance. The advantage of RForest is consistent with previous studies on wheat [18,19]. For example, Charmet et al. [10] compared the performances of five GS models on three elite breeding populations (each with approximately 350 lines) for three years and identified RForest to be the best model for predicting GY. Heslot et al. [7] evaluated seven GS models using eight datasets in wheat, barley, and maize, and demonstrated RF as a promising method to increase prediction accuracy. The superiority of RF for yield and yield-related was also reported in chickpea [33]. The superiority of RForest could be explained by its appealing properties for genomic predictions. RForest includes minor-effect QTLs and interacting and correlated markers with no distributional assumptions in the training model [8,35]. However, these three models may not show superiority in some other traits or populations. Further investigation is required to select the best GS model. In most cases, high-heritability traits result in high prediction accuracy, whereas low-heritability traits result in low prediction accuracy, regardless of the GS model [36]. Nevertheless, prediction accuracy relies not only on heritability but also on the genetic architecture of the target trait [37]. This phenomenon was also observed in this study.

3.4. Effect of Imputation and GWAS on Prediction Accuracy

The QC to keep missing rate levels <40% and MAF values >5% was used as an example to investigate the effect of imputation and GWAS on GS, and which QC scenario prediction accuracy was the highest. Missing values were imputed based on the empirical distribution of genotypes obtained from observed values, because this method requires less computational burden. Poland et al. [27] reported that imputation of missing values resulted in slightly higher prediction accuracy for yield (under drought condition), TKW, and HD, regardless of imputation methods (e.g., random forest, heterozygote, and expected maximization) in a panel of 254 advanced breeding lines from CIMMYT. Jarquín et al. [15] reported that imputation increased marker number and could improve prediction accuracy. However, in this study, it was concluded that imputation had little effect on the prediction accuracy of GS (Table 4), possibly because imputation and GS were conducted on the same dataset, and no additional information was provided to increase the prediction accuracy.

GWAS-derived markers improved prediction accuracy of GS, whether the imputation was conducted before GWAS or not. This was similar to the results by Lozada et al. [12], who reported that GS for GY using GWAS-derived markers had higher prediction accuracy than that using all markers in soft and red winter wheat. The possible reasons for the improvement in prediction accuracy included: (1) the number of GWAS-derived markers was smaller than the total marker number, which ultimately reduced multicollinearity and complexity of models for estimation of GEBVs; (2) the selected markers were all correlated with the traits. A low threshold ($-\log_{10} P = 1$) was adopted in this study to identify more significant markers. The low threshold definitely increased the false positive of GWAS, but it is generally accepted that false positive does not have a significant effect on prediction accuracy. In this study, both TP and VP were used in GWAS, and GWAS and GS were conducted on the same population. In practice, the GS model constructed from TP would be used for other breeding populations only having genotypic data. More evidence is still needed before concluding that GWAS always improves the accuracy of GS.

4. Materials and Methods

4.1. Plant Materials, Field Trials, and Phenotypic Evaluation

The wheat population was comprised of 166 diverse accessions, namely, 143 accessions from the Yellow and Huai River Valley Facultative Wheat Zone of China and 23 varieties from five other countries (Argentina, Australia, Italy, Japan, and Turkey). The names and origins of these accessions are presented in Table S3. These accessions were grown in Anyang in Henan province for three cropping seasons (i.e., from 2013 to 2015), in Shangqiu for two seasons (i.e., 2013 and 2014), and in Shijiazhuang for one season (i.e., 2015). All field trials were conducted in a randomized complete block design. Each trial had three replications, and each plot had three rows that were 2 m in length and 0.2 m in width. The genotypic and phenotypic data used in this study can be downloaded from http://www.isbreeding.net/wheatGS/.

Six yield and yield-related traits, namely, GY, SN, TKW, SL, HD, and PH, were evaluated at each location. GY was measured as the weight of grain harvested kg·ha^{-1}. SNs were counted for each plot and converted to spike number per square meter. TKW was measured by weighting 1000 random kernels from each plot after harvest. SLs were measured from the base of the rachis to the top spikelet, excluding awns. HD was recorded on 50% emergence of the spike; PH was measured as the distance between the soil surface and top of spike, excluding awns after physiological maturity. These traits were considered to represent a wide range of heritability and genetic architecture.

4.2. DNA Extraction, Genotyping, and Quality Control

Five fresh leaves of each accession were sampled, and DNA extraction was carried out by the modified CTAB method [38]. The genotypic data for the wheat accessions were obtained using a high-density Illumina 90K iSelect SNP array [39] featuring 81,587 SNPs. SNP genotyping was conducted by Genome Studio. A total of 21,856 SNPs remained for each accession using the genotypic data conversion function of QTL IciMapping V4.2 (freely available from https://www.isbreeding.net/) [40]. For QC on SNPs, the BIN function of QTL IciMapping v4.2 was used to remove the redundant markers, resulting in 14,043 non-redundant SNPs. Average missing rate of these markers was 28.20%. After binning, SNPs with more than 80% missing data were removed, and a total of 11,997 SNPs were employed to evaluate the prediction accuracy of the GS models.

4.3. Phenotypic Data Analysis and Analysis of Variance (ANOVA)

Descriptive statistics of phenotypic data were performed with Microsoft Excel 2016. Best linear unbiased estimates (BLUE) for each line across multiyear trials, ANOVA, and phenotypic correlation analysis were conducted using QTL IciMapping V4.2 [40]. BLUE values were used for correlation analysis. The ANOVA model across three locations is shown in Equation (1).

$$y_{ijk} = \mu + R_{k/j} + G_i + E_j + GE_{ij} + \varepsilon_{ijk} \text{ and } \varepsilon_{ijk} \sim N\left(0, \sigma_\varepsilon^2\right) \quad (1)$$

where y_{ijk} is the phenotypic value, μ is the overall mean, G_i is the genotypic effect, E_j is the environmental effect, GE_{ij} is the genotype-by-environment interaction effect, $R_{k/j}$ is the kth replication effect in the

jth environment, and ε_{ijk} is the residual effect. From the theoretical expectation of mean square (MS), genetic variance (σ_G^2), interaction variance (σ_{GE}^2), and error variance (σ_ε^2) were calculated using Equation (2), where environment number e = 3, and replication number r = 3 in the present study.

$$\sigma_G^2 = \frac{1}{e \times r}(MS_G - MS_\varepsilon),\ \sigma_{GE}^2 = \frac{1}{r}(MS_{GE} - MS_\varepsilon),\ \text{and}\ \sigma_\varepsilon^2 = MS_\varepsilon \tag{2}$$

Heritability at the plot level and mean level was calculated using Equation (3).

$$H^2_{perplot} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{GE}^2 + \sigma_\varepsilon^2}\ \text{and}\ H^2_{per\ mean} = \frac{\sigma_G^2}{\sigma_G^2 + \frac{1}{e}\sigma_{GE}^2 + \frac{1}{er}\sigma_\varepsilon^2} \tag{3}$$

4.4. Genotypic Data Analysis

LD between markers measured as r^2 was calculated by the TASSEL software (freely available from https://tassel.bitbucket.io/) using the full matrix and sliding window options [41] and plotted against physical distance. Markers with a missing rate lower than 40% were used for LD analysis. The PIC and GD (also known as expected heterozygosity) at each locus were estimated using PowerMarker v.3.5 (freely available from https://brcwebportal.cos.ncsu.edu/powermarker/) [42]. Plot visualizations of these parameters were generated by the ggplot2 package in R (freely available from https://cran.r-project.org/web/packages/ggplot2/index.html) [43].

4.5. GS Models and Factors Affecting Prediction Accuracy

Seven GS models were implemented in the Intel FORTRAN program for estimating prediction accuracies (code written by L.Z. and J.W.), i.e., BLUP [44], GBLUP [45], RRBLUP [46], RidgeReg, GenInv, LASSO [47], and one machine learning method, i.e., RForest [8]. Five-fold cross-validation was employed and replicated 50 times to avoid biases in the estimation of prediction accuracy. The averaged prediction accuracy across the 50 replicates was calculated. To assess the impact of the missing rate and MAF on prediction accuracy, five missing rate thresholds (i.e., 0%, <20%, <40%, <60%, and <80%) and three MAF thresholds (i.e., >0%, >5%, and >10%) were considered. This produced 15 marker datasets (e.g., 5 missing marker levels × 3 MAF levels). Prediction accuracy was evaluated in terms of Pearson's correlation between the observed adjusted phenotypic values (i.e., BLUE) and predicted values (i.e., GEBVs). In order to investigate the effect of QC on GS, QC with missing rate levels <40% and MAF values >5% was used as an example, and compared with non-QC, in which all polymorphic markers were used. *T*-test was conducted to compare the significance of difference between QC and non-QC cases. The detailed descriptions of these seven models are described in the next subsection. Comparison of prediction accuracy among traits and GS models was also conducted under this QC scenario, i.e., missing rate levels <40% and MAF >5%.

4.5.1. BLUP Model

The BLUP mixed model is described as follows:

$$y = X\beta + Zu + \varepsilon \tag{4}$$

where y is a $(n \times 1)$ vector of phenotypic values; X is a $(n \times p)$ incidence matrix for fixed effects; β is a $(p \times 1)$ vector of fixed effects; Z is a $(n \times 1)$ incidence matrix for random effects; u is a vector of random effects; ε is a $(n \times 1)$ vector of independently random residual following distribution $N\left(0, I\sigma_u^2\right)$, with I being the identity matrix in the present case [44]. In addition, p is the number of fixed effects, n is the number of genotypes, and m is the number of markers.

4.5.2. GBLUP Model

The standard GBLUP model is described as follows:

$$y = 1_n\mu + Zu + \varepsilon \tag{5}$$

where y is the vector of phenotypic values, 1_n is the vector of n ones, μ is the overall mean, Z is the design matrix for random effects, u is the random effect with $u \sim N\left(0, G\sigma_u^2\right)$, G is a genomic relationship matrix between individuals estimated from genotypes, and ε is the vector of random residuals following distribution $N\left(0, I\sigma_\varepsilon^2\right)$ [45,46].

4.5.3. RRBLUP Model

The rrBLUP model is described as follows:

$$y = 1_n\mu + Zu + \varepsilon \tag{6}$$

where y is the vector of phenotypic values; 1_n is the vector of n ones; μ is the overall mean; Z is the design matrix for random effects; u is the random e1ffect with $u \sim N\left(0, K\sigma_u^2\right)$; K is the additive relationship matrix, which is the density matrix in the present study; and ε is the vector of random residuals following distribution $N\left(0, I\sigma_\varepsilon^2\right)$ [46].

4.5.4. RidgeReg Model

The model is fitted by:

$$y = X\beta + Zu + \varepsilon \tag{7}$$

where y is the vector of phenotypic values, X is the design matrix for fixed effects, β is the vector of marker fixed effects, Z is the design matrix for random effects, u is the vector of random effects, ε is the vector of random residual following distribution $N\left(0, I\sigma_u^2\right)$, and I is the identity matrix. The estimator of β is $(X'X + \lambda I)^{-1}X'y$, and the estimator of λ can be expressed as arg $\underset{\beta}{\min}\left(\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2\right)$, with the notation $\|.\|_2$ for L_2 norm. The "arg $\underset{\beta}{\min}$" notation expresses the determination of coefficient β minimizing the expression inside the brackets.

4.5.5. GenInv Model

The model is the same as the model described in Section 4.5.4, but the estimation of β is $(X'X)^{+}X'y$. Here, "+" is the Moore–Penrose generalized inverse of matrix.

4.5.6. LASSO Model

The LASSO was developed by Tibshirani [47]. λ refers to the shrinkage and regularization parameter calculated by arg $\underset{\beta}{\min}\left(\frac{\|y-X\beta\|_2^2}{2\sigma^2} + \lambda\|\beta\|_1\right)$ and $\|\beta\| = \sum_i \|\beta_i\|$ is the L_1 norm.

4.5.7. RForest Model

The RForest model is integrated with classification or regression trees that rely on bootstrap samples and splits original data into multiple subsets of non-overlapping sets [16]. The predictions for observations are calculated as the averages of predicted values over the trees.

4.6. Imputation for Missing Genotypes

To evaluate the effect of genotype imputation on prediction accuracy of GS, missing values in genotypic data were imputed using samples from the empirical distribution of marker genotypes [48].

Imputed markers were employed for genomic prediction analysis. The QC to keep missing rate levels <40% and MAF values >5% was used as an example, i.e., 5513 markers were retained after QC and employed for imputation and GS.

4.7. GWAS-Derived Genomic Selection

To evaluate the effect of GWAS-based marker selection on prediction accuracy of GS, the "GWAS" function of R package rrBLUP V4.6 (freely available from https://cran.r-project.org/web/packages/rrBLUP/index.html/) was used to perform GWAS before GS [46]. To avoid spurious marker-trait associations due to population structure, a realized additive relationship matrix and the first five principal components were included in the model (i.e., PCA + K model). The additive relationship matrix was computed using "A.mat" function of rrBLUP. Manhattan plots were also generated by rrBLUP [46]. The threshold of $-\log_{10} P$ was set at 1, in order not to miss any small-effect quantitative trait nucleotides. The QC to keep missing rate levels <40% and MAF values >5% was used as an example. Markers with unknown physical positions were excluded before GWAS. As a result, a total of 5201 markers were employed for GWAS. Two GWAS-derived GS scenarios were designed. In the first scenario, imputation for missing genotypes was conducted before GWAS; in the second scenario, non-imputed genotypic data was used to perform GWAS.

5. Conclusions

In this study, a diverse Chinese winter wheat panel was used to compare prediction accuracy of seven GS models (BLUP, GBLUP, GenInv, LASSO, RForest, RidgeReg, and RRBLUP) under different marker QC scenarios (five missing rate levels and three MAF values) for yield and yield-related traits. No single QC combination or GS model can yield better performance in prediction accuracy for all traits. In general, a combination of moderate missing rate levels (20% to 40%) and MAF (5%) yielded better prediction accuracy, regardless of the traits and GS models. Prediction accuracy of the six traits was affected by the heritability, genetic architecture, and GS models. GenInv, RidgeReg, and RForest models yielded higher prediction accuracy than other models across traits. The effect of genotype imputation and GWAS-based marker selection was also evaluated in this study. The results showed that imputation had marginal effect on GS but using GWAS-derived markers improved the prediction accuracy of GS.

Supplementary Materials: Supplementary Materials can be found at http://www.mdpi.com/1422-0067/21/4/1342/s1.

Author Contributions: J.W. and Y.Z. conceived and designed the research; M.A. and L.Z. conducted data analysis and wrote the paper; A.R. assisted in improving the content of the manuscript. All authors read the final version of the manuscript and approved the manuscript for publication.

Funding: This work was supported by the National Natural Science Foundation of China (project no. 31861143003) and the National Key Basic Research Program of China (project no. 2014CB138105).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ANOVA	Analysis of variance
BLUE	Best linear unbiased estimates
BLUP	Best linear unbiased predictors
GEBV	Genomic estimated breeding values
GenInv	Moore-Penrose generalized inverse
GBS	Genotyping by sequencing
GS	Genomic selection
GWAS	Genome-wide association studies
GY	Grain yield
HD	Heading days
LASSO	Least absolute shrinkage and selection operator
LD	Linkage disequilibrium
MAF	Minor allele frequency
PH	Plant height
QC	Quality control
RForest	Random forest
RidgeReg	Ridge regression
RRBLUP	Ridge regression best linear unbiased predictors
SL	Spike length
SN	Spike number
SNP	Single nucleotide polymorphism
TKW	Thousand-kernel weight
TP	Training population
VP	Validating population

References

1. FAO FAOSTAT. Available online: http://www.fao.org/faostat/en/#data/QC (accessed on 2 August 2017).
2. Voss-Fels, K.P.; Cooper, M.; Hayes, B.J. Accelerating crop genetic gains with genomic selection. *Theor. Appl. Genet.* **2019**, *132*, 669–686. [CrossRef] [PubMed]
3. Bassi, F.M.; Bentley, A.R.; Charmet, G.; Ortiz, R.; Crossa, J. Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). *Plant Sci.* **2016**, *242*, 23–36. [CrossRef] [PubMed]
4. Rasheed, A.; Hao, Y.; Xia, X.; Khan, A.; Xu, Y.; Varshney, R.K.; He, Z. Crop breeding chips and genotyping platforms: Progress, challenges, and perspectives. *Mol. Plant* **2017**, *10*, 1047–1064. [CrossRef] [PubMed]
5. Meuwissen, T.H.; Hayes, B.J.; Goddard, M.E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **2001**, *157*, 1819–1829.
6. Crossa, J.; Pérez-Rodríguez, P.; Cuevas, J.; Montesinos-López, O.; Jarquín, D.; de los Campos, G.; Burgueño, J.; González-Camacho, J.M.; Pérez-Elizalde, S.; Beyene, Y. Genomic selection in plant breeding: Methods, models, and perspectives. *Trends Plant Sci.* **2017**, *22*, 961–975. [CrossRef]
7. Heslot, N.; Yang, H.P.; Sorrells, M.E.; Jannink, J.L. Genomic selection in plant breeding: A comparison of models. *Crop Sci.* **2012**, *52*, 146–160. [CrossRef]
8. Breiman, L. Random Forests. *Machine Learn.* **2001**, *45*, 5–32. [CrossRef]
9. Pérez-Rodríguez, P.; Gianola, D.; González-Camacho, J.M.; Crossa, J.; Manès, Y.; Dreisigacker, S. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3 Genes Genomes Genet.* **2012**, *2*, 1595–1605. [CrossRef]
10. Charmet, G.; Storlie, E.; Oury, F.X.; Laurent, V.; Beghin, D.; Chevarin, L.; Lapierre, A.; Perretant, M.R.; Rolland, B.; Heumez, E. Genome-wide prediction of three important traits in bread wheat. *Mol. Breeding* **2014**, *34*, 1843–1852. [CrossRef]
11. Jannink, J.L.; Lorenz, A.J.; Iwata, H. Genomic selection in plant breeding: From theory to practice. *Brief. Funct. Genomics* **2010**, *9*, 166–177. [CrossRef]
12. Lozada, D.N.; Mason, R.E.; Sarinelli, J.M.; Brown-Guedira, G. Accuracy of genomic selection for grain yield and agronomic traits in soft red winter wheat. *BMC Genet.* **2019**, *20*, 82. [CrossRef] [PubMed]

13. Norman, A.; Taylor, J.; Edwards, J.; Kuchel, H. Optimising genomic selection in wheat: Effect of marker density, population size and population structure on prediction accuracy. *G3 Genes Genomes Genet.* **2018**, *8*, 2889–2899. [CrossRef] [PubMed]
14. Bresolin, T.; de Magalhães Rosa, G.J.; Valente, B.D.; Espigolan, R.; Gordo, D.G.M.; Braz, C.U.; Fernandes, G.A.; Magalhães, A.F.B.; Garcia, D.A.; Frezarim, G.B. Effect of quality control, density and allele frequency of markers on the accuracy of genomic prediction for complex traits in Nellore cattle. *Anim. Prod. Sci.* **2019**, *59*, 48–54. [CrossRef]
15. Jarquín, D.; Howard, R.; Graef, G.; Lorenz, A. Response surface analysis of genomic prediction accuracy values using quality control covariates in soybean. *Evol. Bioinfrom.* **2019**, *15*, 4–10. [CrossRef]
16. Juliana, P.; Singh, R.P.; Singh, P.K.; Crossa, J.; Huerta-Espino, J.; Lan, C.; Bhavani, S.; Rutkoski, J.E.; Poland, J.A.; Bergstrom, G.C. Genomic and pedigree-based prediction for leaf, stem, and stripe rust resistance in wheat. *Theor. Appl. Genet.* **2017**, *130*, 1415–1430. [CrossRef]
17. Hayes, B.; Panozzo, J.; Walker, C.; Choy, A.; Kant, S.; Wong, D.; Tibbits, J.; Daetwyler, H.; Rochfort, S.; Hayden, M. Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes. *Theor. Appl. Genet.* **2017**, *130*, 2505–2519. [CrossRef]
18. Velu, G.; Crossa, J.; Singh, R.P.; Hao, Y.; Dreisigacker, S.; Perez-Rodriguez, P.; Joshi, A.K.; Chatrath, R.; Gupta, V.; Balasubramaniam, A. Genomic prediction for grain zinc and iron concentrations in spring wheat. *Theor. Appl. Genet.* **2016**, *129*, 1595–1605. [CrossRef]
19. Norman, A.; Taylor, J.; Tanaka, E.; Telfer, P.; Edwards, J.; Martinant, J.P.; Kuchel, H. Increased genomic prediction accuracy in wheat breeding using a large Australian panel. *Theor. Appl. Genet.* **2017**, *130*, 2543–2555. [CrossRef]
20. Beyene, Y.; Semagn, K.; Mugo, S.; Tarekegne, A.; Babu, R.; Meisel, B.; Sehabiague, P.; Makumbi, D.; Magorokosho, C.; Oikeh, S. Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. *Crop Sci.* **2015**, *55*, 154–163. [CrossRef]
21. Yao, J.; Zhao, D.; Chen, X.; Zhang, Y.; Wang, J. Use of genomic selection and breeding simulation in cross prediction for improvement of yield and quality in wheat (*Triticum aestivum* L.). *Crop J.* **2018**, *6*, 353–365. [CrossRef]
22. Zhao, Y.; Li, Z.; Liu, G.; Jiang, Y.; Maurer, H.P.; Würschum, T.; Mock, H.P.; Matros, A.; Ebmeyer, E.; Schachschneider, R. Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 15624. [CrossRef] [PubMed]
23. Liu, J.; He, Z.; Rasheed, A.; Wen, W.; Yan, J.; Zhang, P.; Wan, Y.; Zhang, Y.; Xie, C.; Xia, X. Genome-wide association mapping of black point reaction in common wheat (*Triticum aestivum* L.). *BMC Plant Biol.* **2017**, *17*, 220. [CrossRef] [PubMed]
24. Edriss, V.; Guldbrandtsen, B.; Lund, M.S.; Su, G. Effect of marker-data editing on the accuracy of genomic prediction. *J. Anim. Breed. Genet.* **2013**, *130*, 128–135. [CrossRef] [PubMed]
25. Anderson, C.A.; Pettersson, F.H.; Clarke, G.M.; Cardon, L.R.; Morris, A.P.; Zondervan, K.T. Data quality control in genetic case-control association studies. *Nat. Protoco.* **2010**, *5*, 1564. [CrossRef]
26. Elbasyoni, I.S.; Lorenz, A.J.; Guttieri, M.; Frels, K.; Baenziger, P.S.; Poland, J.; Akhunov, E. A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. *Plant Sci.* **2018**, *270*, 123–130. [CrossRef]
27. Poland, J.; Endelman, J.; Dawson, J.; Rutkoski, J.; Wu, S.; Manes, Y.; Dreisigacker, S.; Crossa, J.; Sánchez-Villeda, H.; Sorrells, M. Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* **2012**, *5*, 103–113. [CrossRef]
28. Cavanagh, C.R.; Chao, S.; Wang, S.; Huang, B.E.; Stephen, S.; Kiani, S.; Forrest, K.; Saintenac, C.; Brown-Guedira, G.L.; Akhunova, A. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 8057–8062. [CrossRef]
29. Allen, A.M.; Winfield, M.O.; Burridge, A.J.; Downie, R.C.; Benbow, H.R.; Barker, G.L.; Wilkinson, P.A.; Coghill, J.; Waterfall, C.; Davassi, A. Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*). *Plant Biotechnol. J.* **2017**, *15*, 390–401. [CrossRef]

30. Liu, H.; Zhou, H.; Wu, Y.; Li, X.; Zhao, J.; Zuo, T.; Zhang, X.; Zhang, Y.; Liu, S.; Shen, Y. The impact of genetic relationship and linkage disequilibrium on genomic selection. *Plos ONE* **2015**, *10*, e0132379. [CrossRef]
31. Hickey, J.M.; Crossa, J.; Babu, R.; de los Campos, G. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. *Crop Sci.* **2012**, *52*, 654–663. [CrossRef]
32. Habier, D.; Tetens, J.; Seefried, F.-R.; Lichtner, P.; Thaller, G. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel. Evol.* **2010**, *42*, 5. [CrossRef]
33. Roorkiwal, M.; Rathore, A.; Das, R.R.; Singh, M.K.; Jain, A.; Srinivasan, S.; Gaur, P.M.; Chellapilla, B.; Tripathi, S.; Li, Y. Genome-enabled prediction models for yield related traits in chickpea. *Front. Plant Sci.* **2016**, *7*, 1666. [CrossRef] [PubMed]
34. Ornella, L.; Singh, S.; Perez, P.; Burgueno, J.; Singh, R.; Tapia, E.; Bhavani, S.; Dreisigacker, S.; Braun, H.J.; Mathews, K.; et al. Genomic prediction of genetic values for resistance to wheat rusts. *Plant Gen.* **2012**, *5*, 136–148. [CrossRef]
35. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R news* **2002**, *2*, 18–22.
36. Thavamanikumar, S.; Dolferus, R.; Thumma, B.R. Comparison of genomic selection models to predict flowering time and spike grain number in two hexaploid wheat doubled haploid populations. *G3 Genes Genomes Genet.* **2015**, *5*, 1991–1998. [CrossRef] [PubMed]
37. Daetwyler, H.D.; Pong-Wong, R.; Villanueva, B.; Woolliams, J.A. The impact of genetic architecture on genome-wide evaluation methods. *Genetics* **2010**, *185*, 1021–1031. [CrossRef]
38. Murray, M.; Thompson, W.F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **1980**, *8*, 4321–4326. [CrossRef]
39. Wang, S.; Wong, D.; Forrest, K.; Allen, A.; Chao, S.; Huang, B.E.; Maccaferri, M.; Salvi, S.; Milner, S.G.; Cattivelli, L. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol. J.* **2014**, *12*, 787–796. [CrossRef]
40. Meng, L.; Li, H.; Zhang, L.; Wang, J. QTL IciMapping: Integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* **2015**, *3*, 269–283. [CrossRef]
41. Bradbury, P.J.; Zhang, Z.; Kroon, D.E.; Casstevens, T.M.; Ramdoss, Y.; Buckler, E.S. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **2007**, *23*, 2633–2635. [CrossRef]
42. Liu, K.; Muse, S.V. PowerMarker: An integrated analysis environment for genetic marker analysis. *Bioinformatics* **2005**, *21*, 2128–2129. [CrossRef] [PubMed]
43. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: Berlin, Germany, 2016; pp. 33–88.
44. Bernardo, R. *Breeding for Quantitative Traits in Plants*; Stemma Press: Minneapolis, MN, USA, 2002; pp. 259–299.
45. VanRaden, P.M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **2008**, *91*, 4414–4423. [CrossRef] [PubMed]
46. Endelman, J.B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* **2011**, *4*, 250–255. [CrossRef]
47. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **1996**, *58*, 267–288. [CrossRef]
48. Pérez-Rodríguez, P.; Crossa, J.; Rutkoski, J.; Poland, J.; Singh, R.P.; Legarra, A.; Autrique, E.; Campos, G.d.l.; Burgueño, J.; Dreisigacker, S. Single-step genomic and pedigree genotype × environment interaction models for predicting wheat lines in international environments. *Plant Genome* **2017**, *10*, 1–15. [CrossRef]

International Journal of
Molecular Sciences

MDPI

Article

Genomic Prediction Accuracy of Seven Breeding Selection Traits Improved by QTL Identification in Flax

Samuel Lan [1,2], Chunfang Zheng [1], Kyle Hauck [1,2], Madison McCausland [1,3], Scott D. Duguid [4], Helen M. Booker [5], Sylvie Cloutier [1,*] and Frank M. You [1,*]

[1] Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, ON K1A 0C6, Canada; slanftw@gmail.com (S.L.); chunfang.zheng@canada.ca (C.Z.); kyle.hauck@canada.ca (K.H.); madison.mccausland@canada.ca (M.M.)

[2] Department of Mathematics and Statistics, University of Waterloo, Waterloo, ON N2L 3G1, Canada

[3] Department of Plant Sciences, University of Manitoba, Winnipeg, MB R3T 2N2, Canada

[4] Morden Research and Development Centre, Agriculture and Agri-Food Canada, Morden, MB R6M 1Y5, Canada; scott.duguid@canada.ca

[5] Crop Development Centre, University of Saskatchewan, Saskatoon, SK S7N 5A8, Canada; helen.booker@usaska.ca

* Correspondence: Frank.You@canada.ca (F.M.Y.); Sylvie.Cloutier@canada.ca (S.C);
Tel.: +1-613-759-1539 (F.M.Y.); +1-613-759-1744 (S.C.)

Received: 28 January 2020; Accepted: 23 February 2020; Published: 25 February 2020

Abstract: Molecular markers are one of the major factors affecting genomic prediction accuracy and the cost of genomic selection (GS). Previous studies have indicated that the use of quantitative trait loci (QTL) as markers in GS significantly increases prediction accuracy compared with genome-wide random single nucleotide polymorphism (SNP) markers. To optimize the selection of QTL markers in GS, a set of 260 lines from bi-parental populations with 17,277 genome-wide SNPs were used to evaluate the prediction accuracy for seed yield (YLD), days to maturity (DTM), iodine value (IOD), protein (PRO), oil (OIL), linoleic acid (LIO), and linolenic acid (LIN) contents. These seven traits were phenotyped over four years at two locations. Identification of quantitative trait nucleotides (QTNs) for the seven traits was performed using three types of statistical models for genome-wide association study: two SNP-based single-locus (SS), seven SNP-based multi-locus (SM), and one haplotype-block-based multi-locus (BM) models. The identified QTNs were then grouped into QTL based on haplotype blocks. For all seven traits, 133, 355, and 1208 unique QTL were identified by SS, SM, and BM, respectively. A total of 1420 unique QTL were obtained by SS+SM+BM, ranging from 254 (OIL, LIO) to 361 (YLD) for individual traits, whereas a total of 427 unique QTL were achieved by SS+SM, ranging from 56 (YLD) to 128 (LIO). SS models alone did not identify sufficient QTL for GS. The highest prediction accuracies were obtained using single-trait QTL identified by SS+SM+BM for OIL (0.929 ± 0.016), PRO (0.893 ± 0.023), YLD (0.892 ± 0.030), and DTM (0.730 ± 0.062), and by SS+SM for LIN (0.837 ± 0.053), LIO (0.835 ± 0.049), and IOD (0.835 ± 0.041). In terms of the number of QTL markers and prediction accuracy, SS+SM outperformed other models or combinations thereof. The use of all SNPs or QTL of all seven traits significantly reduced the prediction accuracy of traits. The results further validated that QTL outperformed high-density genome-wide random markers, and demonstrated that the combined use of single and multi-locus models can effectively identify a comprehensive set of QTL that improve prediction accuracy, but further studies on detection and removal of redundant or false-positive QTL to maximize prediction accuracy and minimize the number of QTL markers in GS are warranted.

Keywords: flax; genome-wide association study (GWAS); single nucleotide polymorphism (SNP); genomic selection; prediction accuracy; quantitative trait loci (QTL); quantitative trait nucleotides (QTNs)

1. Introduction

Genomic selection (GS) is a form of marker-assisted selection (MAS) that predicts genomic estimated breeding values (GEBVs) of test individuals through the use of genome-wide markers [1,2]. GS has been implemented in crop breeding to increase selection accuracy, reduce breeding cost, and speed-up genetic progress [3,4]. In a practical GS scheme, many factors affect its accuracy: training populations, statistical models, molecular markers, relatedness of the training populations and selection (test) populations, and so on [1,3]. Markers are one of the critical factors. In the initial concept of GS, high-density genome-wide random markers were used in genomic modeling [2]. With advances in next generation sequencing technologies and genotyping methods such as genotyping-by-sequencing (GBS) and single nucleotide polymorphism (SNP) arrays, a sufficiently large set of high-density genome-wide markers for a genetic panel can be easily generated at a low cost. However, the cost associated with obtaining such a large number of markers in the test lines can be excessive considering their generally large number. In fact, only a few markers may be associated with the traits of interest in a set of high-density genome-wide markers. This not only leads to the "large *p*, small *n*" problem [1], where a high number of marker effects need to be estimated using a population of very small sample size ($p >> n$), but also results in background noise in model construction because of uncorrelated markers, contrarily decreasing the genomic prediction accuracy of GS models [5]. Previous studies have confirmed that increasing marker density ensures the maintenance of association between markers and quantitative trait loci (QTL) to obtain a high prediction accuracy, but prediction accuracy plateaus when marker density increases to a certain threshold [5–7]. Using QTL associated with traits of interest, instead of using a full set of random SNPs in a GS model, greatly reduces the number of markers, which in turns reduces the cost of genotyping large breeding populations. Additionally, the exclusive use of markers associated with traits in GS models can increase prediction accuracy through reducing the background noise in the model construction [5,8]. Our previous study on pasmo resistance in flax has showed that using 500 QTL identified through single-locus and multi-locus genome-wide association study (GWAS) models [9] from a flax core collection (a germplasm population) [10,11] was highly effective for GS and generated a prediction accuracy as high as 0.92 compared with 0.67 when using 52,347 random SNPs [5].

The traditional GWAS methods, such as the general linear model (GLM) [12] and the mixed linear model (MLM) [13], are single-locus models that test the significance of marker–trait association one marker at a time and declare significant associations based on a stringent multiple-test correction (most often Bonferroni). Because of the high significance stringency, these methods only detect a few relatively large-effect quantitative trait nucleotides (QTNs) and, they lack the power to identify small-effect polygenes for more complex quantitative traits. Thus, alternative multi-locus methods have been proposed [14], including the multi-locus random-SNP-effect mixed linear model (mrMLM) [9,15], the FAST multi-locus random-SNP-effect EMMA (FASTmrEMMA) [16], the polygene-background-control-based least angle regression plus empirical Bayes (pLARmEB) [17], the iterative modified-sure independence screening EM-Bayesian LASSO (ISIS EM-BLASSO) [18], and the integration of the Kruskal–Wallis test with empirical Bayes under polygenic background control (pKWmEB). These methods adapt statistical models that simultaneously test multiple markers and, doing so, substantially increase the statistical power while simultaneously reducing Type 1 errors and running time [9,15–19]. These methods also usually adapt LOD scores (usually LOD ≥ 3), rather than the stringent Bonferroni correction (0.05/number of SNPs) [19], thus empowering the detection of more large and small effect QTNs [10]. In contrast to these multi-locus models, the fixed and random model circulating probability unification (FarmCPU) [20] still uses Bonferroni correction and mostly detects a few large-effect QTNs [10]. The above two types of GWAS models can be described as SNP-based single-locus (SS) and SNP-based multi-locus (SM) models. Another type of GWAS is haplotype-block-based (BM) GWAS models. Close SNPs are more likely to be inherited together; haplotype blocks are important in genetic studies [21], such as diversity studies [22], GWAS, and genomic selection [23–25]. The use of haplotypes in the genomic prediction of traits of allogamous plants can increase its predictive ability by 20% [23]. A restricted two-stage multi-locus multi-allele GWAS (RTM-GWAS) procedure [26] is one recently proposed BM [27–29]. This method first generates SNP LD blocks (SNPLDB) and then groups SNPs into an SNPLDB based on LD blocks. Each block as a marker may contain one or more SNPs that result in two or more haplotypes as its alleles for QTL mapping [26]. Thus, the significantly associated SNPLDB markers (blocks or singletons) are directly considered QTL. All these methods offer promise to identify an exhaustive set of QTNs/QTL for breeding selection.

The objectives of this study were to evaluate GS prediction accuracies for seven major breeding selection traits using QTL identified by different GWAS models of a genetic panel of 260 flax breeding lines derived from bi-parental populations. Ten statistical GWAS models belonging to the SS, SM, and BM classes were compared to first optimize QTL identification and second to maximize prediction accuracy.

2. Results

2.1. Phenotyping of the Population

Seven breeding selection traits in flax, namely, seed yield (YLD), days to maturity (DTM), iodine value (IOD), protein content (PRO), oil content (OIL), linoleic acid content (LIO) and linolenic acid content (LIN) were measured from 260 lines from bi-parental populations grown in the field for four years at two locations (Figure 1). Less variability was observed in 2009 at both locations across all traits because only 96 of the 260 lines were evaluated that year at the two locations. DTM, PRO, and YLD showed significant differences across four years and both locations, whereas the seed quality traits (IOD, LIN, LIO, and OIL) had relatively similar performance at the two locations. All traits, with the exception of PRO, had significantly higher values in Saskatoon than Morden ($p < 2 \times 10^{-16}$ for all six traits except for PRO). The analysis of variance also showed a significant interaction between years and locations for all traits except for LIO (p = 0.97; Table S1). The performance of the seven traits in different years and locations suggested that the phenotypic data of each environment (years and locations) should be used to identify all potential stable and environment-specific QTNs associated with the traits.

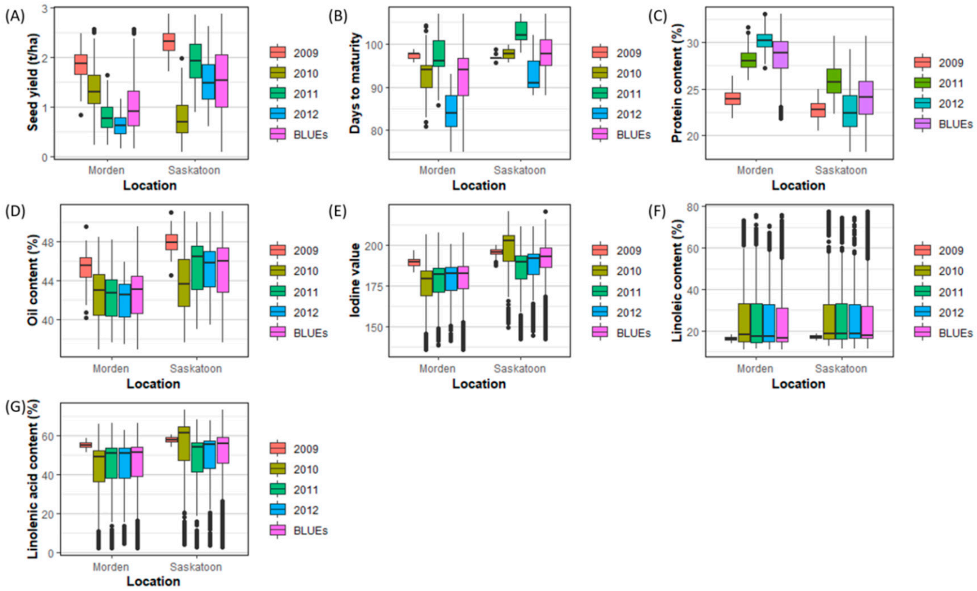


Figure 1. Boxplots of phenotypic data of the seven traits: seed yield (YLD) (**A**), days to maturity (DTM) (**B**), protein content (PRO) (**C**), oil content (OIL) (**D**), iodine value (IOD) (**E**), linoleic acid content (LIO) (**F**), and linolenic acid content (LIN) (**G**). BLUEs, best linear unbiased estimates across four years.

2.2. Haplotype Blocks

RTM-GWAS was used to identify haplotype blocks of 17,277 SNPs in the 260 lines [26]. A total of 2776 haplotype blocks with two or more SNPs per block and 2852 singletons were generated. Although a singleton has only one SNP, it can be treated as an independent block. As such, a total of 5628 haplotype blocks were considered for further QTL mapping and analyses. The number of blocks ranged from 231 in chromosome 11 (Lu11) to 500 in chromosome 1 (Lu1) with an average block size of 20.09–29.78 kb (Table 1).

Table 1. The haplotype blocks identified from 17,277 single nucleotide polimorphsims (SNPs) in the 260 lines and association with quantitative trait loci (QTL) of traits.

Chr	No of Blocks (Including Singletons)	No of Singletons	Average SNPs Per Block	Average Block Size (Kb)	No of Blocks with QTL
Lu1	500	257	3.02	27.61 ± 32.99	126
Lu2	374	178	4.10	28.07 ± 34.68	101
Lu3	472	242	2.81	23.96 ± 30.24	116
Lu4	337	182	2.45	23.31 ± 32.50	108
Lu5	308	133	3.48	29.78 ± 35.16	57
Lu6	419	227	2.80	26.11 ± 32.91	80
Lu7	296	157	2.86	29.15 ± 35.21	116
Lu8	433	244	2.52	20.05 ± 27.18	126
Lu9	443	208	3.19	24.89 ± 31.83	95
Lu10	389	210	2.89	25.79 ± 31.75	80
Lu11	231	127	2.60	26.50 ± 33.37	44
Lu12	355	149	3.90	26.70 ± 32.72	112
Lu13	448	216	3.51	29.50 ± 34.34	111
Lu14	381	208	2.82	23.04 ± 31.60	89
Lu15	242	114	3.07	27.81 ± 33.42	59
Total	5628	2852	3.07	26.12 ± 32.64	1420

2.3. QTNs/QTL

To compare the performance of different statistical models to identify QTNs in GWAS, three types of models were evaluated: (1) two SS models, including GLM [12] and MLM [13], (2) seven SM models, including the six models implemented in the mrMLM package and FarmCPU implemented in the MVP package, and (3) the BM model, RTM-GWAS [26].

A total of 268 and 407 unique QTNs for the seven traits were identified using SS and SM, totaling 608 unique QTNs, while 1208 significant haplotype blocks or singletons were detected using BM (RTM-GWAS) (Table 2, Tables S2 and S3). The QTNs from SS and SM were further grouped based on haplotype blocks; that is, the QTNs located in the same haplotype block were grouped into a QTN cluster or a QTL. As such, 608 QTNs for the seven traits identified using SS and SM were grouped into 427 unique QTN clusters or QTL for the seven traits. Since the results from RTM-GWAS were haplotype-block-based, they were directly treated as QTL. Therefore, 1420 unique QTL were identified for the seven traits when all models (SS+SM+BM) were considered, including 361, 351, 269, 254, 283, 254, and 256 QTL for YLD, DTM, PRO, OIL, LOD, LIO, and LIN, respectively (Table 2, Figure 2). For each QTL, a tag QTN was selected to represent the QTL.

Table 2. Quantitative trait nucleotides (QTNs)/quantitative trait loci (QTL) identified from 17,277 single nucleotide polymorphisms (SNPs) in the 260 lines for the seven traits using three types of genome-wide association study (GWAS) models.

Trait	QTNs		QTL								
	SS	SM	SS	SM	SS+SM	BM	All (SS+SM+BM)	Major QTL	Major QTL Effect (R^2, %)	Minor QTL Effect (R^2, %)	All QTL Effect (R^2, %)
YLD	13	58	8	53	56	323	361	110	11.03 ± 6.75	1.32 ± 1.24	4.64 ± 6.14
DTM	43	76	28	71	87	301	351	39	6.99 ± 2.11	1.12 ± 1.25	1.70 ± 2.22
PRO	66	56	31	51	74	220	269	77	16.55 ± 12.50	1.24 ± 1.25	5.48 ± 9.54
OIL	17	88	10	84	87	186	254	111	15.80 ± 10.26	1.43 ± 1.30	7.88 ± 9.96
IOD	153	82	71	72	123	190	283	55	9.47 ± 3.79	1.30 ± 1.40	2.96 ± 3.91
LIO	146	102	68	87	128	152	254	70	9.86 ± 3.98	1.40 ± 144	3.50 ± 4.34
LIN	189	127	70	67	118	170	256	53	10.21 ± 4.10	1.25 ± 1.37	3.06 ± 4.22
All	268	407	133	355	427	1,208	1,420	520	12.06 ± 8.24	1.28 ± 1.33	3.99 ± 6.34

SS, SNP-based single-locus models; SM, SNP-based multi-locus models; BM, haplotype-block-based multi-locus model. Major QTL are defined as $R^2 \geq 5\%$, while minor QTL as $R^2 < 5\%$.

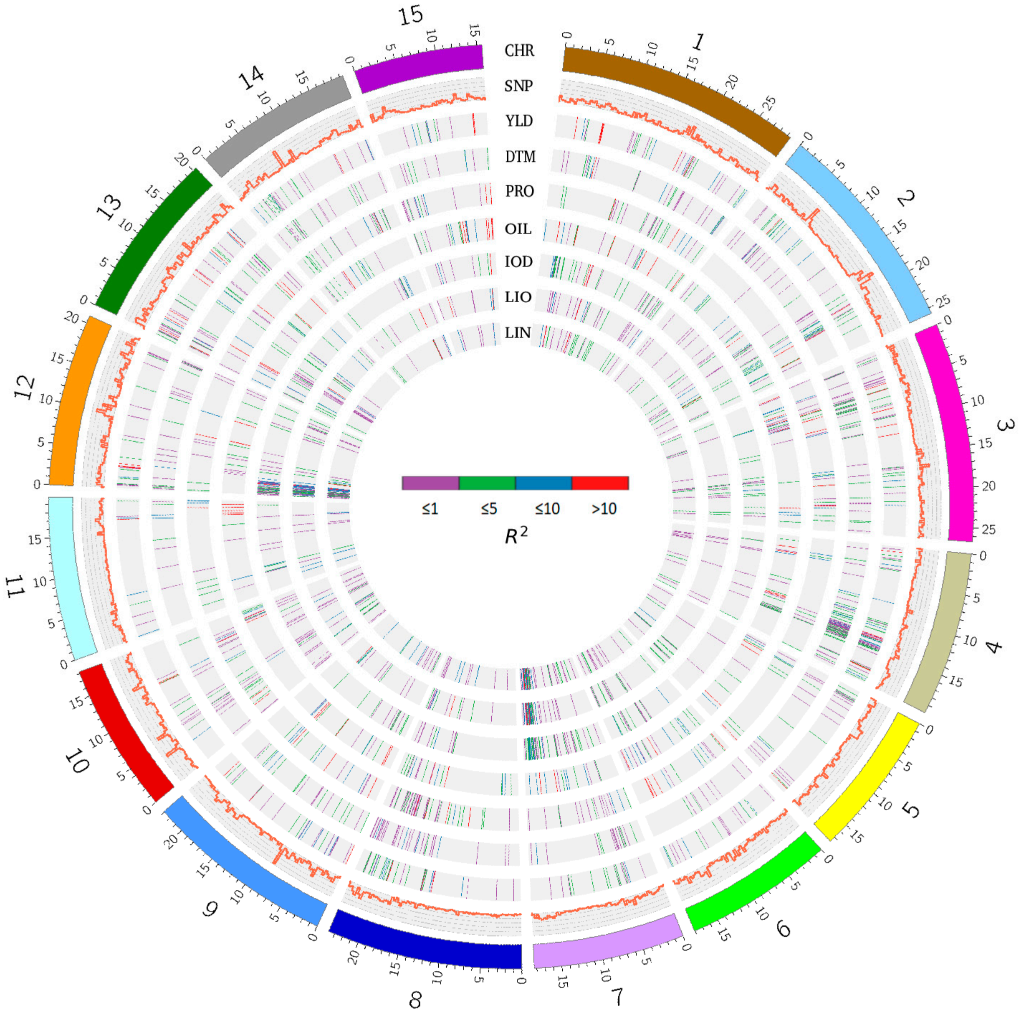


Figure 2. Circos map of quantitative trait nucleotides (QTNs) associated with seven traits in the 260 lines. Track 1 (from outer), chromosomes; Track 2, density of 17,277 SNPs (bin size of 300 kb); Track 3, QTNs for YLD; Track 4, QTNs for DTM; Track 5, QTNs for PRO; Track 6, QTNs for OIL; Track 7, QTNs for IOD; Track 8, QTNs for LIO; Track 9, QTNs for LIN. The effects of QTNs are represented by different colors. $R^2 \leq 1\%$, purple; $1\% < R^2 \leq 5\%$, green; $5\% < R^2 \leq 10\%$, blue; $R^2 > 10\%$, red. YLD, seed yield; DTM, days to maturity; PRO, protein content; OIL, oil content; IOD, iodine value; LIO, linoleic acid content; LIN, linolenic acid content; SNP, single nucleotide polymorphism.

The allelic effects of all QTL are illustrated and summarized in Figures 2 and 3, and Table 2, Tables S2 and S3. Similar QTL effects were observed among the ten statistical models (Figure 3A, Table S3). Using $R^2 \geq 5\%$ as the criterion to define major QTL, 520 of the 1420 unique QTL would be considered major, explaining 12.06 ± 8.24% of the variance. QTL for PRO, OIL, and YLD had relatively larger effects than those of the other four traits (Figure 3B and Table 2). The number of QTL for YLD and OIL exceeded that of the other traits, being 110 (30.5%) and 111 (43.7%), respectively, while the smallest number of major QTL belonged to DTM with 36 out of 351 (10.3%).

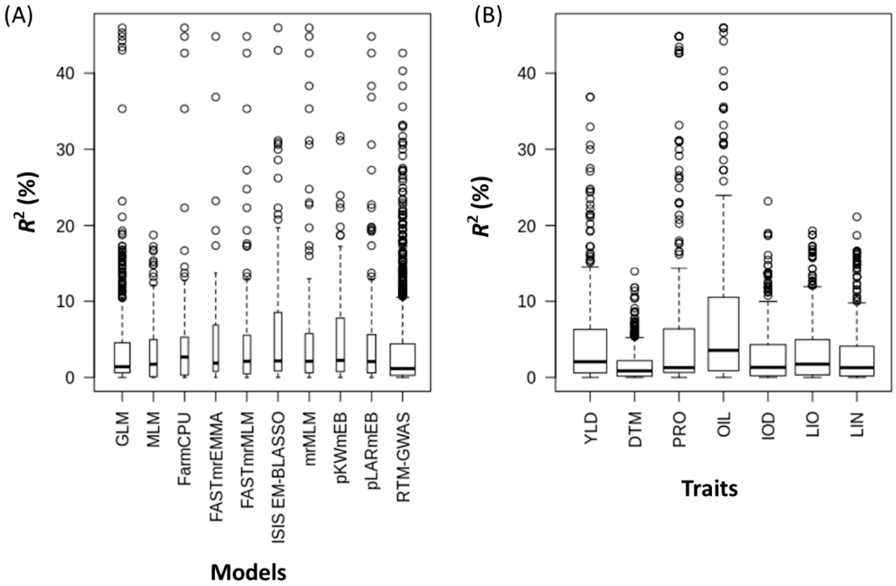


Figure 3. Boxplots of allele effects (R^2) of quntitativ trait loci (QTL) for ten genome-wide association study (GWAS) models (**A**) and seven phenotypic traits (**B**). YLD, seed yield; DTM, days to maturity; PRO, protein content; OIL, oil content; IOD, iodine value; LIO, linoleic acid content; LIN, linolenic acid content.

The GWAS models identified different sets of QTL (Figure 4, Tables S2 and S4). BM detected four times more QTL than the SS+SM and most differed from one another. Of the 1420 QTL, only 215 QTL were shared by both SS+SM and BM, ranging from 18 out of 361 QTL for YLD (5%) to 32 out of 256 QTL for LIN (12.5%). The average allele effect (R^2) of the shared QTL among the three types of models was 2.75%, whereas QTL that were not shared had R^2 of 2.73% for BM, 3.16% for SM, and 2.62% for SS, showing that the shared QTL did not necessarily have greater QTL effects. Between the SNP-based models (SS and SM), the six SM models had more QTL in common with BM than the two SS models (GLM and MLM). SS identified fewer QTL for YLD, DTM, PRO, OIL, and LIO than SM, but a similar number was identified by the two model types for IOD and LIN.

Similarly, seven SNP-based multi-locus models also identified different sets of QTL (Figure 5, Tables S3 and S4). For all seven traits, a total of 355 unique QTL were obtained using the seven SM models (Table 2). Models pKWmEB, pLARmEB and pLARmEB identified 133, 130, and 121 QTL, respectively, followed by ISIS EM-BLASSO (133), FASTmrMLM (96), and FarmCPU (96). FASTmrEMMA identified the fewest QTL (52). More than half of the QTL (an average 58% across the seven traits) identified by the seven SM models were detected by different single models, varying from different traits, ranging from 47.6% (OIL) to 72.4% (LIO). The remaining 42% of the QTL were simultaneously identified by two or more models. Out of 355 QTL, 194 (54.7%), 55 (15.5%), 45 (12.7%), 26 (7.3%), 16 (4.5%), 14 (3.9%), and 5 (1.4%) were identified by a single, two, three, four, five, six, and seven models, respectively. These results indicated that the seven SM models are complementary in QTL identification.

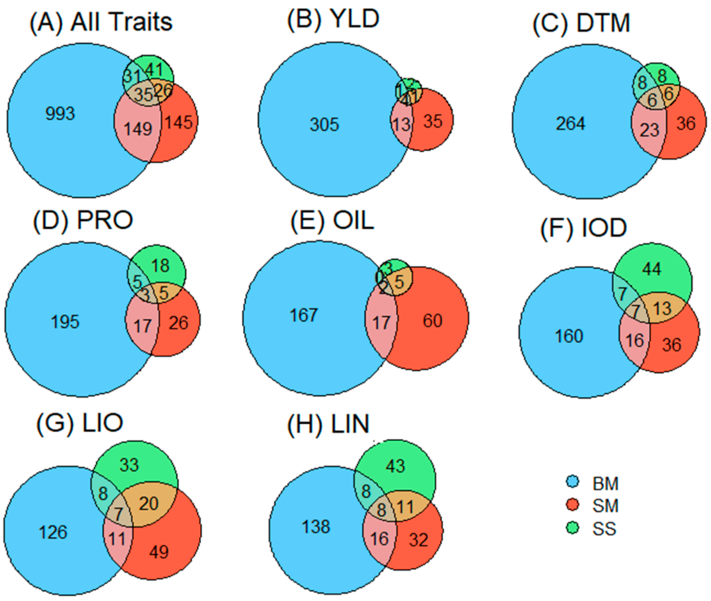


Figure 4. Venn diagrams of quantitative trait loci (QTL) identified by three types of genome-wide association study (GWAS) models for all seven traits (**A**) and individual traits (**B**–**H**). SS, SNP-based single-locus models; SM, SNP-based multi-locus models; BM, haplotype-block-based multi-locus model. YLD, seed yield; DTM, days to maturity; PRO, protein content; OIL, oil content; IOD, iodine value; LIO, linoleic acid content; LIN, linolenic acid content; SNP, single nucleotide polimorphsm.

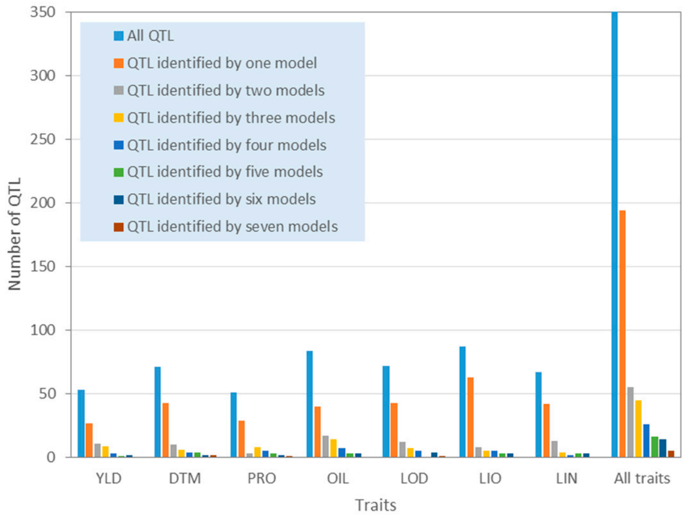


Figure 5. Histograms of quantitative trait loci (QTL) that were identified by one of the seven SNP-based multi-locus models or simultaneously by two or more models for the seven traits. YLD, seed yield; DTM, days to maturity; PRO, protein content; OIL, oil content; IOD, iodine value; LIO, linoleic acid content; LIN, linolenic acid content; SNP, single nucleotide polimorphsm.

2.4. Pleiotropic QTL

Of the 1420 unique QTL identified with all models, 407 were pleiotropic with effects on two or more traits, of which, 239, 139, 25, and 4 QTL were simultaneously associated with 2, 3, 4, and 5 traits, respectively. Some QTL for YLD were associated with DTM as well as PRO and OIL, while many QTL for IOD, LIO, and LIN were co-located (Figure 6). Table 3 lists the number of QTL shared between any two traits. More than 50% of the QTL were shared between any two of LIO, LIN, and IOD. YLD and DTM also had 19% of their respective QTL in common.

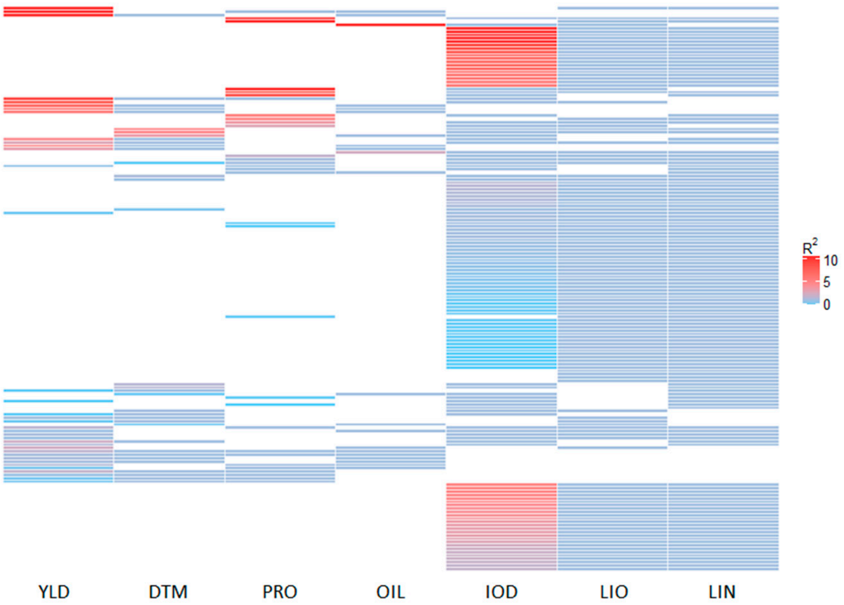


Figure 6. Heatmap of pleiotropic effects of 168 quantitative trait loci (QTL) associated with three or more traits. YLD, seed yield; DTM, days to maturity; PRO, protein content; OIL, oil content; IOD, iodine value; LIO, linoleic acid content; LIN, linolenic acid content.

Table 3. Numbers of quantitative trait loci (QTL) that were pleiotropic on any two of the seven traits.

Trait	YLD	DTM	PRO	OIL	IOD	LIO	LIN
YLD	361	69(19.1,19.7)	28(7.8,10.4)	30(8.3,11.8)	23(6.4,8.1)	17(4.7,6.7)	21(5.8,8.2)
DTM		351	26(7.4,9.7)	29(8.3,11.4)	23(6.6,8.1)	13(3.7,5.1)	14(4.0,5.5)
PRO			269	19(7.1,7.5)	21(7.8,7.4)	17(6.3,6.7)	22(8.2,8.6)
OIL				254	11(4.3,3.9)	9(3.5,3.5)	10(3.9,3.9)
IOD					283	133(47.0,52.4)	162(57.2,63.3)
LIO						254	149(58.7,58.2)
LIN							256

The diagonal values show the number of QTL for individual traits. The two values in parenthesis show percentages of pleiotropic QTL of the two traits of the corresponding row and column. YLD, seed yield; DTM, days to maturity; PRO, protein content; OIL, oil content; IOD, iodine value; LIO, linoleic acid content; LIN, linolenic acid content.

2.5. Genomic Prediction Accuracy

To define the marker sets that generate the best prediction accuracy, we constructed GS models for the seven traits using GBLUP with three types of markers (all SNPs, QTL of all the traits, and QTL of single traits). The QTL marker sets were obtained from four different combinations of GWAS models (SS, SS+SM, BM, and all models, i.e., SS+SM+BM). For the marker type “All SNPs” or the “QTL of all traits”, the same 17,277 SNPs or the same set of QTL of all seven traits (133, 427, 1208, and 1420 QTL for SS, SS+SM, BM, and SS+SM+BM, respectively; Table 2) were used for GS model construction of each trait. However, for the marker type “QTL of single traits”, the specific QTL sets for the respective traits were used as marker sets (Table 2). A joint analysis of variance (ANOVA) of prediction accuracy (*r*) for three factors, namely, traits, GWAS models, and types of markers, was performed. The ANOVA results showed significant differences among traits, marker types, or marker sets due to GWAS models, as well as interactions between the three factors (Table S5).

Among the seven traits, the GS models generated the highest *r* for OIL (0.887 ± 0.058), following by PRO (0.838 ± 0.072), YLD (0.808 ± 0.126), LIO (0.776 ± 0.074), LIN (0.765 ± 0.083), IOD (0.753 ± 0.085), and DTM (0.588 ± 0.150). They were all significantly different from each other at a 0.05 probability level. This trend was consistently observed in terms of QTL identified by different GWAS models (Figure 7) and in terms of QTL of all or single traits (Figure 8).

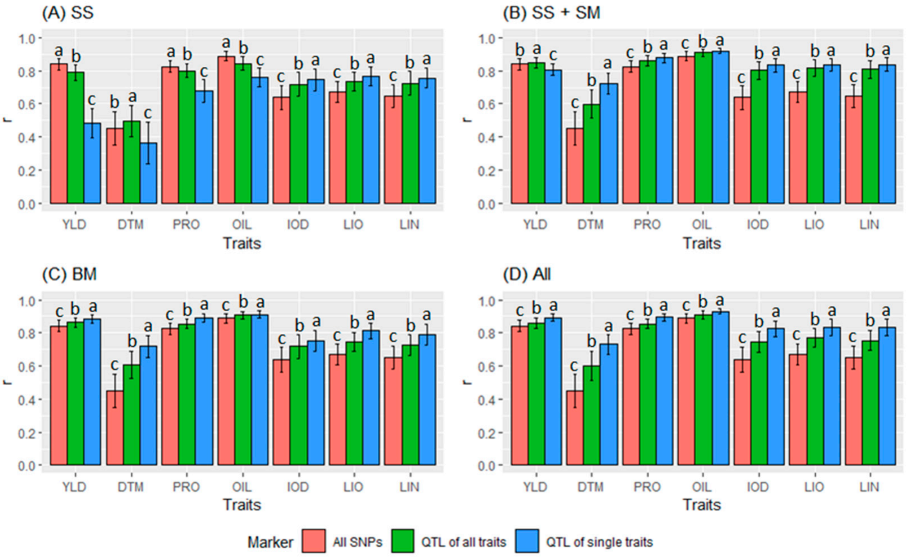


Figure 7. Comparisons of genomic prediction accuracy (*r* ± *s*) using different marker sets, including all single nucleotide polymorphisms (SNPs) and quantitative trait locus (QTL) sets identified by different statistical models: (**A**) SNP based single-locus model (SS), (**B**) SS + SNP based multi-locus model (SM), (**C**) haplotype-block-based model (BM), and (**D**) all three models of SS+SM+BM (All). For each trait, three marker sets were compared for prediction accuracy: All SNPs, QTL of all traits (QTL together for all seven traits), and QTL of single traits (QTL for individual traits). Different letters represent statistical significance of *r* values among different types of markers within each trait. A tag quantitative trait nucleotide (QTN) for each QTL was used for analyses. YLD, seed yield; DTM, days to maturity; PRO, protein content; OIL, oil content; IOD, iodine value; LIO, linoleic acid content (LIO); LIN, linolenic acid content.

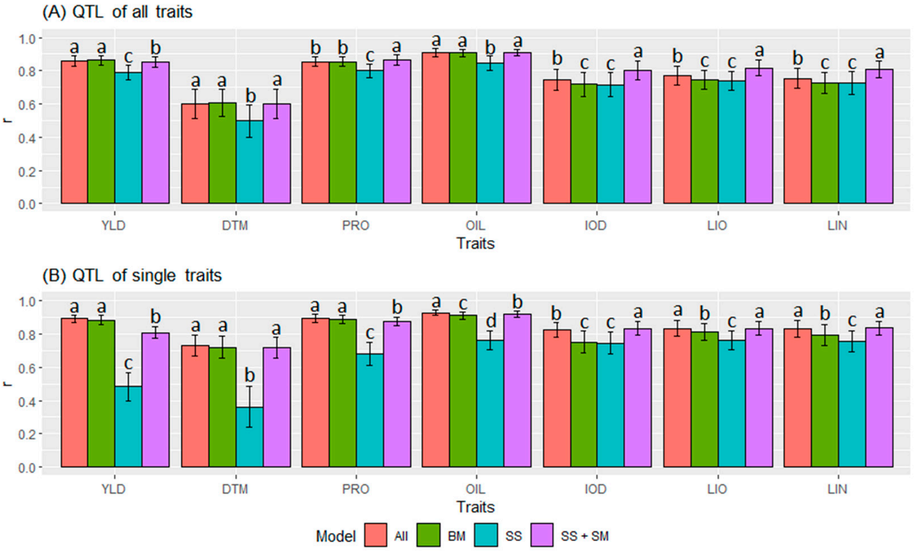


Figure 8. Comparisons of genomic prediction accuracy (*r* ± *s*) by different statistical models, including SNP-based single-locus model (SS), SS+SNP-based multi-locus model (SM), haplotype-block-based model (BM), and all three models of SS+SM+BM (All), which were used for quantitative trait locus (QTL) identification. (**A**) QTL of all traits were used for GS, and (**B**) QTL of single traits were used for GS. A tag quantitative trait nucleotide (QTN) for each QTL was used for analyses. For each trait, different letters represent statistical significance of *r* values among different GWAS models. YLD, seed yield; DTM, days to maturity; PRO, protein content; OIL, oil content; IOD, iodine value; LIO, linoleic acid content (LIO); LIN, linolenic acid content; SNP, single nucleotide polymorphism.

Among the three types of markers, the GS models with the QTL markers (either QTL of all traits or QTL of single traits) identified by SS+SM, BM or all models had significantly greater *r* values than those with all SNPs for all seven traits (Figure 7B–D). An exception was for YLD, DTM, PRO, and OIL when QTL identified by SS were used (Figure 7A). The GS models using single-trait QTL identified by SS+SM (Figure 7B), BM (Figure 7C) or all models (SS+SM+BM) (Figure 7D) performed significantly better than those using QTL of all traits. The average *r* values of the seven traits were 0.789 ± 0.155, 0.774 ± 0.116, and 0.709 ± 0.134 when using QTL of single traits, QTL for all traits, and all SNPs, respectively, and they all significantly differed from each other.

Since more pleiotropic QTL were found between YLD and DTM, between PRO and OIL, and among IOD, LIO, and LIN, we also compared prediction accuracy for all SNPs, single-trait QTL, and the combined QTL of YLD+DTM, PRO+OIL, and IOD+LIO+LIN identified by all statistical models (Table 4). The results showed that the combined marker sets of two or three traits yielded a slightly higher *r* estimates for LIO only, but similar or slightly lower estimates than the ones obtained using the single-trait QTL markers. This indicated that using QTL from more traits did not improve prediction accuracy. Using single-trait QTL marker sets in GS yielded significantly better prediction accuracy.

In terms of QTL marker sets generated by different GWAS models, SS did not identify sufficient QTL markers from YLD, DTM, PRO, and OIL, thus, resulting in low *r* values for these four traits (Table 4, Figure 7A). All GS models using QTL by SS generated lower *r* values than those using QTL by BM, SS+SM, or all models for all seven traits (Table 4, Figure 8) except IOD, LIO, and LIN with all-trait QTL (Figure 8A) and IOD with single-trait QTL (Figure 8B).

BM and SS+SM are two different types of GWAS models. The GS models with QTL identified by SS+SM outperformed BM for IOD, LIN, LIO, and OIL or had similar prediction accuracy for DTM with BM. However, for YLD, BM consistently outperformed SS+SM. For PRO, SS+SM had similar or better performance when all-trait QTL were used (Figure 8A). For the most part, the all-model (SS+SM+BM) had similar to or better results than SS+SM or BM independently (Figure 8, Table 4). Due to significant interactions between marker types and marker sets (Table S5), the GS models with the best prediction accuracy were those using QTL of single traits identified by all GWAS models (SS+SM+BM) for OIL (0.929 ± 0.016), PRO (0.893 ± 0.023), YLD (0.892 ± 0.030), and DTM (0.730 ± 0.062), and by SS+SM for LIN (0.837 ± 0.053), LIO (0.835 ± 0.049), and IOD (0.835 ± 0.041).

In this study, the seven traits were phenotyped in two locations, Morden and Saskatoon, which are representative of the production areas of oilseed flax in Western Canada. To assess the effect of location on genomic prediction and whether or not separate GS models should be constructed in terms of different locations, we compared the prediction accuracy of models using the phenotypic values obtained in Morden and Saskatoon as well as the BLUEs calculated over both locations for the three different types of markers and the seven traits. Only the GS models for YLD at Saskatoon and PRO at Morden performed significantly better than the others. For all other traits, the prediction accuracies were similar regardless of the location-based data set (Table 5 and Table S6). Single-trait QTL for all seven traits as markers significantly improved prediction accuracy compared to all SNPs or all-trait QTL in terms of different locations (Table 5). For all seven traits, the GS models with single-trait QTL had significantly greater prediction accuracy than those with all SNPs or all-trait QTL (Table 5).

Table 4. Prediction accuracy (*r* ± *s*) of seven traits using all single nucleotide polymorphisms (SNPs) and different combinations of quantitative trait loci (QTL) identified by different combinations of statistical models. GBLUP was used to estimate *r* values.

Traits	Models	Marker Sets	No. of Markers	*r* ± *s*
YLD	All	QTL of YLD	361	**0.892 ± 0.023a**
	BM	QTL of YLD	323	**0.885 ± 0.027a**
	All	QTL for YLD + DTM	643	**0.879 ± 0.026a**
	BM	QTL of all traits	1208	0.862 ± 0.030b
	All	QTL of all traits	1420	0.860 ± 0.030b
	SS+SM	QTL of all traits	427	0.850 ± 0.031c
	-	All SNPs	17,277	0.841 ± 0.035d
	SS+SM	QTL of YLD	53	0.807 ± 0.034e
	SS	QTL of all traits	133	0.789 ± 0.045f
	SS	QTL of YLD	8	0.483 ± 0.085g
DTM	All	QTL of DTM	351	**0.730 ± 0.062a**
	SS+SM	QTL of DTM	71	**0.720 ± 0.063a**
	BM	QTL of DTM	301	**0.719 ± 0.066a**
	All	QTL for DTM + YLD	643	0.689 ± 0.076b
	BM	QTL of all traits	1208	0.608 ± 0.083b
	All	QTL of all traits	1420	0.603 ± 0.088b
	SS+SM	QTL of all traits	427	0.599 ± 0.087b
	SS	QTL of all traits	133	0.497 ± 0.095c
	-	All SNPs	17,277	0.449 ± 0.101d
	SS	QTL of DTM	28	0.362 ± 0.125e
PRO	All	QTL of PRO	269	**0.894 ± 0.023a**
	BM	QTL of PRO	220	**0.890 ± 0.024a**
	All	QTL for PRO +OIL	504	**0.879 ± 0.026ab**
	SS+SM	QTL of PRO	51	0.877 ± 0.026b
	SS+SM	QTL of all traits	427	0.864 ± 0.031c
	All	QTL of all traits	1420	0.855 ± 0.031d
	BM	QTL of all traits	1208	0.854 ± 0.030d
	-	All SNPs	17,277	0.825 ± 0.034e
	SS	QTL of all traits	133	0.800 ± 0.042f
	SS	QTL of PRO	31	0.681 ± 0.069g
OIL	All	QTL of OIL	254	**0.929 ± 0.016a**
	All	QTL for PRO + OIL	504	**0.927 ± 0.018a**
	SS+SM	QTL of OIL	84	0.919 ± 0.017b
	BM	QTL of OIL	186	0.911 ± 0.023c
	SS+SM	QTL of all traits	427	0.909 ± 0.021c
	All	QTL of all traits	1420	0.909 ± 0.023c
	BM	QTL of all traits	1208	0.907 ± 0.023c
	-	All SNPs	17,277	0.889 ± 0.028d
	SS	QTL of all traits	133	0.845 ± 0.042e
	SS	QTL of OIL	10	0.762 ± 0.058f
IOD	SS+SM	QTL of IOD	72	**0.835 ± 0.041a**
	All	QTL of IOD	283	**0.824 ± 0.046a**
	All	QTL for IOD + LIO + LIN	468	**0.825 ± 0.051a**
	SS+SM	QTL of all traits	427	0.801 ± 0.055b
	BM	QTL of IOD	190	0.752 ± 0.066c
	SS	QTL of IOD	71	0.746 ± 0.065c
	All	QTL of all traits	1420	0.745 ± 0.066c
	BM	QTL of all traits	1208	0.717 ± 0.072d
	SS	QTL of all traits	133	0.717 ± 0.072d
	-	All SNPs	17,277	0.639 ± 0.073e
LIO	All	QTL for IOD + LIO + LIN	468	**0.836 ± 0.043a**
	SS+SM	QTL of LIO	87	**0.835 ± 0.039a**
	All	QTL of LIO	254	**0.834 ± 0.048a**
	SS+SM	QTL of all traits	427	0.817 ± 0.049b
	BM	QTL of LIO	152	0.812 ± 0.049b
	All	QTL of all traits	1420	0.770 ± 0.055c
	SS	QTL of LIO	68	0.765 ± 0.056c
	BM	QTL of all traits	1208	0.744 ± 0.058d
	SS	QTL of all traits	133	0.736 ± 0.066d
	-	All SNPs	17,277	0.672 ± 0.063e
LIN	SS+SM	QTL of LIN	67	**0.837 ± 0.041a**
	All	QTL of LIN	256	**0.833 ± 0.051a**
	All	QTL for IOD + LIO + LIN	468	**0.830 ± 0.047a**
	SS+SM	QTL of all traits	427	0.809 ± 0.053b
	BM	QTL of LIN	170	0.792 ± 0.062c
	SS	QTL of LIN	70	0.756 ± 0.062d
	All	QTL of all traits	1420	0.755 ± 0.061d
	BM	QTL of all traits	1208	0.727 ± 0.066e
	SS	QTL of all traits	133	0.725 ± 0.070e
	-	All SNPs	17,277	0.649 ± 0.069f

Letters indicate significant difference at α= 0.05 level. Tukey's multiple range test was used. The highest prediction accuracy of each trait is highlighted in bold font. SS, SNP-based single-locus model; SM, SNP-based multi-locus model; BM, block-based model; All, SS+SM+BM; seed yield; DTM, days to maturity; PRO, protein content; OIL, oil content; IOD, iodine value; LIO, linoleic acid content; LIN, linolenic acid content.

Table 5. Genomic prediction accuracy (*r* ± *s*) of seven traits affected by different locations.

Trait	Overall			Saskatoon, Saskatchewan			Morden, Manitoba		
	17,277 SNPs	**All-Trait QTL**	**Single-Trait QTL**	**17,277 SNPs**	**All-Trait QTL**	**Single-Trait QTL**	**17,277 SNPs**	**All-Trait QTL**	**Single-Trait QTL**
YLD	0.84 ± 0.03 ij	0.86 ±0.03 h	**0.89 ± 0.02 efg**	0.88 ± 0.02 g	0.89 ± 0.02 defg	**0.91 ± 0.02 cde**	0.79 ± 0.04 n	0.82 ± 0.04 lm	**0.85 ± 0.04 hij**
DTM	0.45 ± 0.10 x	0.60 ± 0.09 v	**0.73 ± 0.06 q**	0.51 ± 0.09 w	0.61 ± 0.08 v	**0.70 ± 0.07 r**	0.32 ± 0.12 y	0.52 ± 0.11 w	**0.67 ± 0.07 s**
PRO	0.82 ± 0.03 klm	0.86 ± 0.03 hi	**0.89 ± 0.02 defg**	0.81 ± 0.04 mn	0.84 ± 0.03 ijk	**0.89 ± 0.02 fg**	0.88 ± 0.02 fg	0.90 ± 0.02 cdef	**0.91 ± 0.02 bcd**
OIL	0.89 ± 0.03 fg	0.91 ± 0.02 cd	**0.93 ± 0.02 a**	0.89 ± 0.03 defg	0.91 ± 0.02 bcd	**0.93 ± 0.02 ab**	0.88 ± 0.03 g	0.90 ± 0.02 cdef	**0.92 ± 0.02 abc**
IOD	0.64 ± 0.07 tu	0.75 ± 0.07 p	**0.82 ± 0.05 klm**	0.63 ± 0.07 u	0.74 ± 0.06 pq	**0.82 ± 0.05 lm**	0.66 ± 0.07 st	0.75 ± 0.06 op	**0.83 ± 0.04 jklm**
LIO	0.67 ± 0.06 s	0.77 ± 0.05 o	**0.83 ± 0.05 jkl**	0.67 ± 0.06 s	0.77 ± 0.05 o	**0.83 ± 0.05 jklm**	0.67 ± 0.06 s	0.77 ± 0.05 o	**0.83 ± 0.05 jkl**
LIN	0.65 ± 0.07 tu	0.75 ± 0.06 op	**0.83 ± 0.05 jkl**	0.65 ± 0.07 tu	0.75 ± 0.06 op	**0.82 ± 0.05 klm**	0.65 ± 0.07 stu	0.76 ± 0.06 op	**0.84 ± 0.05 jkl**

The highest prediction accuracy among different marker types is highlighted in bold font. Single-trait QTL, quantitative trait loci (QTL) identified using all models for a specific trait, i.e., a different marker set for each trait; All-trait QTL, all unique QTL identified using all models from all seven traits, i.e., the same marker set for all seven trait; Overall, phenotype BLUEs over four years and two locations, Morden, Manitoba and Saskatoon, Saskatchewan; YLD, seed yield; DTM, days to maturity; PRO, protein content; OIL, oil content; IOD, iodine value; LIO, linoleic acid content; LIN, linolenic acid content; SNP, single nucleotide polymorphism. The letters after *r* ± s values represent statistical significance of *r* values among 63 combinations of seven traits, three marker sets, and three location levels (two locations plus overall BLUEs over two locations).

3. Discussion

A good training population in GS has a strong relationship with the test populations in breeding and may include germplasm genotypes for parent selection or breeding lines for offspring selection. In the present study, all lines used for GS evaluation were derived from three bi-parental crosses [30,31]. The two parents of the first cross were Canadian high-yielding conventional linseed cultivars with high LIN of 55–57% (CDC Bethune and Macbeth). The second population resulted from a cross between a low LIN breeding line (E1747) and a European fiber flax cultivar with ~55% LIN (Viking). The third cross had two parents of a yellow-seeded and low LIN (2–3%) cultivar (Solin^{TM} SP2047) and a high LIN breeding line with 63–66% LIN (UGG5-5). Therefore, this genetic panel exhibited diversity in genetic variation in major breeding selection traits [30,31]. Although these breeding lines were derived from a few parents, they are close to breeding populations. Therefore the results obtained herein apply to practical breeding.

Given a training population in practical breeding, markers will be a critical factor for improving prediction accuracy since GS predicts breeding values of selection traits using a set of markers [2]. Prediction accuracy directly assesses the efficiency of a marker set in GS. Here, using prediction accuracy, we consistently demonstrated that QTL markers outperformed genome-wide random SNPs for GS of any traits, further confirming and validating the results observed for pasmo resistance using a flax core germplasm collection of 370 accessions [5]. The use of QTL identified by GWAS models significantly increased prediction accuracy for all seven traits, from 4% for OIL (from 0.89 to 0.93) to 29% for DTM (from 0.45 to 0.73) compared to genome-wide random SNPs (Table 4). The reasons that QTL outperformed genome-wide random SNPs are likely a reduction in background noises or as a consequence of reduced multi-collinearity due to the removal of unrelated markers.

Many statistical models of GWAS have been proposed to identify QTL. In this study, we investigated three types of models, including two SS, seven SM, and one BM, totaling ten different models. However, it seemed that different models generated varying sets of QTL in which only a small portion of QTL was shared by two or more models (Figures 4 and 5, Table S4). Similar results were also obtained in the previous study of QTL identification for pasmo resistance in flax, where the same SS and SM models were used [10]. The two SS methods (GLM and MLM) identified only 133 QTL for all seven traits, accounting for 9% of 1420 QTL, whereas the seven SM methods identified 355 QTL, accounting for 25% of the total QTL. One haplotype block-based model, RTM-GWAS, identified a total of 1208 QTL alone (85%), three times the total QTL identified by the nine SNP-based models (SS+SM). A haplotype-block-based GWAS is expected to increase power relative to SNP-based approaches, resulting in a higher number of QTL identified. First, the block-based approach reduces the dimension of association testing when a single global test for a block is used and thus preserves power and helps maintain reasonable false-positive rates. Second, a haplotype method also captures associations of nearby SNPs that would have been otherwise missed with an SNP-by-SNP approach [32]. Because different algorithms and assumptions are adopted in different models, their QTL results may be complementary in GS.

We evaluated the performance of different sets of QTL markers identified by different models via prediction accuracy. The results indicated that two SS models did not identify sufficient QTL for YLD, DTM, PRO, and OIL, resulting in low prediction accuracy as compared with all SNPs, whereas SS+SM+BM or SS+SM identified sufficient QTL to yield the highest prediction accuracies for all seven traits, strongly suggesting that the advantages of different statistical models are complementary and the combined results from different models improve prediction accuracy. In terms of the number of QTL identified and prediction accuracy, the combined use of SNP-based models (SS+SM) was superior to other models or their combinations since only a small number of QTL were identified by SS+SM compared to BM, but similar or better prediction accuracies were obtained for most traits. The QTL identified by BM was three times greater than those identified by SS+SM, but BM significantly outperformed SS+SM only for YLD and PRO. While BM and SS+SM had similar prediction accuracies for DTM, SS+SM was significantly superior to BM for the remaining four traits: OIL, IOD, LIO, and

LIN (Table 4). These results implied that the combined use of different GWAS models facilitates the identification of a potentially complete set of QTL associated with the traits, but some of them may be redundant or possibly false positives. Therefore, further investigations to design a methodology to identify and remove the redundant or false-positive QTL that would maximize prediction accuracy and minimize the number of QTL markers in GS are warranted.

The heritability of a trait is an important factor that affects the efficiency of genomic selection over traditional phenotypic selection. Generally, high prediction accuracies are more easily achieved with high heritability traits [8]. Conversely, genomic selection is likely more beneficial for traits with low heritability [33,34]. In this study, the broad-sense heritability (H^2), representing the extent with which the performance of a trait is affected by the environment, was estimated for the seven traits (Table 6). Compared to the maximum prediction accuracy (r) of each trait, the H^2 of the traits did not exhibit a consistent relationship with prediction accuracy. OIL with a moderate estimate (0.69) produced the highest prediction accuracy (0.93). The three fatty acid composition related traits, LIO, LIN, and IOD, had a relatively high H^2 values (0.81–0.83) and a similarly high prediction accuracy. Albeit with low heritability, YLD (0.44) and PRO (0.20) generated the second-highest prediction accuracy (0.89). However, considering the relative efficiency of genomic prediction over phenotypic selection (*RE*), which is defined as r/ H^2 [35], the traits with a low H^2 had a high *RE*, exhibiting a strong negative correlation (Table 6). Especially YLD with H^2 of 0.2 generated as high as 4.45 times selection efficiency over phenotypic selection, demonstrating more benefits of GS for low heritability traits. Based on *RE*, GS for YLD, DTM, PRO and OIL outperformed phenotypic selection, whereas GS for IOD, LIO and LIN were equivalent to or slightly better than phenotypic selection. A similar trend for YLD, OIL, IOD, LIO and LIN was also observed when a limited number of microsatellite markers were used [30]. Compared to H^2, the prediction accuracy of a trait was more dependent on genomic heritability that represents a proportion of additive genetic variation explained by the markers (Table 6). In other words, prediction accuracy mostly depends on whether the marker set contains sufficient QTL to contribute to the total variation of the phenotypes, or whether all related QTL have been identified from the marker set if QTL markers are used in GS models.

Table 6. Broad-sense and genomic heritability of seven traits.

Trait	Broad-Sense Heritability (H^2)	Genomic Heritability Based on Single Trait QTL (h^2)	Genomic Heritability Based on 1420 QTL of 7 Traits (h^2)	Genomic Heritability Based on 17,277 SNPs (h^2)	Maximum Perdition Accuracy (r)	Relative Efficiency (r/H^2)
YLD	0.20 ± 0.02	0.68 ± 0.06	0.62 ± 0.08	0.62 ± 0.09	0.89 ± 0.02	4.45
DTM	0.49 ± 0.03	0.58 ± 0.08	0.59 ± 0.09	0.46 ± 0.11	0.73 ± 0.06	1.49
PRO	0.44 ± 0.04	0.71 ± 0.08	0.62 ± 0.08	0.62 ± 0.09	0.89 ± 0.02	2.02
OIL	0.69 ± 0.03	0.66 ± 0.06	0.72 ± 0.07	0.73 ± 0.07	0.93 ± 0.02	1.35
IOD	0.81 ± 0.02	0.73 ± 0.05	0.73 ± 0.07	0.72 ± 0.07	0.84 ± 0.04	1.04
LIO	0.84 ± 0.02	0.73 ± 0.05	0.74 ± 0.07	0.74 ± 0.07	0.84 ± 0.04	1.00
LIN	0.83 ± 0.02	0.76 ± 0.05	0.73 ± 0.07	0.73 ± 0.07	0.84 ± 0.04	1.01

YLD, seed yield; DTM, days to maturity; PRO, protein content; OIL, oil content; IOD, iodine value; LIO, linoleic acid content; LIN, linolenic acid content; SNP, single nucleotide polymorphism; QTL, quantitative trait loci.

Pleiotropy of genes has been thought to be the molecular basis of trait genetic correlation. We have identified highly significant correlations between YLD and DTM, between PRO and OIL, and among IOD, LIO, and LIN (Table S7) [30,31]. Correspondingly, we also identified many pleiotropic QTL between these traits in the present (Table S2 and Table 3, Figure 5) and previous studies [31], suggesting that different traits may be genetically controlled by the same or tightly linked genes/QTL. Our hypothesis is that if some QTL are pleiotropic to two or more traits, all the QTL identified from genetically-related traits could be used as markers in GS to improve prediction accuracy. Therefore, we evaluated GS accuracy of different marker sets, including QTL of single traits, QTL of all seven traits, and QTL of some combinations of related traits (YLD+DTM, PRO+OIL, IOD+LIO+LIN). Our results rejected the hypothesis, indicating that QTL from pleiotropic traits did not improve GS accuracy for any of the seven traits. However, this does not necessarily signify that the pleiotropic QTL do not have a role in improving GS accuracy because QTL identified from each single trait already includes QTL pleiotropic to other traits and additional unrelated QTL from other traits thereby reducing prediction accuracy as a consequence of redundancy or background noise. Thus, our results strongly suggest that QTL from single traits can not only significantly improve prediction accuracy but also reduce the number of markers, which in turn would decrease genotyping cost in practical breeding programs compared with the use of all SNPs or QTL of all traits or QTL of any trait combinations.

Significant genotype by environment (GXE) interactions (Table S1, Figure 1) hinted at the potential need for separate GS models for different breeding target regions in order to maximize GS accuracy. We constructed separate GS models for two locations: Saskatoon and Morden, using phenotypic data observed from the two locations as well as GS models using BLUEs over years and locations. Only the GS models for YLD at Saskatoon and PRO at Morden had higher GS accuracies than any of the other models because these two traits had the largest GXE interaction, although significant GXE interactions also existed for the other five traits (Table S1). This suggested that genomic selection based on BLUEs over years and locations is suitable for traits with moderate or no GXE, but higher accuracies are obtained if GS is performed using by location for traits with high GXE.

GS applied in practical breeding requires not only a high prediction accuracy but also an acceptable cost. Although GBS is a most popular genotyping approach to obtain high density genome-wide random SNPs, it is not an efficient genotyping approach for GS. It generates a large number of unused SNPs. The cost is also a limiting factor for a GS scheme with a large genome, such as wheat. In addition, it is prone to generate missing data in low-coverage sequencing. Recently, some new target-oriented genotyping methods have been developed for breeding, such as genotyping by target sequencing (GBTS) [36], and RAD capture (Rapture) [37]. These methods enable low-cost, high-read coverage genotyping of target loci, and also allow previous training data based on non-captured GBS to be fully compatible with new rapture data [38]. Using GBTS, for example, only USD 12.36 per sample for 5000 target markers of the 2.3 Gb maize genome was needed [36], a much cheaper option than GBS [4,39]. The Rapture assay consistently outperformed the GBS assay, and its cost per sample was approximately 40% less than GBS in oat, a crop with a genome size of 12.5 GB [38]. Therefore, QTL identification by single-locus and multi-locus GWAS models combined with new target-oriented genotyping methods facilitate the implementation of a highly efficient genomic selection scheme in modern plant molecular breeding.

4. Materials and Methods

4.1. Plant Materials, SNPs and Phenotypic Data

A total of 260 lines derived from three different bi-parental populations was used as a genotype panel for the association study and genomic selection evaluation. These lines consisted of 97 F6-derived recombinant inbred lines (RILs) generated by single seed descent from a cross between two Canadian high-yielding conventional linseed cultivars CDC Bethune and Macbeth, 91 F6-derived RILs from a cross between a low LIN breeding line E1747 and a French fiber flax cultivar Viking, and 72 F1-derived doubled haploid (DH) lines obtained from a cross between two breeding lines SP2047 (low LIN, 2–3%) and UGG5-5 (high LIN, 63–66%). The details have been previously described [30,31].

Reduced representation libraries from the 260 lines were re-sequenced by the Michael Smith Genome Sciences Centre of the BC Cancer Agency, Genome British Columbia (Vancouver, BC, Canada) using 100-bp paired-end reads on an Illumina HiSeq 2000 platform (Illumina Inc., San Diego, CA, USA) as previously described [40]. The short reads were aligned to the flax scaffold sequences of cultivar CDC Bethune [41], and SNPs were called and filtered using the revised AGSNP pipeline [40,42,43]. Final SNPs with a MAF ≥ 0.01 and a genotyping rate ≥ 60% were used for further imputation using Beagle v.4.2 [44] to estimate missing data. The coordinates of all SNPs based on scaffolds were converted to the new chromosome-based flax pseudomolecules v2.0 [45].

All lines were evaluated in field trials over four years (2009–2012) at two sites, Morden Research and Development Centre, Manitoba (MD) and Kernen Crop Research Farm near Saskatoon, Saskatchewan (SAS) in Canada. A type-2 modified augmented design (MAD) [46] was used for the field experiments from which phenotypic data were collected. The detailed experimental design was previously described [30,31,47]. Seven major breeding selection traits were evaluated, including YLD, DTM, PRO, OIL, IOD, LIO, and LIN. The methods and criteria used for the evaluation of these traits are detailed in [31]. All phenotypic data from the field experiments and laboratory measurements were adjusted for soil heterogeneity, as previously described, based on the MAD pipeline [47]. The BLUE values over multiple environmental phenotypes estimated using TASSEL [48] were used for further association study analyses. The Shapiro–Wilk normality test was performed for all traits using the R function "shapiro.test". All seven traits followed approximately a normal or mixed normal distribution.

4.2. Identification of Haplotype Blocks

The software RTM-GWAS [26] was used in identifying haplotype blocks. RTM-GWAS provides a function module to group sequential SNPs into linkage disequilibrium blocks (SNPBDBs), using the block-partitioning approach with confidence interval based on genome-wide D' pattern [49]. The software requires SNP data in VCF format. The default values for all the other parameters were used, including the minimum minor haplotype frequency (0.01), and the maximum length of blocks (100 kb).

4.3. QTL Identification

Three types of GWAS models were used to identify putative QTNs associated with the seven traits. These models included two traditional SNP-based single-locus models (GLM [12] and MLM [13]), seven SNP-based multi-locus models (pLARmEB, pKWmEB, FASTmrMLM, FASTmrEMMA, ISIS EM-BLASSO, and mrMLM implemented in the R package mrMLM, https://cran.r-project.org/web/packages/mrMLM/index.html, and FarmCPU [20] implemented in the R package MVP, https://github.com/XiaoleiLiuBio/MVP), and one haplotype block-based model RTM-GWAS [26]. Kinship genetic relationship matrix was estimated using the protocol suggested by each GWAS software package. The population structure of the 260 lines was estimated using principal component analysis (PCA) using TASSEL [48], and the first five principal components (PCs) accounting for 72.35% of the total variation were chosen as covariates in all GWAS models. GWAS were conducted separately for each phenotype data sets from the four individual years and two locations and the BLUE dataset over years and locations for each trait to identify all stable or environment-specific QTL. Thus, all QTNs from different phenotype data sets were merged for analyses.

For GLM, MLM, and FarmCPU, the threshold of significant marker-trait associations was determined by a critical p-value ($\alpha = 0.05$) subjected to Bonferroni correction, i.e., the corrected p-value $= 2.89 \times 10^{-6}$ (0.05/17,277 SNPs). For the six models implemented in the mrMLM R package, a log of odds (LOD) score of three was used to detect robust marker–trait association signals for these six methods.

The identified QTNs were further grouped into QTN clusters or QTL based on the haplotype blocks generated by RTM-GWAS. The SNPs within the same block were treated as a QTN cluster or a QTL. The QTN with the largest R^2 within a QTN cluster was selected as a tag QTN for that cluster or QTL.

4.4. Genomic Selection (GS) Models and Evaluation

The statistical model Genomic BLUP (GBLUP) implemented in the R package BGLR [50] was used to evaluate prediction accuracy for different marker sets. The computation procedures of GBLUP have been described in detail [51,52]. When preparing QTL marker data for model construction, the positive-effect allele of the tag QTN/SNP of a QTL was coded "1" and the alternative allele "−1". Similarly for the SNP marker set, the reference allele of an SNP was coded "1" and the alternative allele

"−1". Missing data were coded "0". The EM algorithm implemented in the R package rrBLUP [53] was used to impute the missing marker data.

The five-fold random cross-validation was used to evaluate GS models. The 260 lines were randomly partitioned into five subsets. For a given partition, each subset was, in turn, used as test data, while the remaining four subsets were used as a training dataset. This partitioning was repeated 50 times. The accuracy of the genomic predictions (r) was defined by the Pearson correlation coefficient between the GBEV values predicted by GS and the observed phenotypic values. To compare GS models constructed from different markers, a joint analysis of variance with Tukey's multiple pairwise-comparisons (HSD.test function) was performed to test the statistical significance of differences in r values using the R package agricolae (https://cran.r-project.org/web/packages/agricolae/index.html).

4.5. Estimation of Broad-sense and Genomic Heritability

Broad-sense heritability of phenotypes for the traits was estimated using the inter-environment correlation method [54]. Genomic heritability of the traits is a molecular marker based heritability parameter that explains a portion of the additive genetic variance (σ_A^2): $h^2 = \sigma_A^2/(\sigma_A^2+\sigma_e^2)$. It was estimated using the R package sommer with the GBLUP model [55].

5. Conclusions

In this study, we adopted a set of genomic and phenotypic data, including 260 lines derived from bi-parental populations, 17,277 genome-wide random SNPs, and phenotypes of seven major breeding selection traits in flax, which were evaluated in four years and two locations, to find optimal markers for maximizing prediction accuracy and minimizing cost of genotyping in breeding selection for these important traits. Our results confirmed and validated that the use of QTL significantly increases prediction accuracy compared to genome-wide random SNPs and cuts down the cost of genotyping of test populations since the number of markers used in GS models have been dramatically reduced to a magnitude of dozens to hundreds rather than a scale of thousands, even hundreds of thousands. In the evaluation of GS models, we compared QTL identified by different types of GWAS models and also QTL from a single trait or QTL from all traits. The results indicated that the highest prediction accuracy of individual traits was obtained by using QTL of respective traits identified by SS+SM+BM or SS+SM, rather than using all genome-wide random markers or QTL of all seven traits. In terms of the number of QTL identified and prediction accuracy, SS+SM outperformed other models or their combinations for most traits. Our work demonstrates that the combined use of single- and multi-locus GWAS models can identify sufficient QTL of traits and significantly improve prediction accuracy, but some redundancy or false-positives may exist in QTL identified by some GWAS models, especially in those by BM. Therefore, further investigation of detection and removal of the redundant or false-positive QTL to maximize prediction accuracy and minimize the number of QTL markers in GS is warranted.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/21/5/1577/s1. Table S1. Results of analysis of variance (ANOVA) for the seven traits; Table S2. Quantitative trait nucleotides (QTNs)/quantitative trait loci (QTL) identified for the seven traits; Table S3. Average allele effects of quantitative trait loci (QTL) identified by different statistical models for all the seven traits; Table S4. Summary statistics of quantitative trait loci (QTL) identified by seven multi-locus models for the seven traits; Table S5. Analysis of variance for genomic prediction accuracy (r) of genomic selection models constructed by different traits, statistical models and marker sets; Table S6. Analysis of variance (ANOVA) for genomic prediction accuracy (r) in terms of locations (Morden and Saskatoon, Canada, average of two locations), traits (YLD, DTM, PRO, OIL, IOD, LIO and LIN), and marker types (all SNPs, QTL of single traits and QTL of all traits); Table S7. Pearson correlation coefficients of phenotypes among the seven traits.

Author Contributions: F.M.Y. and S.C. designed and supervised the study. S.C. performed sequencing. S.D.D. and H.M.B. performed the phenotyping. S.L., C.Z., K.H., M.M. and F.M.Y. analyzed the data. F.M.Y. and S.L. drafted and S.C. edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Genome Canada and other industrial stakeholders for the Total Utilization Flax GENomics (TUFGEN) project, by Genome Canada and Agriculture and Agri-Food Canada (AAFC) for the 4DWheat project, and by Agriculture and Agri-Food Canada for the TUGBOAT project.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Abbreviations

DTM	days to maturity
GBS	genotype by sequencing
GEBV	genomic estimate of breeding value
GWAS	genome-wide association study
IOD	iodine value
LD	linkage disequilibrium
LIN	linolenic acid
LIO	linoleic acid
MAF	minor allele frequency
OIL	oil content
QTN	quantitative trait nucleotide
QTL	quantitative trait locus/loci
SNP	single nucleotide polymorphism
YLD	seed yield

References

1. Desta, Z.A.; Ortiz, R. Genomic selection: Genome-wide prediction in plant improvement. *Trends Plant Sci.* **2014**, *19*, 592–601. [CrossRef] [PubMed]
2. Meuwissen, T.H.; Hayes, B.J.; Goddard, M.E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **2001**, *157*, 1819–1829. [PubMed]
3. Lipka, A.E.; Kandianis, C.B.; Hudson, M.E.; Yu, J.; Drnevich, J.; Bradbury, P.J.; Gore, M.A. From association to prediction: Statistical methods for the dissection and selection of complex traits in plants. *Curr. Opin. Plant. Biol.* **2015**, *24*, 110–118. [CrossRef]
4. Bassi, F.M.; Bentley, A.R.; Charmet, G.; Ortiz, R.; Crossa, J. Breeding schemes for the implementation of genomic selection in wheat (*triticum* spp.). *Plant Sci.* **2016**, *242*, 23–36. [CrossRef]
5. He, L.; Xiao, J.; Rashid, K.Y.; Jia, G.; Li, P.; Yao, Z.; Wang, X.; Cloutier, S.; You, F.M. Evaluation of genomic prediction for pasmo resistance in flax. *Int. J. Mol. Sci.* **2019**, *20*, 359. [CrossRef]
6. Wang, Q.; Yu, Y.; Yuan, J.; Zhang, X.; Huang, H.; Li, F.; Xiang, J. Effects of marker density and population structure on the genomic prediction accuracy for growth trait in pacific white shrimp *litopenaeus vannamei*. *BMC Genet.* **2017**, *18*, 45. [CrossRef]
7. Norman, A.; Taylor, J.; Edwards, J.; Kuchel, H. Optimising genomic selection in wheat: Effect of marker density, population size and population structure on prediction accuracy. *G3 (Bethesda)* **2018**, *8*, 2889–2899. [CrossRef]
8. Ali, M.; Zhang, Y.; Rasheed, A.; Wang, J.; Zhang, L. Genomic prediction for grain yield and yield-related traits in chinese winter wheat. *Int. J. Mol. Sci.* **2020**, *21*, 1342. [CrossRef]
9. Wang, S.B.; Feng, J.Y.; Ren, W.L.; Huang, B.; Zhou, L.; Wen, Y.J.; Zhang, J.; Dunwell, J.M.; Xu, S.; Zhang, Y.M. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* **2016**, *6*, 19444. [CrossRef]
10. He, L.; Xiao, J.; Rashid, K.Y.; Yao, Z.; Li, P.; Jia, G.; Wang, X.; Cloutier, S.; You, F.M. Genome-wide association studies for pasmo resistance in flax (*linum usitatissimum* L.). *Front. Plant Sci.* **2019**, *9*, 1982. [CrossRef]
11. You, F.M.; Jia, G.; Xiao, J.; Duguid, S.D.; Rashid, K.Y.; Booker, H.M.; Cloutier, S. Genetic variability of 27 traits in a core collection of flax (*linum usitatissimum* L.). *Front. Plant Sci.* **2017**, *8*, 1636. [CrossRef] [PubMed]
12. Price, A.L.; Patterson, N.J.; Plenge, R.M.; Weinblatt, M.E.; Shadick, N.A.; Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **2006**, *38*, 904–909. [CrossRef] [PubMed]
13. Yu, J.; Pressoir, G.; Briggs, W.H.; Vroh Bi, I.; Yamasaki, M.; Doebley, J.F.; McMullen, M.D.; Gaut, B.S.; Nielsen, D.M.; Holland, J.B.; et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **2006**, *38*, 203–208. [CrossRef] [PubMed]

14. Segura, V.; Vilhjalmsson, B.J.; Platt, A.; Korte, A.; Seren, U.; Long, Q.; Nordborg, M. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* **2012**, *44*, 825–830. [CrossRef] [PubMed]
15. Li, H.; Zhang, L.; Hu, J.; Zhang, F.; Chen, B.; Xu, K.; Gao, G.; Li, H.; Zhang, T.; Li, Z.; et al. Genome-wide association mapping reveals the genetic control underlying branch angle in rapeseed (*brassica napus* L.). *Front. Plant Sci.* **2017**, *8*, 1054. [CrossRef] [PubMed]
16. Wen, Y.J.; Zhang, H.; Ni, Y.L.; Huang, B.; Zhang, J.; Feng, J.Y.; Wang, S.B.; Dunwell, J.M.; Zhang, Y.M.; Wu, R. Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* **2017**, *19*, 700–712. [CrossRef]
17. Zhang, J.; Feng, J.Y.; Ni, Y.L.; Wen, Y.J.; Niu, Y.; Tamba, C.L.; Yue, C.; Song, Q.; Zhang, Y.M. Plarmeb: Integration of least angle regression with empirical bayes for multilocus genome-wide association studies. *Heredity* **2017**, *118*, 517–524. [CrossRef]
18. Tamba, C.L.; Ni, Y.L.; Zhang, Y.M. Iterative sure independence screening em-bayesian lasso algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* **2017**, *13*, e1005357. [CrossRef]
19. Ren, W.L.; Wen, Y.J.; Dunwell, J.M.; Zhang, Y.M. Pkwmeb: Integration of kruskal-wallis test with empirical bayes under polygenic background control for multi-locus genome-wide association study. *Heredity* **2017**, *120*, 208–218. [CrossRef]
20. Liu, X.; Huang, M.; Fan, B.; Buckler, E.S.; Zhang, Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* **2016**, *12*, e1005767. [CrossRef]
21. Pan, L.; He, J.; Zhao, T.; Xing, G.; Wang, Y.; Yu, D.; Chen, S.; Gai, J. Efficient qtl detection of flowering date in a soybean ril population using the novel restricted two-stage multi-locus gwas procedure. *Theor. Appl. Genet.* **2018**, *131*, 2581–2599. [CrossRef] [PubMed]
22. Zhang, K.; Calabrese, P.; Nordborg, M.; Sun, F. Haplotype block structure and its applications to association studies: Power and study designs. *Am. J. Hum. Genet.* **2002**, *71*, 1386–1394. [CrossRef] [PubMed]
23. Matias, F.I.; Galli, G.; Correia Granato, I.S.; Fritsche-Neto, R. Genomic prediction of autogamous and allogamous plants by snps and haplotypes. *Crop Sci.* **2017**, *57*, 2951–2958. [CrossRef]
24. Bekele, W.A.; Wight, C.P.; Chao, S.; Howarth, C.J.; Tinker, N.A. Haplotype-based genotyping-by-sequencing in oat genome research. *Plant Biotechnol. J.* **2018**, *16*, 1452–1463. [CrossRef]
25. Jiang, Y.; Schmidt, R.H.; Reif, J.C. Haplotype-based genome-wide prediction models exploit local epistatic interactions among markers. *G3 (Bethesda)* **2018**, *8*, 1687–1699. [CrossRef]
26. He, J.; Meng, S.; Zhao, T.; Xing, G.; Yang, S.; Li, Y.; Guan, R.; Lu, J.; Wang, Y.; Xia, Q.; et al. An innovative procedure of genome-wide association analysis fits studies on germplasm population and plant breeding. *Theor. Appl. Genet.* **2017**, *130*, 2327–2343. [CrossRef]
27. Zhang, Y.; He, J.; Wang, H.; Meng, S.; Xing, G.; Li, Y.; Yang, S.; Zhao, J.; Zhao, T.; Gai, J. Detecting the qtl-allele system of seed oil traits using multi-locus genome-wide association analysis for population characterization and optimal cross prediction in soybean. *Front. Plant Sci.* **2018**, *9*, 1793. [CrossRef]
28. Khan, M.A.; Tong, F.; Wang, W.; He, J.; Zhao, T.; Gai, J. Using the rtm-gwas procedure to detect the drought tolerance qtl-allele system at the seedling stage under sand culture in a half-sib population of soybean [*glycine max* (L.) merr.]. *Can. J. Plant Sci.* **2019**, *99*, 801–814. [CrossRef]
29. Li, S.; Xu, H.; Yang, J.; Zhao, T. Dissecting the genetic architecture of seed protein and oil content in soybean from the yangtze and huaihe river valleys using multi-locus genome-wide association studies. *Int. J. Mol. Sci.* **2019**, *20*, 3041. [CrossRef]
30. You, F.M.; Booker, M.H.; Duguid, D.S.; Jia, G.; Cloutier, S. Accuracy of genomic selection in biparental populations of flax (*linum usitatissimum* L.). *Crop J.* **2016**, *4*, 290–303. [CrossRef]
31. You, F.M.; Xiao, J.; Li, P.; Yao, Z.; Jia, G.; He, L.; Kumar, S.; Soto-Cerda, B.; Duguid, S.D.; Booker, H.M.; et al. Genome-wide association study and selection signatures detect genomic regions associated with seed yield and oil quality in flax. *Int. J. Mol. Sci.* **2018**, *19*, 2303. [CrossRef]
32. Shim, H.; Chun, H.; Engelman, C.D.; Payseur, B.A. Genome-wide association studies using single-nucleotide polymorphisms versus haplotypes: An empirical comparison with data from the north american rheumatoid arthritis consortium. *BMC Proc.* **2009**, *3* (Suppl. 7), S35. [CrossRef]
33. Goddard, M. Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica* **2009**, *136*, 245–257. [CrossRef]

34. Li, Y.; Telfer, E.; Wilcox, P.L. New zealand forestry enters the genomics era—Applications of genomics in tree breeding. *N. Z. J. For.* **2015**, *60*, 23–25.
35. Dekkers, J.C. Prediction of response to marker-assisted and genomic selection using selection index theory. *J. Anim. Breed. Genet.* **2007**, *124*, 331–341. [CrossRef]
36. Guo, Z.; Wang, H.; Tao, J.; Ren, Y.; Xu, C.; Wu, K.; Zou, C.; Zhang, J.; Xu, Y. Development of multiple snp marker panels affordable to breeders through genotyping by target sequencing (gbts) in maize. *Mol. Breed.* **2019**, *39*, 37. [CrossRef]
37. Ali, O.A.; O'Rourke, S.M.; Amish, S.J.; Meek, M.H.; Luikart, G.; Jeffres, C.; Miller, M.R. Rad capture (rapture): Flexible and efficient sequence-based genotyping. *Genetics* **2016**, *202*, 389–400. [CrossRef]
38. Bekele, W.A.; Itaya, A.; Boyle, B.; Yan, W.; Mitchell Fetch, J.; Tinker, N.A. A targeted genotyping-by-sequencing tool (rapture) for genomics-assisted breeding in oat. *Theor. Appl. Genet.* **2019**, *133*, 653–664. [CrossRef]
39. Poland, J.A.; Rife, T.W. Genotyping-by-sequencing for plant breeding and genetics. *Crop Sci.* **2012**, *5*, 92–102. [CrossRef]
40. Kumar, S.; You, F.M.; Cloutier, S. Genome wide snp discovery in flax through next generation sequencing of reduced representation libraries. *BMC Genom.* **2012**, *13*, 684. [CrossRef]
41. Wang, Z.; Hobson, N.; Galindo, L.; Zhu, S.; Shi, D.; McDill, J.; Yang, L.; Hawkins, S.; Neutelings, G.; Datla, R.; et al. The genome of flax (*linum usitatissimum*) assembled *de novo* from short shotgun sequence reads. *Plant J.* **2012**, *72*, 461–473. [CrossRef]
42. You, F.M.; Deal, K.R.; Wang, J.; Britton, M.T.; Fass, J.N.; Lin, D.; Dandekar, A.M.; Leslie, C.A.; Aradhya, M.; Luo, M.C.; et al. Genome-wide snp discovery in walnut with an agsnp pipeline updated for snp discovery in allogamous organisms. *BMC Genom.* **2012**, *13*, 354. [CrossRef]
43. You, F.M.; Huo, N.; Deal, K.R.; Gu, Y.Q.; Luo, M.C.; McGuire, P.E.; Dvorak, J.; Anderson, O.D. Annotation-based genome-wide snp discovery in the large and complex *aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. *BMC Genom.* **2011**, *12*, 59. [CrossRef]
44. Browning, S.R.; Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **2007**, *81*, 1084–1097. [CrossRef]
45. You, F.M.; Xiao, J.; Li, P.; Yao, Z.; Jia, G.; He, L.; Zhu, T.; Luo, M.C.; Wang, X.; Deyholos, M.K.; et al. Chromosome-scale pseudomolecules refined by optical, physical and genetic maps in flax. *Plant J.* **2018**, *95*, 371–384. [CrossRef]
46. Lin, C.S.; Poushinsky, G. A modified augmented design (type 2) for rectangular plots. *Can. J. Plant Sci.* **1985**, *65*, 743–749. [CrossRef]
47. You, F.M.; Duguid, S.D.; Thambugala, D.; Cloutier, S. Statistical analysis and field evaluation of the type 2 modified augmented design (mad) in phenotyping of flax (*linum usitatissimum*) germplasms in multiple environments. *Aust. J. Crop Sci.* **2013**, *7*, 1789–1800.
48. Bradbury, P.J.; Zhang, Z.; Kroon, D.E.; Casstevens, T.M.; Ramdoss, Y.; Buckler, E.S. Tassel: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **2007**, *23*, 2633–2635. [CrossRef]
49. Gabriel, S.B.; Schaffner, S.F.; Nguyen, H.; Moore, J.M.; Roy, J.; Blumenstiel, B.; Higgins, J.; DeFelice, M.; Lochner, A.; Faggart, M.; et al. The structure of haplotype blocks in the human genome. *Science* **2002**, *296*, 2225–2229. [CrossRef]
50. Perez, P.; de los Campos, G. Genome-wide regression and prediction with the bglr statistical package. *Genetics* **2014**, *198*, 483–495. [CrossRef]
51. de Los Campos, G.; Hickey, J.M.; Pong-Wong, R.; Daetwyler, H.D.; Calus, M.P. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **2013**, *193*, 327–345. [CrossRef]
52. Lorenz, A.J.; Chao, S.; Asoro, F.G.; Heffner, E.L.; Hayashi, T.; Iwata, H.; Smith, K.P.; Sorrells, M.E.; Jannink, J.L. Genomic selection in plant breeding. In *Advances in Agronom*; Academic Press: Cambridge, MA, USA, 2011; Volume 110, pp. 77–123.
53. Endelman, J.B. Ridge regression and other kernels for genomic selection with r package rrblup. *Plant Genome* **2011**, *4*, 250–255. [CrossRef]

54. You, F.M.; Jia, G.; Cloutier, S.; Booker, H.M.; Duguid, S.D.; Rashid, K.Y. A method of estimating broad-sense heritability for quantitative traits in the type 2 modified augmented design. *J. Plant Breed. Crop Sci.* **2016**, *8*, 257–272.
55. Covarrubias-Pazaran, G. Genome-assisted prediction of quantitative traits using the r package sommer. *PLoS ONE* **2016**, *11*, e0156744. [CrossRef]

International Journal of
Molecular Sciences

MDPI

Article

Genetic and Methylome Variation in Turkish *Brachypodium Distachyon* Accessions Differentiate Two Geographically Distinct Subpopulations

Aleksandra Skalska [1], Christoph Stritt [2], Michele Wyler [2], Hefin W. Williams [3], Martin Vickers [4], Jiwan Han [5], Metin Tuna [6], Gulsemin Savas Tuna [7], Karolina Susek [8], Martin Swain [3], Rafał K. Wóycicki [9], Saurabh Chaudhary [10], Fiona Corke [11], John H. Doonan [11], Anne C. Roulin [2], Robert Hasterok [1,*] and Luis A. J. Mur [3,5,*]

[1] Plant Cytogenetics and Molecular Biology Group, Institute of Biology, Biotechnology and Environmental Protection, Faculty of Natural Sciences, University of Silesia in Katowice, 40–032 Katowice, Poland; askalska@us.edu.pl

[2] Department of Plant and Microbial Biology, University of Zürich, 8008 Zürich, Switzerland; christoph.stritt@uzh.ch (C.S.); michele.wyler@botinst.uzh.ch (M.W.); anne.roulin@botinst.uzh.ch (A.C.R.)

[3] Institute of Biological, Environmental and Rural Sciences (IBERS), Aberystwyth University, Aberystwyth SY23 3DA, UK; hew05@aber.ac.uk (H.W.W.); mts11@aber.ac.uk (M.S.)

[4] The John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK; martinj.vickers@gmail.com

[5] Shanxi Agricultural University, Taigu, Shanxi 030801, China; hanjiwan@sxau.edu.cn

[6] Department of Field Crops, Faculty of Agriculture, Tekirdag Namik Kemal University, Suleymanpasa 59030, Tekirdag, Turkey; mtuna@nku.edu.tr

[7] Tekirdag Anatolian High School, 59030 Suleymanpasa, Tekirdag, Turkey; glsvs@yahoo.com

[8] Department of Genomics, Institute of Plant Genetics, Polish Academy of Sciences, 60–479 Poznan, Poland; ksus@igr.poznan.pl

[9] Applied Omics—Rafał Wóycicki, 31–510 Kraków, Poland; rafal.woycicki@appliedomics.com

[10] School of Biosciences, Cardiff University, Cardiff CF10 3AX, UK; s.chaudhary84@outlook.com

[11] National Plant Phenomics Centre, Institute of Biological, Environmental and Rural Sciences (IBERS), Aberystwyth University, Aberystwyth SY23 3EB, UK; fic5@aber.ac.uk (F.C.); john.doonan@aber.ac.uk (J.H.D.)

* Correspondence: robert.hasterok@us.edu.pl (R.H.); lum@aber.ac.uk (L.A.J.M.)

Received: 12 July 2020; Accepted: 9 September 2020; Published: 13 September 2020

Abstract: *Brachypodium distachyon* (Brachypodium) is a non-domesticated model grass species that can be used to test if variation in genetic sequence or methylation are linked to environmental differences. To assess this, we collected seeds from 12 sites within five climatically distinct regions of Turkey. Seeds from each region were grown under standardized growth conditions in the UK to preserve methylated sequence variation. At six weeks following germination, leaves were sampled and assessed for genomic and DNA methylation variation. In a follow-up experiment, phenomic approaches were used to describe plant growth and drought responses. Genome sequencing and population structure analysis suggested three ancestral clusters across the Mediterranean, two of which were geographically separated in Turkey into coastal and central subpopulations. Phenotypic analyses showed that the coastal subpopulation tended to exhibit relatively delayed flowering and the central, increased drought tolerance as indicated by reduced yellowing. Genome-wide methylation analyses in GpC, CHG and CHH contexts also showed variation which aligned with the separation into coastal and central subpopulations. The climate niche modelling of both subpopulations showed a significant influence from the "Precipitation in the Driest Quarter" on the central subpopulation and "Temperature of the Coldest Month" on the coastal subpopulation. Our work demonstrates genetic diversity and variation in DNA methylation in Turkish accessions of Brachypodium that may be associated with climate variables and the molecular basis of which will feature in ongoing analyses.

Keywords: Brachypodium; DNA methylation; drought; flowering; genome; phenomics

1. Introduction

Brachypodium distachyon (hereafter Brachypodium) is a well-established grass model species, to be found mostly in countries bordering the Mediterranean. With its relatively small (~270 Mb) [1] nuclear genome it possesses most of the genomic and functional genomic infrastructure seen in *Arabidopsis thaliana* (hereafter Arabidopsis) [2,3]. The model has been developed to foster our understanding of such as grass cell wall biology e.g., [4] and flowering control e.g., [5]. Brachypodium has also been the subject of many studies into drought e.g., [6], salt and cold e.g., [7] and defining tolerance mechanisms of relevance to grasses and cereal crops.

Although ecologically important, until recently, the geographical variation of Brachypodium has been relatively poorly characterized [2]. Genome-wide variation in single nucleotide polymorphisms (SNPs) revealed a link to the geographical locations of accessions, with SNP variation suggesting a total of 15 genes being significantly related to environmental adaptation [8]. Genotyping by sequencing of >1400 accessions allowed the definition of a geographical split between the Western and Eastern Mediterranean populations but within each population there existed the same A and B subspecies. It was proposed that both subspecies re-colonized the Mediterranean basin after glaciation followed by lesser, allopatric genetic diversification [9]. Further information on variation was revealed from the pan-genome based on the resequencing of 54 inbred Brachypodium accessions. This study focused on existing inbred Turkish (T^+) and Spanish (S^+) accessions and identified 3,933,264 high-confidence SNPs. Phenotypically, the populations could be split between Extremely Delayed Flowering (EDF^+) phenotype, which was most common in the T^+ populations, and those where flowering was more rapid [10]. This genomic and phenotypic variation in the Turkish population was not associated with any precise geographical areas within Turkey.

Recently, increasing attention has focused on the possible contribution of DNA methylation, as a component of the epigenomic variation in response to environmental adaptation. DNA methylation forms distinct patterns on cytosines; 5′C-phosphate-G3′ (CpG), CHG, and CHH contexts (where H is any nucleotide except for G) [11] which together represent the methylome. Revealing such variations could identify features linked to the evolution of ecotypes [12]. This idea is supported by reports of stress-associated changes in the epigenome. For example, analyses of methylation sensitive amplified fragment length polymorphisms have suggested that whole methylome variation in plants correlates with environmental variables such as salt concentration [13] or the degree of plant isolation [14,15]. Now, with the widespread use of bisulfite sequencing (BS-Seq), finer scale mapping of the methylome is possible and responses to stress have been further suggested in, for example, transgenerational acquired resistance to disease [16].

Arabidopsis has proven to be especially useful in examining epigenetic variation related to the environment. A study of 150 Swedish Arabidopsis accessions demonstrated considerable epigenomic variation, particularly around transposable elements (TEs), when Arabidopsis was grown at 10 °C or 16 °C [17]. When the geographical origins of the accessions were considered, variation correlated with the relative degree of photosynthetically active radiation in spring and the strongest association was between CpG methylation and latitude [17]. The 1001 epigenome project assessed a global collection of Arabidopsis and further indicated that variation in methylation was related to geographical origin [18]. The methylome appeared to be shaped greatly by the genomic architecture of TEs which can influence the expression of nearby genes. Examining variably expressed genes indicated the prominence of genes linked to defense; for example, resistance genes, which were enriched in methylation in CpG and CHG and/or CHH contexts [18]. In segregating populations derived from an Arabidopsis Cvi × *Ler* cross, phenotypic differences, e.g., flowering time, were linked to patterns of DNA methylation [19]. Other analyses have linked differentially methylated regions with patterns of glucosinolate production [20]. In Brachypodium, the methylation patterns of seven inbred resequenced lines were found to correlate with the degree of genetic variation [21]. A recent more extensive assessment of DNA methylation was based on 83 inbred Turkish accessions and some invasive Australian accessions [22]. Considerable phenotypic variation was mostly correlated with SNP and DNA methylation patterns. There were some limited effects of CG methylation on certain phenotypic features [23].

In this current study, we adopted a different strategy to assess the variation in genomic sequences, DNA methylation and phenotypes in Brachypodium. Thus, we established a new bespoke collection of 55 Brachypodium accessions from Turkey, a center of Brachypodium diversification [3]. The sampling sites were selected to conform to the distinctive climatic regions of Turkey. Crucially, in order to maintain variation in DNA methylation, the seeds were used immediately in experimentation without inbreeding. This makes our Brachypodium collection unique. Seeds were transferred to the UK, and germinated under controlled environment conditions to avoid the introduction of variation in DNA methylation due to intergenerational changes [24]. We observed two major subpopulations in Turkey which could be distinguished based on variation in genome sequences and DNA methylation. Further, our study suggested that these subpopulations can be geographically separated into those from "coastal" and "central" regions. Physiologically, these subpopulations were distinguishable based on flowering requirements and relative drought tolerance as defined using phenomic approaches. This represents a foundational study based on which the nature of possible adaptive changes will be defined.

2. Results

2.1. Genomic Diversity Reveals Two Subpopulations in Turkey

To establish if genomic variation can be linked to climatic variables, a new collection of Brachypodium accessions was required where seed sampling reflected the different prevailing conditions. Köppen climate classifications can be used to divide Turkey into seven different climatic environments [25]. These were used to define our sampling strategy where 12 accessions were obtained from the regions designated 1a, 1c, 2, 3 and 4 (Supplementary Materials Figure S1). On advice from local collectors, region 1d was amalgamated with 3, as the collecting sites on the 3/1d border could not be precisely geographically defined. The collection sites are listed in Table 1. No Brachypodium accessions could be collected from region 1b as this was a high-altitude region where Brachypodium is not commonly found [26].

Following genomic sequencing of each accession and Bd21 as the canonical reference, the newly obtained data were joined with publicly available sequencing data for accessions from Spain and Turkey and individual accessions from France (ABR2) and Slovenia (ABR9) [10]. From the initial 8,556,181 hard-filtered SNPs identified among the 111 accessions, a set of 5,792 unlinked SNPs at synonymous positions were obtained for the characterization of genetic structure. Initial analyses compared the genetic variation in the Turkish accessions using a cluster analysis in SNPRelate package implemented in R, this indicated two main groups (Figure 1A). One branch contained all of the accessions from regions 3 and 4, but also some from regions 1c (1c_25_14, 1c_25_15, 1c_35_1, 1c_35.7). Conversely, the other group contained all the accessions from region 1a, some from 1c and included Bd21, which originated from Iraq and was therefore geographically close to the 1a/1c regions. Reflecting their geographical origins, we designated these groups as coastal (regions 2, 3 and 4) and central (regions 1a and 1c) subpopulations. Within the coastal subpopulation, we observed a separate clade of accessions 2_20_16, 2_14_15 and 2_15_20. Principal component analyses (PCA) with these SNPs shows that the two subpopulations of the newly collected accessions aligned with the T^{+} and the EDF^{+} previously defined by Gordon et al. [10] (Supplementary Materials Figure S2). Those accessions from the 1c region which were found within the coastal subpopulation (1c_25_14, 1c_25_15, 1c_35_1, 1c_35_7) were genotypically EDF^{+}. Surprisingly, accessions 2_20_16, 2_14_15 and 2_15_20 belonged to the S$^+$ cluster, which otherwise comprises only accessions from Spain and France (Supplementary Materials Figure S2). These various designations based on genetic, phenotypic and geographical variation are listed in Supplementary Materials Table S1.

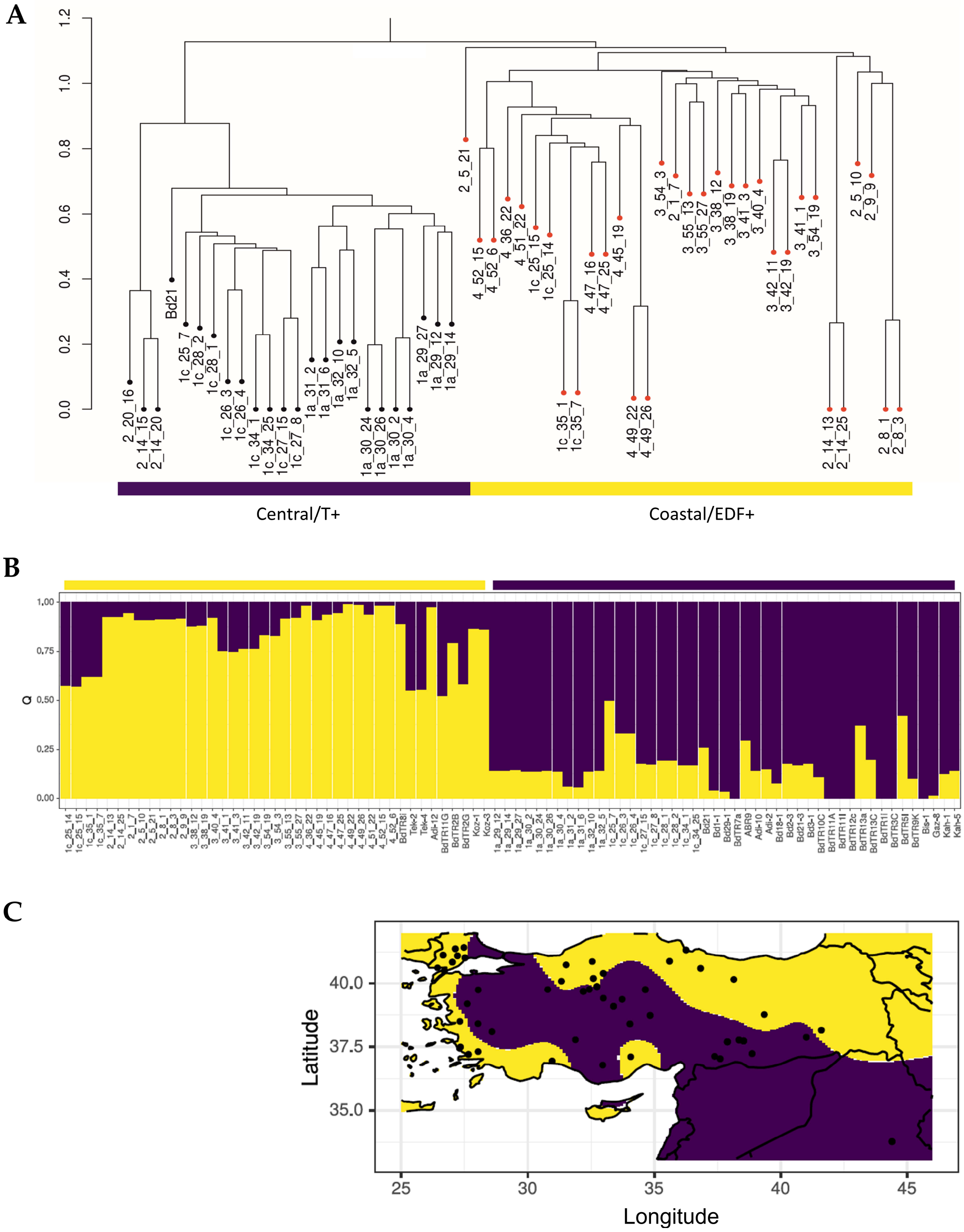


Figure 1. (**A**) Genetic diversity of Brachypodium germplasm from different environmental regions of Turkey as indicated using hierarchical clustering. (**B**) Ancestry coefficients were estimated with TESS3 two ancestral groups amongst Brachypodium accessions (listed in Supplementary Materials Table S1) and (**C**) mapped to Turkey by purple and yellow colors. The yellow and purple horizontal regions indicate accessions corresponding to the ancestral groups and geographically distinguishable as coastal and central subpopulations, respectively. The correspondence between the coastal and central subpopulations and the previously defined Extremely Delayed Flowering (EDF$^+$) and Turkish (T$^+$) populations [10], respectively, is indicated.

Table 1. Geographical origins of Brachypodium accessions used in this study (sorted by collecting date).

Region	Station	Collecting Date	Latitude	Longitude	Altitude (m) *
2	1	22-May-2016	38.5055	27.31671667	349
2	5	23-May-2016	37.49243333	27.3395	67
2	8	24-May-2016	37.20506667	27.65306667	40
2	9	24-May-2016	37.31233333	28.03705	626
2	14	25-May-2016	36.94201667	30.96305	10
2	20	27-May-2016	36.95461667	34.7507	161
1c	25	28-June-2016	39.86911667	32.7329	1042
1c	26	28-June-2016	39.68008333	32.19811667	879
1c	27	29-June-2016	38.40685	34.03873333	1122
1c	28	30-June-2016	38.738	34.83881667	1063
1a	29	1-July-2016	37.73385	38.53376667	668
1a	30	1-July-2016	37.69656667	37.89476667	696
1a	31	2-July-2016	37.03268333	37.60995	735
1a	32	2-July-2016	37.23601667	38.87008333	605
1c	34	3-July-2016	39.09433333	33.39311667	933
1c	35	3-July-2016	40.19286667	32.59326667	1059
4	36	3-July-2016	40.73106667	31.51755	865
3	38	7-July-2016	41.12035	26.65313333	58
3	40	23-July-2016	40.61361667	26.43273333	63
3	41	27-July-2016	41.0926	27.22096667	97
3	42	8-August-2016	41.3691	27.13661667	50
4	45	15-August-2016	40.86231667	32.54991667	1242
4	47	16-July-2016	40.87441667	35.60698333	605
4	49	16-July-2016	40.59275	36.83505	283
4	51	16-July-2016	40.15375	38.14713333	920
4	52	19-July-2016	41.32165	36.25826667	128
3	54	29-July-2016	40.83846667	27.02001667	205
3	55	29-July-2016	40.5003	26.70376667	88

Five distinct regions (1a, 1c, 2, 3, and 4) were selected for Brachypodium sampling defined by Köppen climate classifications [25]. There were at least five sampling sites ("stations") within each region and individual stations were sampled at least 12 times to derive individual accessions. Thus, for example, an accession designed 2_14_13 refers respectively to region, station, individual plant sample. *—meters above sea level.

In order to obtain a more detailed picture of the geographic distribution of the two genetic clusters present in Turkey, TESS3 was used to estimate ancestry components and project them onto a map (Figure 1B). Model fit improves as the number of ancestral populations (K) in the model increases, showing that population structure is strongly hierarchical (Supplementary Materials Figure S3). After K = 2, however, increase in model fit is marginal, indicating that a K of 2 describes the most important level of population subdivision. In agreement with the results obtained using PCA (Supplementary Materials Figure S2), at K = 2 one subpopulation corresponds to the EDF$^+$ cluster and contains all the accessions from the Köppen regions 3 and 4, but it should be noted some from regions 1c and 2 (Figure 1C, yellow). Conversely, the other subpopulation corresponds to the T$^+$ cluster and contains all accessions from region 1a, some from 1c and also Bd21 (Figure 1C, purple). Previous studies did not observe this geographic pattern because the coastal/ EDF$^+$ subpopulation was hugely under-represented: with only seven accessions of this subpopulation being sequenced before, compared to 27 of the central/ T$^+$ subpopulation [10].

2.2. *Whole Genome Methylation Assessments Also Indicate Two Subpopulations in the Turkish Population of Brachypodium*

We next assessed how the methylome could also vary across the Turkish regions that were sampled. DNA extracted from sample T_0 plant material was subjected to BS-Seq to reveal genome-wide cytosine methylation. The extent of methylation in different contexts across the population is given in Supplementary Materials Table S2. This indicated that CpG was the most common form of methylation (ranging between 54.3 and 67.8% of bases), followed by CHG (ranging between 28.4 and 43.7% of bases) and CHH (ranging between 1.3 and 10.2% of bases). Visualization of the variation in CpG methylation at a genomic level by hierarchical cluster analysis indicated geographic region-specific clustering in the Turkish accessions (Figure 2A). These clusters exhibited a significant geographic bias with the 1a southern group including the reference accession Bd21 from Iraq. Crucially, the major separation in CpG was between regions which represented the coastal and central subpopulations. With the other methylation contexts; CHG and CHH, also suggest epigenomic separation of the coastal and central subpopulations. This was most prominent in the CHG context compared to CHH (Figure 2B,C). Those EDF$^+$ genotypes from region 1c (1c_25_14, 1c_25_15, 1c_35_1, 1c_35_7) also exhibited methylomic variation which placed them in the same coastal subpopulations. Further, in all contexts, the S$^+$ accessions 2_20_16, 2_14_15 and 2_15_20 in our collection had distinctive features of their methylome compared to the other Turkish accessions. Therefore, genetic and methylomic variation in the Turkish accessions were closely aligned.

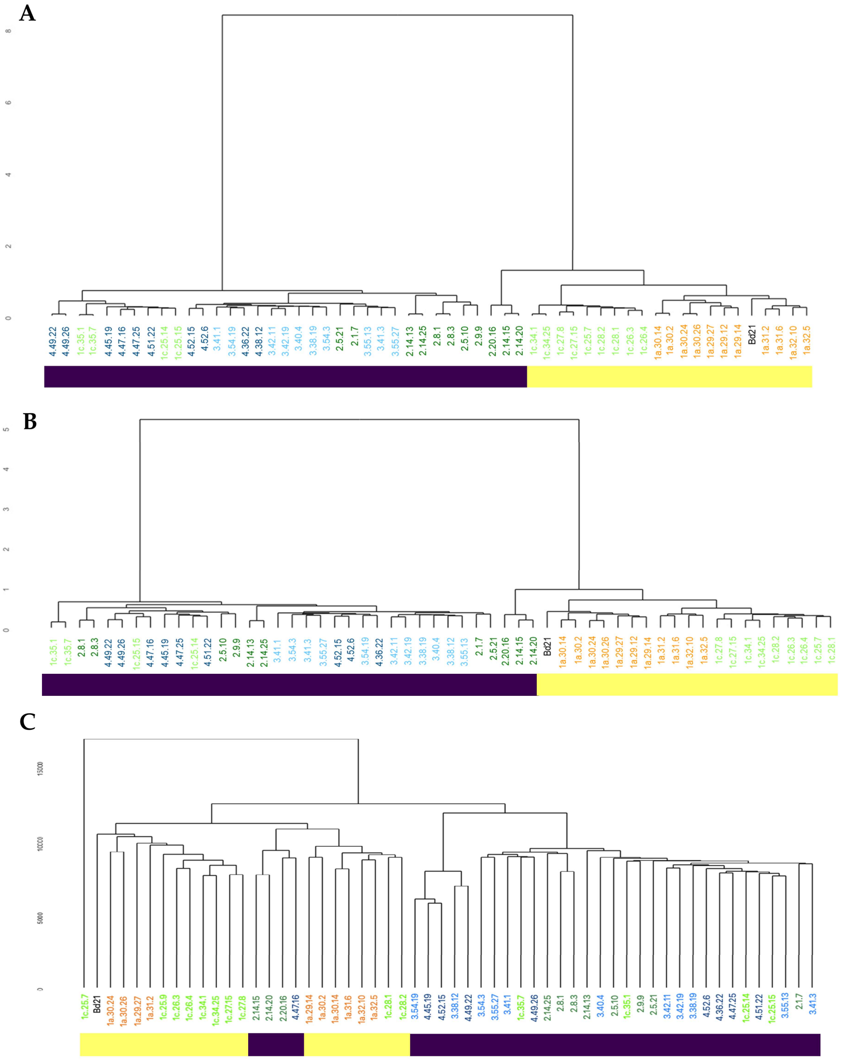


Figure 2. Hierarchical clustering analysis showing variation in whole genome (**A**) CpG, (**B**) CHG and (**C**) CHH methylation patterns based on the similarity of the accession's methylation profiles. Accessions from particular regions are color-coded (1a—orange, 1c—light green, 2—dark green, 3—light blue, 4—dark blue). Bd21 (from Iraq) is indicated in black. The yellow and purple horizontal bars indicate accessions broadly classified as central and coastal subpopulations, respectively. Region 1c accessions (1c_25_14, 1c_25_15, 1c_35_1, 1c_35_7) with an Extremely Delayed Flowering (EDF$^+$) [10] genotype are also located with the coastal subpopulation clade and are given that classification. Although classified as part of the coastal subpopulation the S$^+$ genotype accessions 2_14_15, 2_14_20, 2_20_16, were found in distinct clades in each methylation context; especially in CHH (**C**).

2.3. Phenomic Assessment of the Turkish Brachypodium Collection

Computerized image analysis approaches were employed to assess phenotypic variation in the sampling sites. Seeds (n = 8) from each accession were sown and at two weeks after sowing were vernalized at 4 °C for six weeks. After two weeks of growth at 22 °C (n = 4) plants of each accession were exposed to drought targeting 15% soil water content (SWC) over a period of 12 days. We had previously shown that this level of SWC was sufficient to impose drought stress on a diversity collection of Brachypodium accessions [6]. The remaining (n = 4) control plants were watered as normal. RGB images were obtained for plants and assessed for height and area as estimated from side view images (Supplementary Materials Figure S4) to provide a proxy for growth [27]. Phenotypic data (Supplementary Materials Table S3) obtained for individual accessions are provided in Supplementary Materials Figure S5). When accessions were considered based on regions, both plant height and side area were significantly reduced ($p < 0.05$) by drought treatment (Figure 3A,B). However, although there was considerable variation in accession height and area, no significant differences (p = 0.92) were observed between the different regions or the previously defined population groups.

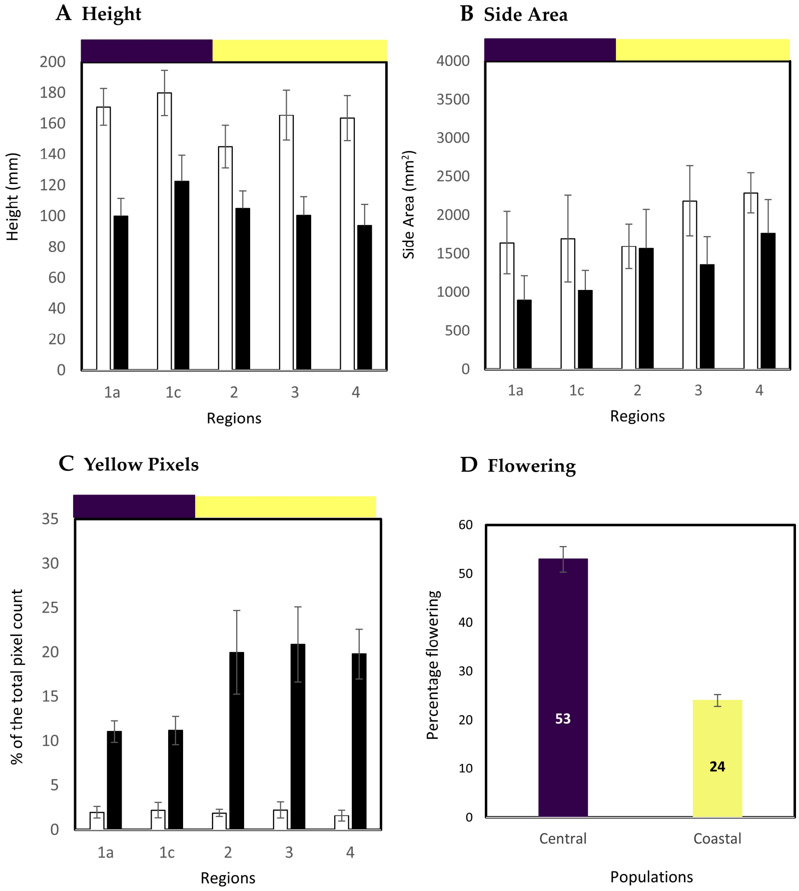


Figure 3. Phenotypic variation in the Turkish collection of Brachypodium. Brachypodium accessions (n = 8 plants) were vernalized for six weeks at 4 °C before being transferred to 22 °C and either maintained with full watering (n = 4 plants, white bars) or at 15% soil water content (n = 4 plants, black bars). At 12 d the plants were imaged at the National Plant Phenomics Centre, Aberystwyth, UK, where (**A**) height and (**B**) side area were derived. Data are grouped based on regional origins (1a, 1c, 2, 3, 4); (**C**) yellow pixels were extracted from the images of plants. Pixel data are presented as % of the total pixel count for the whole plant. The purple and yellow horizontal bars on (**A**–**C**) indicate data from accessions sampled from central (1a, 1c) and coastal (2, 3, 4) subpopulations, respectively; (**D**) after a further eight weeks the percentage of control plants originating from the coastal and central subpopulations which had flowered was measured.

Pixel colors were extracted from the images and the percentage of yellow pixels (indicative of stress associated leaf senescence) was significantly less in plants from region 1a and 1c as compared to plants from other regions ($p < 0.001$) (Figure 3C). In our previous publication we associated yellow pixel percentage with the extend of tolerance to stress [6]. This indicated a difference between the coastal and central subpopulation in terms of responses to drought stress with the central being more tolerant. The control plants were maintained for a further eight weeks after which accessions were assessed for flowering. Only 24% of accessions from the coastal subpopulation showed evidence of flowering (all from region 2) compared to 53% of the central subpopulation (Figure 3D). This aligned with the slower flowering EDF$^+$ phenotype which could be predicted to dominate the coastal subpopulation [10]. Too many of the plants that experienced drought stress subsequently died to allow the impact of stress on flowering to be determined.

2.4. Relating Climatic Niches to the Two Subpopulations in the Turkish Population of Brachypodium

Phenotypic, genetic and methylation analyses indicated that vernalization and drought tolerance could differentiate between the coastal and central subpopulation. Given this, we tested a series of environmental variables ("Bioclim") to see if they aligned with the distribution patterns of the subpopulations. This analysis was based on the derivation of Maxent [28] climate niche models. Model fitting using the two Bioclim variables identified a beta multiplier of "2" as producing the most parsimonious models and this was used as the setting for the comparison against the null models. Comparing the Maxent against the null models showed that the central region subpopulation had a median area under the curve (AUC) >98 and the coastal subpopulation >97. The Maxent model for the central region subpopulation was influenced solely by the response to the "Precipitation of the Driest Quarter" and the response curve (Figure 4A) clearly showed that the probability of presence is at its highest in areas of low precipitation in the study area. The variable Minimum Temperature of the Coldest Month does not have any influence on the model. Instead, this variable was specific to coastal areas around the Black and Mediterranean Seas (Figure 4B). The Maxent model response curves show that the highest probabilities were observed towards the higher end of Minimum Temperature during the Coldest month (Figure 4C) and at mid-ranges of precipitation during the driest quarter (Figure 4D,E). The metrics for niche similarity computed for the two subpopulations (Supplementary Materials Figure S6) indicate that there is no significant climate niche overlap, compared to the pseudoreplicates from the pooled location points from the two subpopulations. This conclusion is supported by "Schoener's D" ($p = 0.01$) and 'I' ($p = 0.01$).

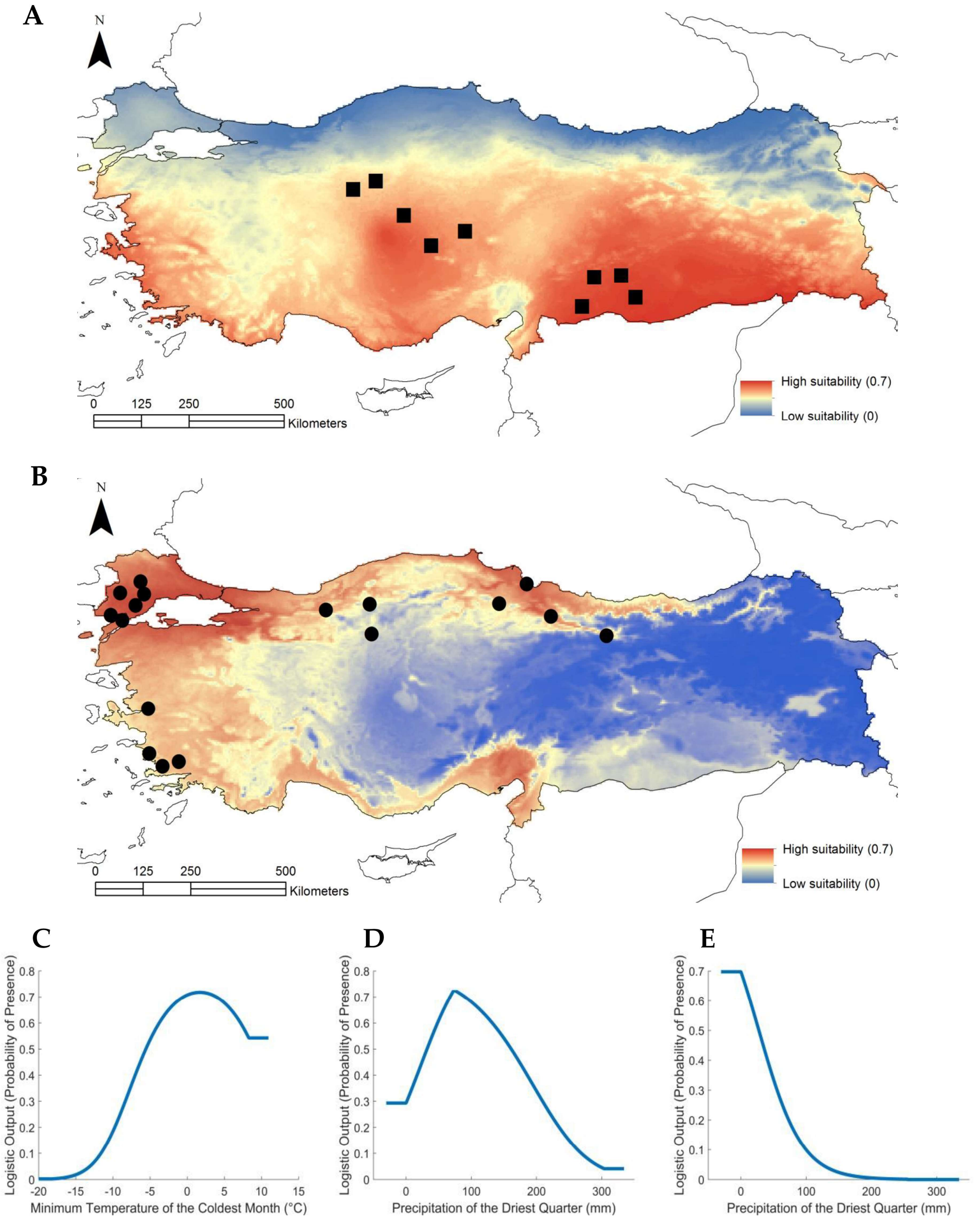


Figure 4. Bioclim modelling of environmental variables. These indicated the climate suitability of Brachypodium subpopulations in Turkey. (**A**) Color-coded, Maxent model climate suitability for the central subpopulation. Black squares represent sampling sites for the central subpopulation; (**B**) Maxent model climate suitability for the coastal subpopulation. Black circles represent sampling sites for the coastal subpopulation; (**C**) Maxent model response curve for the relationship between the probability of presence of the coastal subpopulation and the Minimum Temperature of the Coldest Month (°C); (**D**) Maxent model response curve for the relationship between the probability of presence of the coastal subpopulation and the Precipitation of the Driest Quarter (mm); (**E**) Maxent model response curve for the relationship between the probability of presence of the central subpopulation and the Precipitation of the Driest Quarter (mm).

3. Discussion

In this study we used Brachypodium to examine how genetic and methylation variation could reflect climatic variation across Turkey. The natural selection of particular alleles is the foundation of evolutionary thinking, but the potential selective role of epigenetic features also needs to be considered [29,30]. There is some evidence that in mangrove plants (*Laguncularia racemosa*) epigenetic patterns can emerge in the absence of major genetic variation and these could be maintained for at least 20 years [14]. Other plant species growing in the presence of stress conditions can exhibit epigenetic variation which is not reflected at the genetic level [31,32]. There are various mechanisms through which epigenetic changes can influence phenotype, for example through CpG methylation of the promoter regions to influence gene expression [19,33] or when differential methylation present around TEs alters gene expression [34]. It is also difficult to untangle the possible contributions of genetic from epigenetic variation to a given phenotype. While genome level genetic and methylation changes can be closely associated as seen in collections of wild Swedish accessions of Arabidopsis and this appears to be equally the case with Brachypodium [9,21].

A recent study described the derivation of a new inbred population from Turkey to assess SNP and methylome variation [23]. This work indicated the close association between genetic and methylation patterns across the population. Phenotypic variation was mostly linked to genetic variation, but some CpG methylation appeared to be associated with some additional effects in certain environments. In this current paper, we also examined the genetic and methylome variation but across a wild-collected population where we used climatic information to govern our sampling strategy and experimental approach. This led to a clear clustering in our sampling sites (Supplementary Materials Figure S1) which differed from the more equidistant sampling sites used by researchers who characterized the other Turkish populations [22,23]. Other studies indicate an East-West split in Brachypodium genetic diversity across the Mediterranean, which is likely to reflect separate refugia from the last ice age [10,35]. Further, within the eastern population in Turkey, two subpopulations have already been described; variously designated EDF$^+$ and T$^+$ [10] or Subspecies A East and B East [9]. Crucially, these differences were not linked to geography. Our assessment reveals additional potential drivers of diversity within climate-environmental regions of Turkey; leading to our definition of central vs coastal subpopulations. Thus, one subpopulation was predominant in regions 1a and 1c and as a result was designated as central (belonging to the T$^+$ cluster discussed above). In the other regions, 2, 3 and 4 the coastal (belonging to the EDF$^+$ cluster) subpopulation was predominant.

To preserve the methylome, seeds gathered from Turkish regions were assessed without undergoing generations of plant growth and meiosis in captivity. We did not use the inbred Turkish lines that are available e.g., [26] as epigenetic landmarks may have altered when propagated over many years [36]. We also used seeds directly sampled from Turkey but germinated in the UK under controlled environmental conditions. The germinated seedlings were used in our experiments in order to maintain, as much as possible, each accessions' genome methylation status. Our methylomic assessments showed a similar separation of the Turkish accessions into coastal and central subpopulations. This was particularly prominent for CpG and CHG contexts but was also observable with CHH. In plants, CpG methylation is maintained by the methyltransferase MET1. However, CHG and CHH are methylated by CHROMOMETHYLASE2 and 3 (CMT2, CMT3) and DNA (cytosine-5)-methyltransferase (DRM2) [37]. DRM2 interacts with its target through an RNA-directed DNA methylation (RdDM) pathway which employs 24-nucleotide small interfering RNAs (24nt-siRNAs). This mechanism requires de novo modification that needs constant targeting by RdDM [38]. Both CMT2 and CMT3 are guided to their targets by histone H3 lysine 9 (H3K9) methylation [39,40]. CpG methylation is predominant in heterochromatic regions with TEs, other repeats as well as coding regions but CHG and CHH methylation is almost only found in heterochromatin [41,42]. Therefore, these different means of establishing the methylome are employed to confer the methylomes patterns that we observed. Eichten et al. [21] also found that the patterns of methylation in accessions were broadly similar but in our case that appeared to be a geographical split. Therefore, the relative role of genetic vs methylation in driving phenotypic differences is unclear. Our results might reflect a role of epigenetic changes in adaptation (for review, see [43]) to the two-contrasting climate-environmental regions mentioned above. This separation might be reinforced by the fact that both subpopulations are likely to harbor distinct sets of transposable element polymorphisms [44], which are strongly methylated, like in many other plant species. Recent assessments of transposable elements in a wild collection of Brachypodium suggest that there is no great variation in copy number so this is unlikely to explain the differences in global methylation seen in our populations [45].

The major differentiation between the two subpopulations could reflect the impact of allopatric separation by a barrier such as a mountain range. Thus, the coastal vs central subpopulation could have arisen from the highland Anatolian plateau which roughly corresponds to the location of the central subpopulation, reducing gene flow. This could be reinforced through localized adaptation to such as stress driving increased tolerance. Equally, reduced gene flow can also arise from a shift in flowering times which would influence the frequency of cross-pollination [9]. To investigate possible phenotypic differences between the coastal and central subpopulations, we employed phenomic approaches to assess drought responses in a diverse collection of Spanish accessions [6]. Eichten et al. [23] used manual approaches to measure plant height, third leaf length and width, tiller count, ear count, and flowering time. Our image analysis measurement concentrated on plant height, width and flowering time but additionally considered yellow pixels percentages as this was an indicator of chlorophyll loss, this being a symptom of stress. The major split seen between coastal and central subpopulations was not reflected in any measured growth characteristic but there was a difference in flowering time after vernalization and relative drought tolerance. The former aligned with the EDF$^+$ phenotype which appeared to predominate in our coastal subpopulation although there was a significant proportion showing the rapid flowering type.

Flowering time in Brachypodium is strongly associated with vernalization period [9] and drought self-evidently with precipitation, so we tested how far the two subpopulation sampling sites could be associated with relevant climatic features. This involved testing Bioclim models which best explained the distribution of the coastal and central subpopulations. The statistically significant association of the coastal subpopulation with the "lowest temperature of the coldest month" variable agreed with the coastal accessions having the EDF$^+$ phenotype. Conversely, the distribution pattern of the central subpopulation was best explained by the "precipitation during driest quarter" Bioclim which would explain its greater degree of drought tolerance. These twin features could be major drivers of the marked genomic differences between the subpopulations. This stated the influence of flowering time as a barrier to gene flow in Brachypodium could be limited given its almost cleistogamous behavior which also results in a high degree of gene homozygosity [10]. Thus, the drought tolerance could play a much larger role in driving adaptive changes in the genome.

More detailed assessments are in progress; however, this current study highlights the importance of sampling strategies based on prevailing environmental conditions in order to better reveal differences between wild populations.

4. Materials and Methods

4.1. Derivation of Turkish Lines

Single seed inflorescences were sampled from 55 Brachypodium accessions from five distinct Turkish environments. The sites are described in Table 1. This (T_0) collection was transferred to Aberystwyth University, UK and three seeds from each accession were germinated under controlled environmental conditions (Levingtons F2 with horticultural grit [1/5 vol] added prior to use, 16 h photoperiod, natural light supplemented with artificial light from 400-W sodium lamps at 22 °C). After six weeks the first three leaves were collected from each accession and frozen in liquid N_2 prior to DNA extraction.

4.2. Whole Genome Sequencing and Single Nucleotide Polymorphism Calling

Genomic DNA was isolated from 10–15 mg of leaf tissue using the cetyl trimethylammonium bromide (CTAB) method [46]. Sequencing libraries were constructed with Illumina TruSeq Nano DNA kit and sequenced using Illumina X-Ten at 10× genome coverage (Macrogen Inc., Seoul, Republic of Korea; quality control (QC) data are provided in Supplementary Materials Table S4) For each accession sequenced paired-end reads were aligned using the BWA-MEM algorithm of Burrows-Wheeler Aligner (v.0.7) [47] to version 3.1 of the Brachypodium reference genome on Phytozome (https://phytozome.jgi.doe.gov/). After removing duplications with Sambamba (v.0.6.8) [48], SNPs were called with the Genome Analysis Toolkit (GATK, v.4.0.2.1) and filtered for quality scores lower than 20, a mean depth lower than 50 and a StrandOddsRatio higher than three. From the SNP set obtained synonymous SNPs were extracted through annotating to the reference *Brachypodium distachyon* synonymous positions in the genome (v.3.1), LD-pruned and filtered for minor allele frequency of 0.05. This generated a final data set of 5021 SNPs. To assess the genetic structure in 55 accessions, a phylogenetic tree was derived using HDClusters in the SNPrelate package (v.1.22.0) [49]. Population genetic and TESS3 structure analyses were performed with the tess3r package (v.0.1) implemented in R [50]. To obtain a wider depiction of genetic clustering, together with our individuals, using GATK (v.4.0.2.1) we combined whole genome sequences (vcf files) of accessions that were previously analyzed by Gordon et al. [10]. The merged file was later annotated to the synonymous position in the reference genome (annotation v.3.1), filtered and LD-pruned as described above. In total we obtained 5792 LD-pruned synonymous SNPs with no missing data. The cross-validation plot for the structure analysis was done using tess3r (v.01). All genomic data are available from https://www.ncbi.nlm.nih.gov/sra/PRJNA605320.

4.3. Bisulfite Sequencing and Data Analysis

Genomic DNA was isolated and BS-Seq libraries were constructed with Illumina TruSeq DNA methylation kit. After sequencing, poor quality reads and adapters were removed using TrimGalore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/, accessed 10 September 2020). Trimmed reads were mapped to the *B. distachyon* v.3.1 chloroplast and genomic reference sequences (QC, Supplementary Materials Table S5). After removing of duplicated reads with Bismark (v.1.3) [51], cytosine methylation in CpG, CHG and CHH contexts was estimated as an average over two duplicates for each accession (Supplementary Material Table S2). Epigenetic structure together with methylation and coverage statistics for each sample and context were performed using methylKit (v.1.8.1) [52] and GenomicRanges (v.1.34) within Bioconductor [53]. Conversion rates for whole-genome BS-Seq data in all three contexts are shown in Supplementary Materials Table S2.

4.4. Phenomic Experiments

To provide sufficient plants for the phenotyping experiments, seeds of second generation (T_1) were used. Seeds of the collected accessions were germinated in pots with 50 g of 4:1 Levington F2: grit sand. After two weeks, seedlings were vernalized for a further six weeks, and then the eight-week-old plants were transferred into the plant screening system (National Plant Phenomics Centre, NPPC, Aberystwyth, UK). The NPPC allows computer regulated watering of each individual plant and watering was withdrawn from four replicates from each genotype to achieve 15% soil water content (SWC) by seven days. This level of SWC was maintained for 12 days; the end of the experiment, the remaining replicates continued to be watered to 75% SWC. Images were captured at 12 days after watering was first restricted using a single-lens reflex camera Nikon D60 (Nikon Corporation, Tokyo, Japan) with an 18–55 mm lens. For uniform processing results and further analysis, images were processed to generate 24-bit RGB color images where each channel had 256 class color levels. Images were segmented from background in RGB color space. Plant growth parameters and color pixel data were extracted as plant height, top view and side view projection area and color information were extracted from the processed images using C++ (Visual Studio 2012) and Open Source Computer Vision Library (Open CV, v.2.4.9) [6]. Yellow pixel percentages as a proportion of total pixel numbers were calculated. Derived phenotype data were subjected to ANOVA using SPSS (v.25) software and residual plots were inspected to confirm normality of the distribution. Significance of differences between means was determined by contrast analysis (Scheffe's).

4.5. Climate Modelling

Potential differences between the climate niches of the different subpopulations by the multi-omic analyses were investigated by using the Identity test as described by Warren et al. [54]. This entailed creating Maxent [28] climate niche models for both subpopulations. We undertook two approaches for selecting predictor climate variables for the models. A PCA approach was undertaken to condense the variance from 19 Bioclim climate variables from the WorldClim climate dataset [55] into a smaller number of principal components. Having identified the final fitted models, we then tested the ability of the models to identify significant associations between the distribution of the subpopulations and the climate variables. This was undertaken by using a method proposed by Beale et al. [56,57] and developed by Williams et al. [58] where null models retaining the spatial structure of the presence points are used to compare against the real models.

Climate niche models that identified significant associations between the distribution of the subpopulations and the climate variables were run with ENMtools (v.1.3) [55] with the beta multiplier settings identified during the model fitting stage. One hundred pseudoreplicates were created for the Identity test from presence points of both subpopulations which yielded 100 values for 'I' and Schoener's D [54] to be compared against the values from the observed models. The hypothesis of niche identity was rejected if more than 95 of the pseudoreplicates had niche overlap values in excess of the niche overlap values from the observed subpopulations.

5. Conclusions

Defining variation amongst populations is important to determining likely evolutionary pressures shaping natural selection. In this study, we use the well-established model grass—Brachypodium—to define variation in Turkish populations. Crucially, our sampling strategy was biased towards representing the major climatic regions of Turkey. Variation was characterized at the genetic and methylome levels, to reflect two of the levels at which selection could act. Both genetic and methylome variation suggested two subpopulations which we designated as coastal and central. Phenotypic assessment suggested the subpopulations exhibited, respectively, a preponderance of differential flowering and drought tolerance phenotypes. This aligned with Bioclim models which suggested that late flowering was linked to the cold month—i.e., vernalization—and drought with relative precipitation in the driest month. Therefore, we provide evidence of climate being associated with genetic and methylome variation. Although Turkey has been extensively sampled by others, our environment bias sampling approach has provided important insights into potential drivers of Brachypodium evolution.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/21/18/6700/s1. Figure S1. Climatic regions of Turkey and Brachypodium sampling sites. Figure S2. Principal component analyses of SNP variation in the Brachypodium accessions. Figure S3. Ancestry coefficient assessment of structure analysis. Figure S4. Exemplar RGB image analysis of Brachypodium accessions in the National Plant Phenomics Centre, Aberystwyth, UK. Figure S5. Phenotypic variation in the Turkish collection of Brachypodium. Figure S6. Climate niche overlap metrics for Brachypodium subpopulations in Turkey. Table S1. Classification of Brachypodium accessions used in this study. Table S2. Conversion rates for whole genome bisulfite sequencing data in CpG, CHG and CHH context. Table S3. Phenotypic data generated at the National Plant Phenomic Centre (UK). Table S4. Whole genome sequencing quality check results. Table S5. Bisulfite sequencing quality check results.

Author Contributions: Conceptualization and study design—L.A.J.M. and R.H.; collecting Brachypodium accessions—M.T., G.S.T. and R.H.; investigation—A.S., C.S., M.W., H.W.W., M.V. and M.S.; formal analyses— all of the authors; supervision—K.S., A.C.R., R.H. and L.A.J.M.; preparing the original draft—A.S., A.C.R., R.H. and L.A.J.M.; reviewing and editing—all of the authors; funding acquisition, resources—J.H.D., A.C.R., R.H. and L.A.J.M.; project administration—R.H. and L.A.J.M. All authors contributed to the intellectual input and assistance to this study. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Centre Poland (grant no. 2015/18/M/NZ2/00394). Access to the National Plant Phenomics, Aberystwyth, UK was provided by the European Plant Phenotyping Network 2020 ("BRACHY-PHENO-DROUGHT" project, grant no. 25) and funded from the European Union's Horizon 2020 research and innovation program under grant agreement no. 731013. ACR and CS were supported by University Research Priority Program Evolution in Action of the University of Zurich. MW was supported by PSC Syngenta Fellowship.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. IBI (The International *Brachypodium* Initiative). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **2010**, *463*, 763–768. [CrossRef]
2. Mur, L.A.; Allainguillaume, J.; Catalan, P.; Hasterok, R.; Jenkins, G.; Lesniewska, K.; Thomas, I.; Vogel, J. Exploiting the *Brachypodium* tool box in cereal and grass research. *New Phytol.* **2011**, *191*, 334–347. [CrossRef] [PubMed]
3. Scholthof, K.B.G.; Irigoyen, S.; Catalan, P.; Mandadi, K.K. *Brachypodium*: A monocot grass model genus for plant biology. *Plant. Cell* **2018**, *30*, 1673–1694. [CrossRef] [PubMed]
4. Rancour, D.M.; Marita, J.M.; Hatfield, R.D. Cell wall composition throughout development for the model grass *Brachypodium distachyon*. *Front. Plant. Sci.* **2012**, *3*, 266. [CrossRef] [PubMed]
5. Woods, D.P.; Ream, T.S.; Bouche, F.; Lee, J.; Thrower, N.; Wilkerson, C.; Amasino, R.M. Establishment of a vernalization requirement in *Brachypodium distachyon* requires repressor of vernalization1. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 6623–6628. [CrossRef] [PubMed]
6. Fisher, L.H.; Han, J.; Corke, F.M.; Akinyemi, A.; Didion, T.; Nielsen, K.K.; Doonan, J.H.; Mur, L.A.; Bosch, M. Linking dynamic phenotyping with metabolite analysis to study natural variation in drought responses of *Brachypodium distachyon*. *Front. Plant. Sci.* **2016**, *7*, 1751. [CrossRef] [PubMed]
7. Priest, H.D.; Fox, S.E.; Rowley, E.R.; Murray, J.R.; Michael, T.P.; Mockler, T.C. Analysis of global gene expression in *Brachypodium distachyon* reveals extensive network plasticity in response to abiotic stress. *PLoS ONE* **2014**, *9*, e87499. [CrossRef]
8. Dell'Acqua, M.; Zuccolo, A.; Tuna, M.; Gianfranceschi, L.; Pe, M.E. Targeting environmental adaptation in the monocot model *Brachypodium distachyon*: A multi-faceted approach. *BMC Genom.* **2014**, *15*, 801. [CrossRef]
9. Wilson, P.; Streich, J.; Borevitz, J. Genomic diversity and climate adaptation in *Brachypodium*. *bioRxiv* **2015**. [CrossRef]
10. Gordon, S.P.; Contreras-Moreira, B.; Woods, D.P.; Des Marais, D.L.; Burgess, D.; Shu, S.; Stritt, C.; Roulin, A.C.; Schackwitz, W.; Tyler, L.; et al. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* **2017**, *8*, 2184. [CrossRef]
11. Finnegan, E.J.; Genger, R.K.; Peacock, W.J.; Dennis, E.S. DNA methylation in plants. *Annu. Rev. Plant. Physiol. Plant. Mol. Biol.* **1998**, *49*, 223–247. [CrossRef] [PubMed]
12. Mitchell-Olds, T.; Willis, J.H.; Goldstein, D.B. Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nat. Rev. Genet.* **2007**, *8*, 845–856. [CrossRef] [PubMed]
13. Foust, C.M.; Preite, V.; Schrey, A.W.; Alvarez, M.; Robertson, M.H.; Verhoeven, K.J.; Richards, C.L. Genetic and epigenetic differences associated with environmental gradients in replicate populations of two salt marsh perennials. *Mol. Ecol.* **2016**, *25*, 1639–1652. [CrossRef] [PubMed]
14. Herrera, C.M.; Bazaga, P. Genetic and epigenetic divergence between disturbed and undisturbed subpopulations of a Mediterranean shrub: A 20-year field experiment. *Ecol. Evol.* **2016**, *6*, 3832–3847. [CrossRef] [PubMed]
15. Herrera, C.M.; Medrano, M.; Bazaga, P. Comparative spatial genetics and epigenetics of plant populations: Heuristic value and a proof of concept. *Mol. Ecol.* **2016**, *25*, 1653–1664. [CrossRef] [PubMed]

16. Stassen, J.H.M.; Lopez, A.; Jain, R.; Pascual-Pardo, D.; Luna, E.; Smith, L.M.; Ton, J. The relationship between transgenerational acquired resistance and global DNA methylation in *Arabidopsis*. *Sci. Rep.* **2018**, *8*, 14761. [CrossRef]
17. Dubin, M.J.; Zhang, P.; Meng, D.; Remigereau, M.S.; Osborne, E.J.; Paolo Casale, F.; Drewe, P.; Kahles, A.; Jean, G.; Vilhjalmsson, B.; et al. DNA methylation in *Arabidopsis* has a genetic basis and shows evidence of local adaptation. *Elife* **2015**, *4*, e05255. [CrossRef]
18. Kawakatsu, T.; Huang, S.C.; Jupe, F.; Sasaki, E.; Schmitz, R.J.; Urich, M.A.; Castanon, R.; Nery, J.R.; Barragan, C.; He, Y.; et al. Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell* **2016**, *166*, 492–505. [CrossRef]
19. Schmid, M.W.; Heichinger, C.; Coman Schmid, D.; Guthorl, D.; Gagliardini, V.; Bruggmann, R.; Aluri, S.; Aquino, C.; Schmid, B.; Turnbull, L.A.; et al. Contribution of epigenetic variation to adaptation in *Arabidopsis*. *Nat. Commun.* **2018**, *9*, 4446. [CrossRef]
20. Aller, E.S.T.; Jagd, L.M.; Kliebenstein, D.J.; Burow, M. Comparison of the relative potential for epigenetic and genetic variation to contribute to trait stability. *G3* **2018**, *8*, 1733–1746. [CrossRef]
21. Eichten, S.R.; Stuart, T.; Srivastava, A.; Lister, R.; Borevitz, J.O. DNA methylation profiles of diverse *Brachypodium distachyon* align with underlying genetic diversity. *Genome Res.* **2016**, *26*, 1520–1531. [CrossRef] [PubMed]
22. Wilson, P.B.; Streich, J.C.; Murray, K.D.; Eichten, S.R.; Cheng, R.; Aitken, N.C.; Spokas, K.; Warthmann, N.; Gordon, S.P.; Vogel, J.P.; et al. Global diversity of the *Brachypodium* species complex as a resource for genome-wide association studies demonstrated for agronomic traits in response to climate. *Genetics* **2019**, *211*, 317–331. [CrossRef] [PubMed]
23. Eichten, S.R.; Srivastava, A.; Reddiex, A.J.; Ganguly, D.R.; Heussler, A.; Streich, J.C.; Wilson, P.B.; Borevitz, J.O. Extending the genotype in *Brachypodium* by including DNA methylation reveals a joint contribution with genetics on adaptive traits. *G3* **2020**, *10*, 1629–1637. [CrossRef]
24. Roessler, K.; Takuno, S.; Gaut, B.S. CG methylation covaries with differential gene expression between leaf and floral bud tissues of *Brachypodium distachyon*. *PLoS ONE* **2016**, *11*, e0150002. [CrossRef] [PubMed]
25. McKnight, T.L.; Hess, D. *Physical Geography: A Landscape Appreciation*, 7th ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2002.
26. Vogel, J.P.; Tuna, M.; Budak, H.; Huo, N.; Gu, Y.Q.; Steinwand, M.A. Development of SSR markers and analysis of diversity in Turkish populations of *Brachypodium distachyon*. *BMC Plant. Biol* **2009**, *9*, 88. [CrossRef] [PubMed]
27. Neilson, E.H.; Edwards, A.M.; Blomstedt, C.K.; Berger, B.; Moller, B.L.; Gleadow, R.M. Utilization of a high-throughput shoot imaging system to examine the dynamic phenotypic responses of a C4 cereal crop plant to nitrogen and water deficiency over time. *J. Exp. Bot.* **2015**, *66*, 1817–1832. [CrossRef]
28. Phillips, S.J.; Anderson, R.P.; Schapire, R.E. Maximum entropy modeling of species geographic distributions. *Ecol. Model.* **2006**, *190*, 231–259. [CrossRef]
29. Baulcombe, D.C.; Dean, C. Epigenetic regulation in plant responses to the environment. *Cold Spring Harb. Perspect Biol.* **2014**, *6*, a019471. [CrossRef]
30. Schmitz, R.J.; Ecker, J.R. Epigenetic and epigenomic variation in *Arabidopsis thaliana*. *Trends Plant. Sci.* **2012**, *17*, 149–154. [CrossRef]
31. Gao, L.; Geng, Y.; Li, B.; Chen, J.; Yang, J. Genome-wide DNA methylation alterations of *Alternanthera philoxeroides* in natural and manipulated habitats: Implications for epigenetic regulation of rapid responses to environmental fluctuation and phenotypic variation. *Plant. Cell Environ.* **2010**, *33*, 1820–1827. [CrossRef]
32. Lira-Medeiros, C.F.; Parisod, C.; Fernandes, R.A.; Mata, C.S.; Cardoso, M.A.; Ferreira, P.C. Epigenetic variation in mangrove plants occurring in contrasting natural environment. *PLoS ONE* **2010**, *5*, e10326. [CrossRef]
33. Schmitz, R.J.; Schultz, M.D.; Lewsey, M.G.; O'Malley, R.C.; Urich, M.A.; Libiger, O.; Schork, N.J.; Ecker, J.R. Transgenerational epigenetic instability is a source of novel methylation variants. *Science* **2011**, *334*, 369–373. [CrossRef] [PubMed]
34. Secco, D.; Wang, C.; Shou, H.; Schultz, M.D.; Chiarenza, S.; Nussaume, L.; Ecker, J.R.; Whelan, J.; Lister, R. Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements. *Elife* **2015**, *4*. [CrossRef] [PubMed]

35. Sharbel, T.F.; Haubold, B.; Mitchell-Olds, T. Genetic isolation by distance in *Arabidopsis thaliana*: Biogeography and postglacial colonization of Europe. *Mol. Ecol.* **2000**, *9*, 2109–2118. [CrossRef] [PubMed]
36. Slotkin, R.K.; Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* **2007**, *8*, 272–285. [CrossRef]
37. Zemach, A.; Kim, M.Y.; Hsieh, P.H.; Coleman-Derr, D.; Eshed-Williams, L.; Thao, K.; Harmer, S.L.; Zilberman, D. The *Arabidopsis* nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell* **2013**, *153*, 193–205. [CrossRef] [PubMed]
38. Matzke, M.A.; Mosher, R.A. RNA-directed DNA methylation: An epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **2014**, *15*, 394–408. [CrossRef]
39. Law, J.A.; Jacobsen, S.E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **2010**, *11*, 204–220. [CrossRef]
40. Stroud, H.; Do, T.; Du, J.; Zhong, X.; Feng, S.; Johnson, L.; Patel, D.J.; Jacobsen, S.E. Non-CG methylation patterns shape the epigenetic landscape in *Arabidopsis*. *Nat. Struct. Mol. Biol.* **2014**, *21*, 64–72. [CrossRef]
41. Cokus, S.J.; Feng, S.; Zhang, X.; Chen, Z.; Merriman, B.; Haudenschild, C.D.; Pradhan, S.; Nelson, S.F.; Pellegrini, M.; Jacobsen, S.E. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **2008**, *452*, 215–219. [CrossRef]
42. Lister, R.; O'Malley, R.C.; Tonti-Filippini, J.; Gregory, B.D.; Berry, C.C.; Millar, A.H.; Ecker, J.R. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **2008**, *133*, 523–536. [CrossRef] [PubMed]
43. Thiebaut, F.; Hemerly, A.S.; Ferreira, P.C.G. A role for epigenetic regulation in the adaptation and stress responses of non-model plants. *Front. Plant. Sci.* **2019**, *10*, 246. [CrossRef] [PubMed]
44. Stritt, C.; Gordon, S.P.; Wicker, T.; Vogel, J.P.; Roulin, A.C. Recent activity in expanding populations and purifying selection have shaped transposable element landscapes across natural accessions of the mediterranean grass *Brachypodium distachyon*. *Genome Biol. Evol.* **2018**, *10*, 304–318. [CrossRef] [PubMed]
45. Wyler, M.; Stritt, C.; Walser, J.-C.; Baroux, C.; Roulin, A.C. Impact of transposable elements on methylation and gene expression across natural accessions of *Brachypodium distachyon*. *Genome Biol. Evol.* **2020**, evaa180. [CrossRef] [PubMed]
46. Doyle, J. DNA protocols for plants. In *Molecular Techniques in Taxonomy*; Hewitt, G.M., Johnston, A.W.B., Young, J.P.W., Eds.; Springer: Berlin/Heidelberg, Germany, 1991; pp. 283–293.
47. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef]
48. Tarasov, A.; Vilella, A.J.; Cuppen, E.; Nijman, I.J.; Prins, P. Sambamba: Fast processing of NGS alignment formats. *Bioinformatics* **2015**, *31*, 2032–2034. [CrossRef]
49. Zheng, X.; Levine, D.; Shen, J.; Gogarten, S.M.; Laurie, C.; Weir, B.S. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **2012**, *28*, 3326–3328. [CrossRef]
50. Caye, K.; Deist, T.M.; Martins, H.; Michel, O.; François, O. TESS3: Fast inference of spatial population structure and genome scans for selection. *Mol. Ecol. Resour.* **2016**, *16*, 540–548. [CrossRef]
51. Krueger, F.; Andrews, S.R. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **2011**, *27*, 1571–1572. [CrossRef]
52. Akalin, A.; Kormaksson, M.; Li, S.; Garrett-Bakelman, F.E.; Figueroa, M.E.; Melnick, A.; Mason, C.E. methylKit: A comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **2012**, *13*, R87. [CrossRef]
53. Gentleman, R.C.; Carey, V.J.; Bates, D.M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5*, R80. [CrossRef] [PubMed]
54. Warren, D.L.; Glor, R.E.; Turelli, M. ENMTools: A toolbox for comparative studies of environmental niche models. *Ecography* **2010**, *33*, 607–611. [CrossRef]
55. Fick, S.E.; Hijmans, R.J. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **2017**, *37*, 4302–4315. [CrossRef]
56. Beale, C.M.; Brewer, M.J.; Lennon, J.J. A new statistical framework for the quantification of covariate associations with species distributions. *Methods Ecol. Evol.* **2014**, *5*, 421–432. [CrossRef]

57. Beale, C.M.; Lennon, J.J.; Gimona, A. Opening the climate envelope reveals no macroscale associations with climate in European birds. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 14908–14912. [CrossRef] [PubMed]
58. Williams, H.W.; Cross, D.E.; Crump, H.L.; Drost, C.J.; Thomas, C.J. Climate suitability for European ticks: Assessing species distribution models against null models and projection under AR5 climate. *Parasites Vectors* **2015**, *8*, 440. [CrossRef] [PubMed]

International Journal of
Molecular Sciences

MDPI

Article

Identification and Expression Analysis of Hormone Biosynthetic and Metabolism Genes in the 2OGD Family for Identifying Genes That May Be Involved in Tomato Fruit Ripening

Qiangqiang Ding [1,†], Feng Wang [1,†], Juan Xue [1], Xinxin Yang [1], Junmiao Fan [1], Hong Chen [2], Yi Li [3] and Han Wu [1,*]

[1] State Key Laboratory of Crop Genetics and Germplasm Enhancement, College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China; 2016204015@njau.edu.cn (Q.D.); wangf@njau.edu.cn (F.W.); 2017104059@njau.edu.cn (J.X.); 2018104055@njau.edu.cn (X.Y.); 2018204022@njau.edu.cn (J.F.)

[2] Jiangsu Key Laboratory for the Research and Utilization of Plant Resources, Institute of Botany, Jiangsu Province and Chinese Academy of Sciences, Nanjing 210014, China; ch198472@163.com

[3] Department of Plant Science and Landscape Architecture, University of Connecticut, Storrs, CT 06269, USA; yi.li@uconn.edu

* Correspondence: wuhan@njau.edu.cn

† These authors contributed equally to this work.

Received: 30 June 2020; Accepted: 23 July 2020; Published: 28 July 2020

Abstract: Phytohormones play important roles in modulating tomato fruit development and ripening. The 2-oxoglutarate-dependent dioxygenase (2OGD) superfamily containing several subfamilies involved in hormone biosynthesis and metabolism. In this study, we aimed to identify hormone biosynthesis and metabolism-related to 2OGD proteins in tomato and explored their roles in fruit development and ripening. We identified nine 2OGD protein subfamilies involved in hormone biosynthesis and metabolism, including the gibberellin (GA) biosynthetic protein families GA20ox and GA3ox, GA degradation protein families C19-GA2ox and C20-GA2ox, ethylene biosynthetic protein family ACO, auxin degradation protein family DAO, jasmonate hydroxylation protein family JOX, salicylic acid degradation protein family DMR6, and strigolactone biosynthetic protein family LBO. These genes were differentially expressed in different tomato organs. The GA degradation gene *SlGA2ox2*, and the auxin degradation gene *SlDAO1*, showed significantly increased expression from the mature-green to the breaker stage during tomato fruit ripening, accompanied by decreased endogenous GA and auxin, indicating that *SlGA2ox2* and *SlDAO1* were responsible for the reduced GA and auxin concentrations. Additionally, exogenous gibberellin 3 (GA_3) and indole-3-acetic acid (IAA) treatment of mature-green fruits delayed fruit ripening and increased the expression of *SlGA2ox2* and *SlDAO1*, respectively. Therefore, *SlGA2ox2* and *SlDAO1* are implicated in the degradation of GAs and auxin during tomato fruit ripening.

Keywords: genome-wide identification; expression analysis; 2OGD family; hormone biosynthetic and metabolism genes; tomato fruit ripening

1. Introduction

Tomato is used as a model to study climacteric fruit ripening, which is mediated by the hormone ethylene. Other hormones are also involved in tomato ripening. For example, exogenous auxin treatment, or increasing the endogenous auxin level by silencing the expression of the auxin-degradation gene *SlGH3.2*, delayed tomato fruit ripening [1–3]. Indeed, exogenous application of gibberellins

(GAs) delayed fruit ripening, while decreasing the endogenous levels of GAs via overexpression of the GA-catabolic gene *SlGA2ox1* accelerated fruit ripening [4]. Therefore, changes in hormone concentrations play important roles in tomato fruit ripening, and identification and functional analysis of hormone biosynthetic and metabolism genes are prerequisites for understanding their roles in tomato fruit ripening.

The 2-oxoglutarate-dependent dioxygenase (2OGD) superfamily is the largest enzyme family and facilitates numerous oxidative reactions, including hydroxylation, halogenation, desaturation, epimerization, etc. [5]. The 2OGD superfamily contains many proteins involved in hormone biosynthesis and metabolism. To date, nine hormone biosynthesis and metabolism-related protein families have been identified in the 2OGD family, including GA biosynthetic protein families GA20-oxidase (GA20ox) and GA3-oxidase (GA3ox), GA degradation protein families GA2-oxidases (C19-GA2ox and C20-GA2ox), auxin degradation protein family Dioxygenase for Auxin Oxidation (DAO), ethylene biosynthetic protein family 1-aminocyclopropane-1-carboxylic acid oxidase (ACO), jasmonate (JA) hydroxylation protein family JASMONATE-INDUCED OXYGENASE (JOX), salicylic acid (SA) degradation protein family Downy Mildew Resistant6 (DMR6) and DMR6-LIKE OXYGENASE (DLO), and strigolactone (SL) biosynthetic protein family LATERAL BRANCHING OXIDOREDUCTASE (LBO). In detail, GA20oxs and GA3oxs catalyze the final two steps of GA biosynthesis: GA20oxs catalyze the conversion of GA_{12} and GA_{53} to GA_9 and GA_{20}, which are converted by GA3oxs to bioactive GA_1 and GA_4 [6]. GA2oxs are GA-oxidation enzymes that convert bioactive GAs or their precursors into inactive forms [6]. DAOs catalyze the irreversible conversion of active auxin into inactive 2-oxindole-3-acetic acid (oxIAA) [7]. ACO proteins function in the last step of ethylene biosynthesis by converting ACC into ethylene [8]. JOX proteins hydroxylate jasmonate (JA) into inactive 12-OH-JA [9]. DMR6s, as SA 5-hydroxylases, hydroxylate active salicylate (SA) at the C5 position of the phenyl ring to produce inactive 2,5-DHBA [10]. In *Arabidopsis*, LBO converts methyl carlactonoate into an unidentified strigolactone (SL)-like compound that may be the final product of SL biosynthesis [11]. All of these 2OGD-family hormone biosynthetic and metabolism genes play key roles in maintaining endogenous hormone homeostasis, thereby regulating plant growth and development, and the response to stresses.

2OGDs are non-heme iron-containing proteins. Their catalytic core contains a double-stranded β-helix fold (DSBH) with a highly conserved His-X-Asp-$(X)_n$-His (HxD … H) motif, which is responsible for binding Fe (II) to form a catalytic triad [12]. Using Fe (II) as a cofactor and 2-oxoglutarate (2OG) as a co-substrate, 2OGD proteins catalyze oxidation of the substrate and concomitant decarboxylation of 2OG to produce succinate and CO_2. In addition, a conserved Arg-X-Ser/Thr (RxS/T) motif at the subfamily-conserved position within the secondary structure of the DSBH fold likely binds the C5-carboxy group of 2OG, which is the co-substrate for all known members of the subfamily except isopenicillin N synthase (IPNS), 1-aminocyclopropane-1-carboxylic acid oxidase (ACO) and (S)-2-hydroxypropylphosphonic acid epoxidase (HPPE) [13]. 2OGD-family proteins have been identified in several species [14]. The 2OGD superfamily can be divided into DOXA, DOXB, and DOXC subfamilies based on the amino acid sequence [14]. DOXA proteins contain a 2OG-FeII_Oxy_2 conserved domain, and the DOXA protein AlkB of *Escherichia coli*, which has homologs in *Arabidopsis* and rice, participates in the oxidative demethylation of alkylated nucleic acids and histones [15]. DOXB proteins typically have a conserved 2OG-FeII_Oxy_1 domain; most studies have focused on prolyl-4-hydroxylase, which is involved in the synthesis of cell-wall proteins in plants and algae [16]. DOXC proteins, including those involved in hormone biosynthesis and metabolism, have a conserved 2OG-FeII_Oxy domain [14].

In this study, we identified hormone biosynthesis- and metabolism-related proteins from DOXC family in tomato. Based on analysis of their structures, we predicted their motifs with the aim of determining their molecular mechanisms of action. We also analyzed the transcript levels of these genes in tomato to assess their roles in tomato growth and development, and focused on the correlations between their expression levels and tomato fruit ripening to identify proteins that degrade GAs and auxin during tomato fruit ripening.

2. Results

2.1. Identification and Phylogenetic Analysis of Hormone Biosynthetic and Metabolism Proteins in 2OGD Superfamily

Currently known hormone biosynthetic and metabolism proteins in the 2OGD superfamily are exclusively present in the DOXC subfamily. To identify all hormone biosynthetic and metabolism proteins of DOXC family in tomato, we used DOXC-specific 2OGD domain 2OG-FeII_Oxy (PF03171) as a key query in hmmersearch to identify all DOXC proteins in *Arabidopsis*, rice, and tomato. The result showed that 99, 90, and 159 proteins were identified in *Arabidopsis*, rice, and tomato, respectively. A phylogenic tree was constructed using the best-fit model in MEGA6.0, based on the complete sequences of the 348 identified proteins (Figure S1). Nine hormone biosynthetic and metabolism protein families in DOXC family were identified: the GA biosynthetic protein families GA20ox and GA3ox, GA degradation protein families C19-GA2ox and C20-GA2ox, auxin degradation protein family DAO, ethylene biosynthetic protein family ACO, JA hydroxylation protein family JOX, SA degradation protein family DMR, and SL biosynthetic protein family LBO. The bootstrap values were >80%, suggesting high reliability of the results. The numbers of these subfamilies in *Arabidopsis*, rice, and tomato were as follows: 20 GA20oxs, 10 GA3oxs, 20 C19-GA2oxs, 8 C20-GA2oxs, 20 ACOs, 6 DAOs, 11 JOXs, 9 DMR6s, and 3 LBOs. A phylogenetic tree constructed using the above proteins showed that there were 11 GA20oxs, 4 GA3oxs, 9 C19-GA2oxs, 3 C20-GA2oxs, 7 ACOs, 3 DAOs, 3 JOXs, 2 DMR6s, and 1 LBO in tomato, of which 10 GA20oxs (SlGA20ox1-SlGA20ox10), 6 GA2oxs (SlGA2ox2, SlGA2ox4, SlGA2ox5, SlGA2ox7, SlGA2ox8, and SlGA2ox9), 3 DAOs (SlDAO1-SlDAO3), and 5 ACOs (SlACO1-SlACO3, SlACO4, and SlACO6) clustered together to form a monophyletic group (Figure 1). Therefore, these genes emerged via lineage-specific expansion events in tomato. In addition, the identified hormone biosynthetic and metabolism proteins in tomato comprised 104–380 amino acids, and most of them also containing a DIOX_N domain (Table S1).

2.2. Synteny and Duplication Analysis of Hormone Biosynthetic and Metabolism Proteins in 2OGD Superfamily

Synteny was performed to assess the relationships of the hormone biosynthetic and metabolism 2OGD genes among *Arabidopsis*, rice, and tomato. The result showed that there were 27 collinear gene pairs, of which 25 were between tomato and *Arabidopsis*: 5 pairs in the *ACO* family, 2 in the *GA3ox* family, 4 in the *C19-GA2ox* family, 2 in the *C20-GA2ox* family, 6 in the *JOX* family, 4 in the *GA20ox* family, and 2 in the *DMR6* family. There was only one collinear gene pair in the *JOX* family between tomato and rice, as and one between rice and *Arabidopsis* (Figure 2, Table S2). This result is consistent with the evolutionary relationship between monocotyledons and dicotyledons.

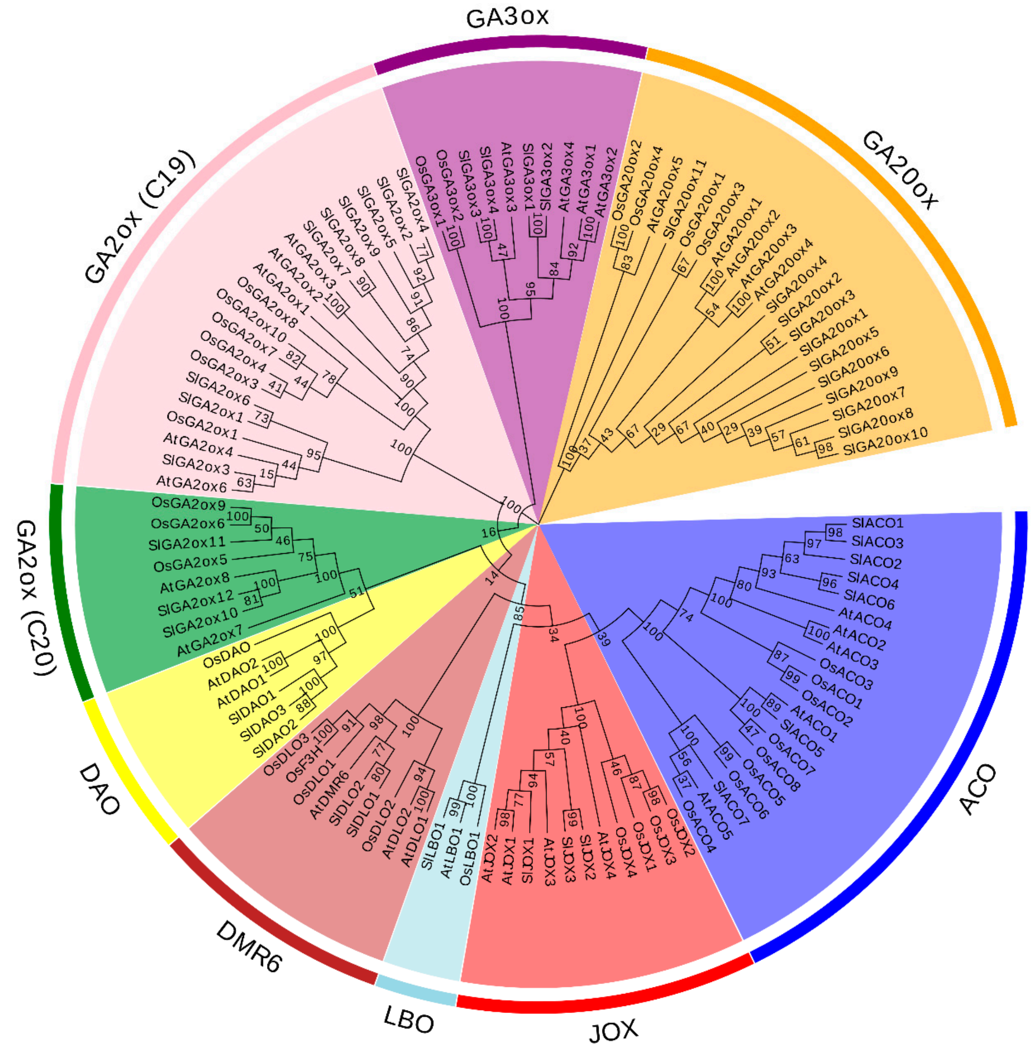


Figure 1. Phylogenetic tree of hormone biosynthetic and metabolism proteins of DOXC family in *Arabidopsis*, rice, and tomato. The phylogenetic tree was constructed by MEGA6 with Maximum likelihood.

The chromosomal location of the hormone biosynthetic and metabolism 2OGD genes in tomato was analyzed based on genome annotation data. The result showed that the identified hormone biosynthetic and metabolism 2OGD genes were unevenly distributed on tomato 12 chromosomes (Figure S2). There was one gene on chromosomes 4, 8, and 12, seven on chromosome 2, and six on chromosome 7. Further, the genes exhibited the following duplication events: nine dispersed gene pairs in *SlGA20ox* (Figure S2, Table S3); two segmental duplication genes (one WGD or segmental duplication events) and two dispersed gene pairs in *SlGA3ox*; seven dispersed gene pairs in *C19-SlGA2ox*; two segmental duplication genes (one WGD or segmental duplication events) in *C20-SlGA2ox*; two tandem duplication events in *SlDAO*; four segmental duplication genes (two WGD or segmental duplication events) in *SlACO*; two segmental duplication genes (one WGD or segmental duplication events) in *SlJOX*; two segmental duplication genes in *SlDMR6* (one WGD or segmental duplication events); and one dispersed gene pair in *SlLBO*.

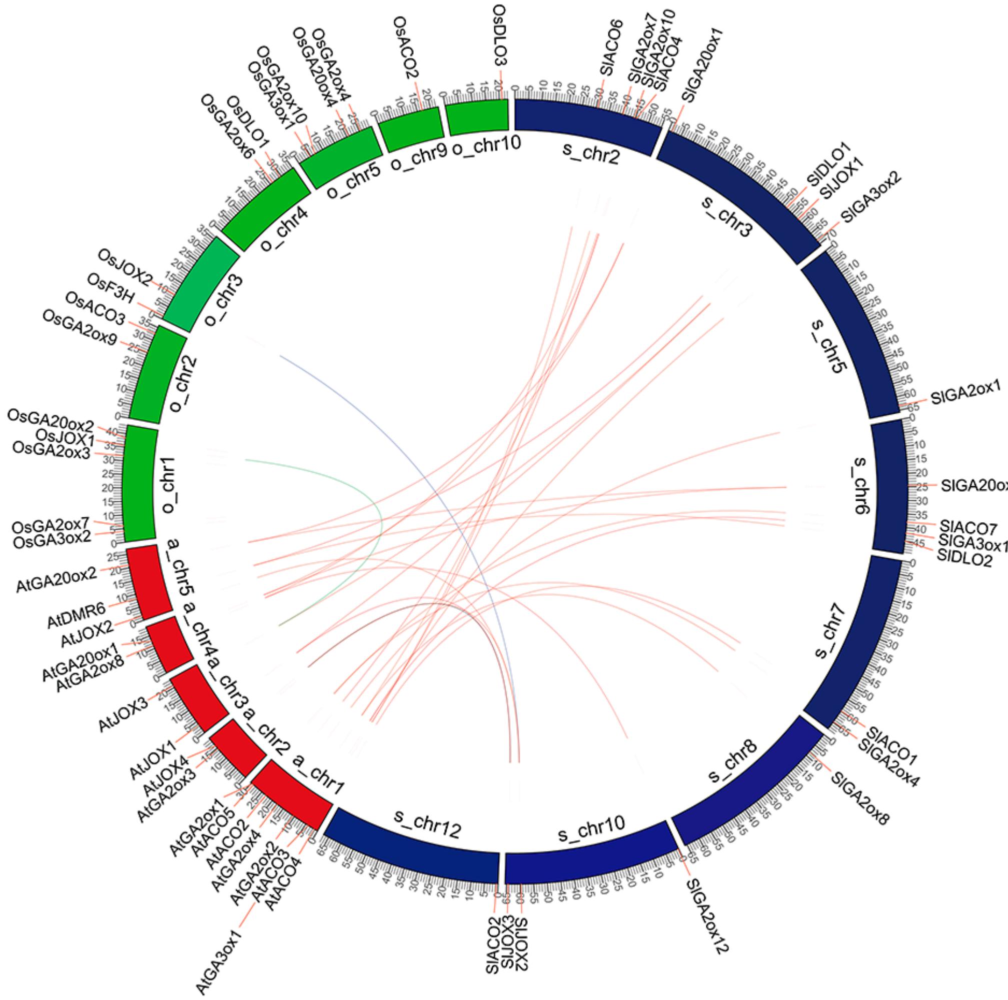


Figure 2. Synteny analysis of hormone biosynthetic and metabolism 2-oxoglutarate-dependent dioxygenase (2OGD) genes among *Arabidopsis*, rice, and tomato. Chromosome numbers of *Arabidopsis* (At), rice (Os), and tomato (Sl) are indicated on the inner side. Red, green, and blue colors represent *Arabidopsis*, rice, and tomato chromosomes. Gene pairs with a collinear relationship are joined by lines. Red lines represent collinear pairs between *Arabidopsis* and tomato, blue lines represent collinear pairs between *Arabidopsis* and rice, green lines represent collinear pairs between rice and tomato.

2.3. Multiple Sequence Alignment and Motif Composition Analysis of Hormone Biosynthetic and Metabolism 2OGD Proteins

To determine the functional similarity of hormone biosynthetic and metabolism 2OGD proteins of tomato with those of *Arabidopsis* and rice, we performed multiple sequence alignments and motif composition analysis. Two 2OGD-family proteins of known three-dimensional structure—OsGA2ox3 and OsDAO [17], and seven hormone biosynthetic and metabolism 2OGD-family proteins—AtGA20ox1, AtGA3ox1, AtGA2ox7, SlACO1, AtJOX1, AtDMR6, and AtLBO1—which have been functionally characterized were aligned to identify conserved domains or motifs in 2OGD family. The result showed that the above 2OGD proteins had the HxD … H and RxS/T conserved motifs in OsGA2ox3 and OsDAO (Figure 3a), which recruit Fe(II) as a cofactor and co-substrate. Further, among the hormone biosynthetic and metabolism 2OGD proteins in *Arabidopsis*, rice, and tomato, SlGA20ox8, SlGA20ox9, SlGA20ox10, SlGA2ox12, and OsACO6 did not have an HxD … H motif, while SlGA20ox7, SlGA20ox10, OsGA2ox10, SlGA2ox12, and OsACO6 lacked an RxS/T motif (Figures S3–S11), suggesting that these proteins do not have 2OGD biological activity.

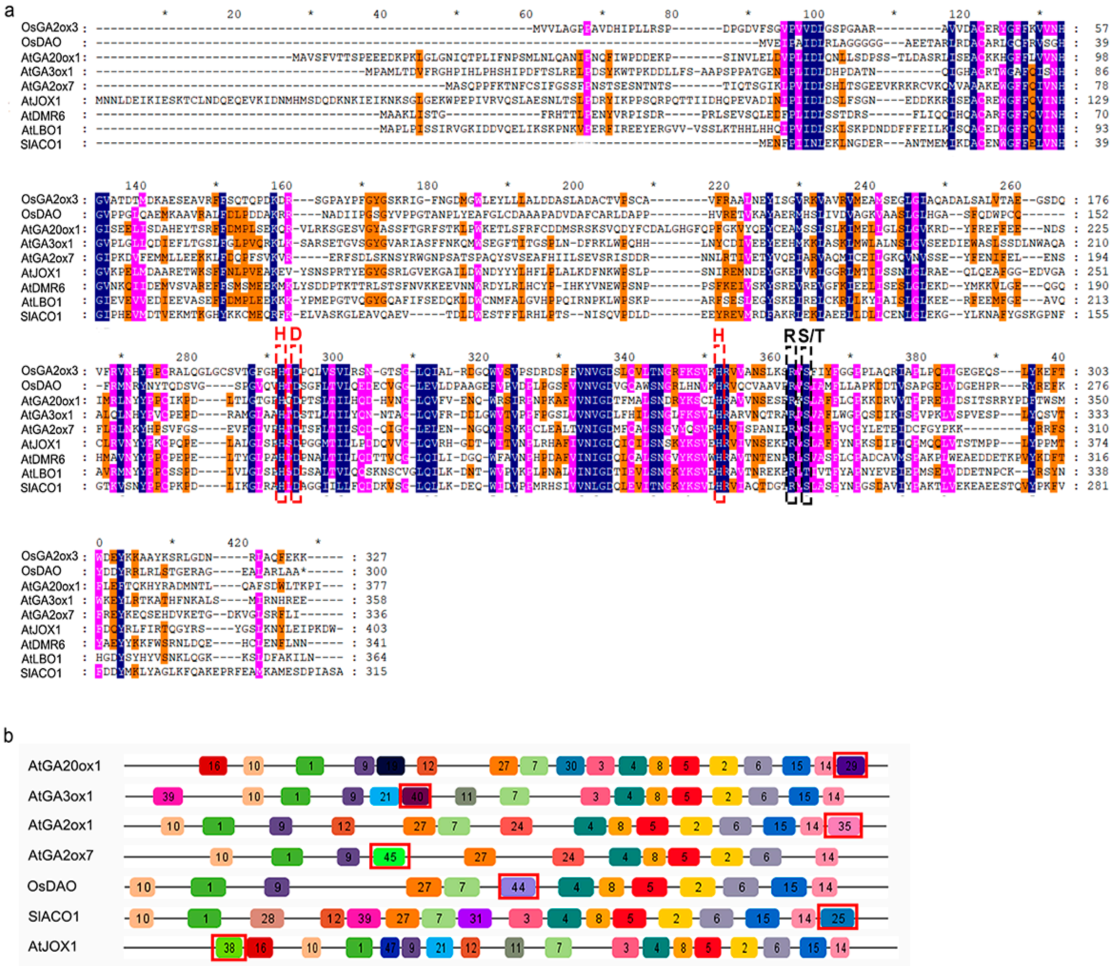


Figure 3. Sequence alignment and conserved motif analysis of functionally characterized hormone biosynthetic and metabolism 2OGD proteins. (**a**) Sequence alignment of functionally characterized hormone biosynthetic and metabolism 2OGD proteins in *Arabidopsis*, rice, tomato. The putative His-X-Asp-(X)$_n$-His (HxD . . . H) and Arg-X-Ser/Thr (RxS/T) motif locations are highlighted in red and black dotted boxes, respectively. (**b**) The motif composition of functionally characterized hormone biosynthetic and metabolism 2OGD proteins. The motif enclosed by red boxes is specific motifs in each group.

However, what is the difference of protein structure among different hormone biosynthetic and metabolism 2OGD protein families? Next, we used MEME to identify conserved motifs in DOXC-family proteins of *Arabidopsis*, rice, and tomato (Tables S4 and S5). The result showed that seven hormones biosynthetic and metabolism 2OGD protein families had uniquely conserved motifs—motifs 29, 40, 35, 45, 25, 44, and 38 were unique to the GA20ox, GA3ox, C19-GA2ox, C20-GA2ox, DAO, ACO, and JOX families, respectively (Figure 3b). However, no specific conserved motif was identified in the DMR6 or LBO families (Table S4). Further, sequence alignments showed that SlGA20ox7, SlGA20ox8, SlGA20ox9, and SlGA20ox10 did not have motif 29 (Figure S3), OsGA2ox10 did not have motif 35 (Figure S5), and OsACO4 did not have motif 25 (Figure S7), suggesting that these six proteins are not related to hormone biosynthesis or metabolism. In addition, SlGA2ox6 and SlGA2ox9 were truncated proteins with several missing amino acids in the N-terminal region (Figure S5). In conclusion, from the result of multiple sequence alignment and motif composition, the results suggesting that SlDAO1-SlDAO3, SlGA20ox1-SlGA20ox6, SlGA3ox1-SlGA3ox4, SlGA2ox1-SlGA2ox5, SlGA2ox7-SlGA2ox8, SlGA2ox10-SlGA2ox11, SlACO1-SlACO7, SlJOX1-SlJOX3, SlDLO1-SlDLO2, and SlLBO1 may have the ability of hormone biosynthesis and metabolism in tomato.

2.4. Expression of Hormone Biosynthetic and Metabolism 2OGD Genes in Tomato

To assess the function of identified hormone biosynthetic and metabolism 2OGD genes in tomato, we analyzed online transcriptome data of tomato roots, leaves, flowers, and developing fruits. Most genes exhibited distinct spatial and temporal expression patterns (Figure 4). Three *GA3ox* genes exhibited the highest expression in flowers, *SlGA3ox1* had moderate expression in roots and early developing fruits, and *SlGA3ox2* had moderate expression in leaves. No *GA3ox* gene was expressed during fruit ripening (Figure 4a). Regarding the *GA20ox* family, *SlGA20ox1*, *SlGA20ox2*, and *SlGA20ox3* were highly expressed in flowers and early developing fruits; *SlGA20ox1* and *SlGA20ox3* were also expressed in roots, and *SlGA20ox1* and *SlGA20ox2* were expressed in leaves. *SlGA20ox4* was specifically expressed in unopened flowers. Only *SlGA20ox3* was expressed during fruit ripening, during which its expression increased continuously (Figure 4b). Five GA2ox-family genes (*SlGA2ox3*, *4*, *5*, *7*, and *10*) showed high expression in roots, three (*SlGA2ox2*, *3*, and *10*) in leaves, and six (*SlGA2ox1*, *2*, *4*, *5*, *7*, and *10*) in flowers. In addition, four genes (*SlGA2ox2*, *4*, *5*, and *7*) had high expression in early developing fruits, which increased during fruit ripening (from the mature-green stage to the breaker stage) (Figure 4c). Among the *DAO* family, the expression of *SlDAO1* was high in ripening fruits, moderate in early fruits, and low in roots, leaves, and flowers. *SlDAO2* was expressed mainly in flowers and early fruits, while the expression of *SlDAO3* was negligible in all organs. Notably, *SlDAO1* expression increased significantly from the mature-green to the breaker stage, suggesting a role in fruit ripening (Figure 4d). The expression of the three *JOX*-family genes was highest in flowers, while that of *SlJOX1* and *SlJOX2* was moderate in roots, leaves, and early developing fruits, and *SlJOX2* was expressed in breaker fruits (Figure 4e). Regarding the *ACO* family, three genes (*SlACO2*, *3*, and *4*) were expressed in roots, two (*SlACO4* and *5*) in leaves, and five (*SlACO1*, *2*, *3*, *4*, and *6*) in flowers. Further, four genes (*SlACO1*, *3*, *4*, and *6*) had high expression in early developing fruits, and the expression of four other genes (*SlACO1*, *3*, *5*, and *6*) increased from mature-green to breaker fruit (Figure 4f). The *DLO*-family gene *SlDLO1* showed high expression in roots, leaves, flowers, and early fruits, and decreased expression in ripening fruits, while *SlDLO2* was expressed only in flowers and early fruits (Figure 4g). The only *LBO* gene in tomato, *SlLBO1,* was expressed mainly in roots and flowers, suggesting roles in root and flower development (Figure 4h). In conclusion, a variety of 2OGD hormone biosynthetic and metabolism genes play roles in organ development and fruit ripening in tomato.

2.5. Expression of SlGA2ox and SlDAO Genes during Tomato Fruit Ripening

Ethylene is the major hormone regulating tomato fruit ripening, while auxin and GAs regulate fruit ripening via the ethylene pathway [2–4]. The endogenous auxin and GA concentration was decreased during tomato fruit ripening (Figure S12) [3,4], so we investigated the roles of auxin- and GA-degradation genes on tomato fruit ripening. Tomato pericarps at four stages (mature-green, breaker, yellow-ripening, and red-ripening) were collected from the tomato cultivars ‘Ai Ji Qiao Li’ and ‘Micro-Tom’ for qPCR analysis (Figure 5a). The *SlDAO1* expression level was higher than that of *SlDAO2* in Ai Ji Qiao Li and Micro-Tom during fruit ripening (Figure 5b,d). Notably, the expression of *SlDAO1* significantly increased, about two-fold, in Ai Ji Qiao Li, and tenfold in Micro-Tom from the mature-green to the breaker stage; its expression level remained elevated in the yellow- and red-ripening stages. However, *SlDAO2* expression did not significantly change from the mature-green to the breaker stage, and remained very low in the yellow- and red-ripening stages (Figure 5b,d). Thus, *SlDAO1*, rather than *SlDAO2*, likely plays a role in the transition from the mature-green to the breaker stage and subsequent fruit ripening. In addition, the expression of *SlGA2ox2* was 100-fold higher than that of *SlGA2ox4* and *SlGA2ox5*, while *SlGA2ox4* and *SlGA2ox5* expression was negligible in Ai Ji Qiao Li and Micro-Tom (Figure 5c,e). *SlGA2ox2* expression was increased threefold in Ai Ji Qiao Li and thirty-fold in Micro-Tom from the mature-green stage to the breaker stage, and decreased slightly in the yellow- and red-ripening stages (Figure 5c,e); this suggested that *SlGA2ox2* participates in tomato fruit ripening.

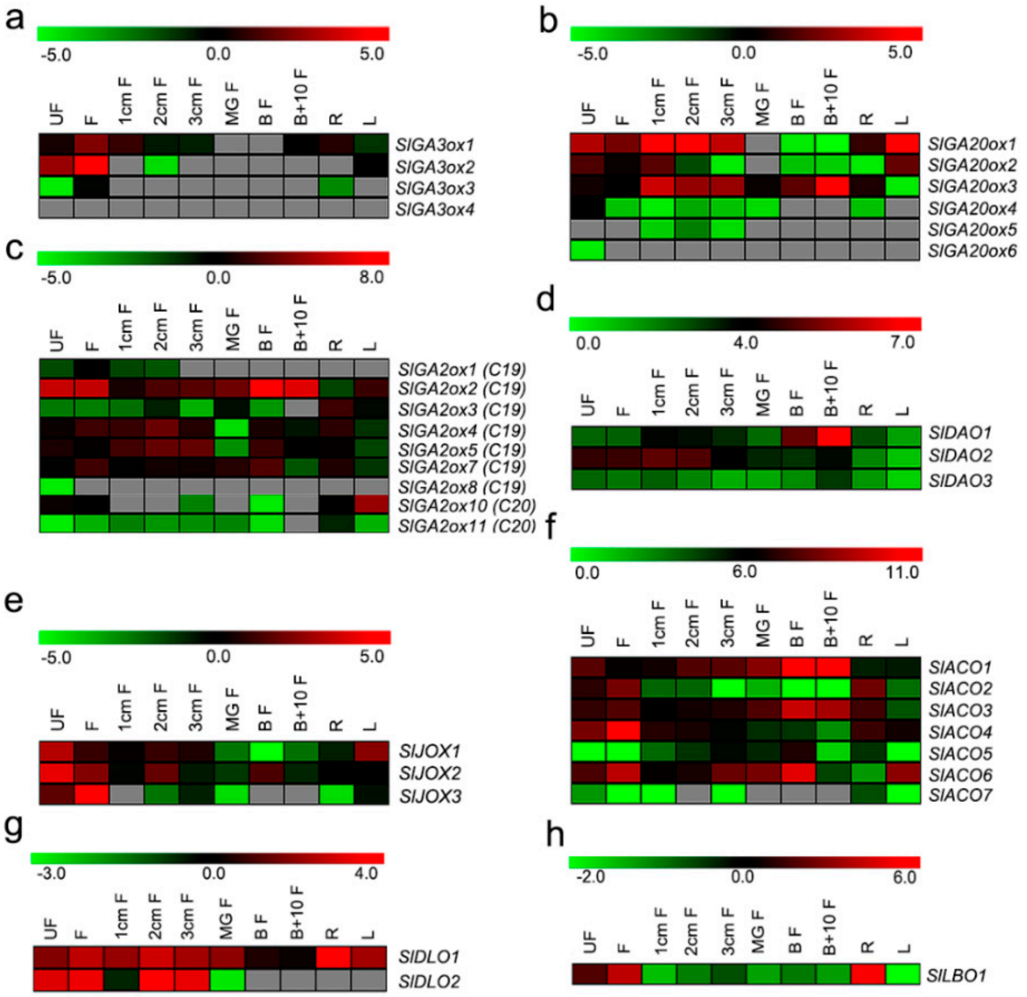


Figure 4. Expression pattern of hormone biosynthetic and metabolism 2OGD genes in tomato. (**a**–**h**) Expression pattern of *SlGA3ox*, *SlGA20ox*, *SlGA2ox*, *SlDAO*, *SlJOX*, *SlACO*, *SlDLO*, and *SlLBO* group genes. Gray boxes represent the expression of genes was undetectable. Unopened flowers (UF); Opened flowers (F); 1 cm fruits (1 cm F); 2 cm fruits (2 cm F); 3 cm fruits (3 cm F); mature-green fruits (Mg F); breaker fruits (Br F); breaker+10 days' fruits (Br+10 F); roots (R); leaves (L). The detailed descriptions of the stages and tissues were on the website (http://ted.bti.cornell.edu/cgi-bin/TFGD/digital/home.cgi).

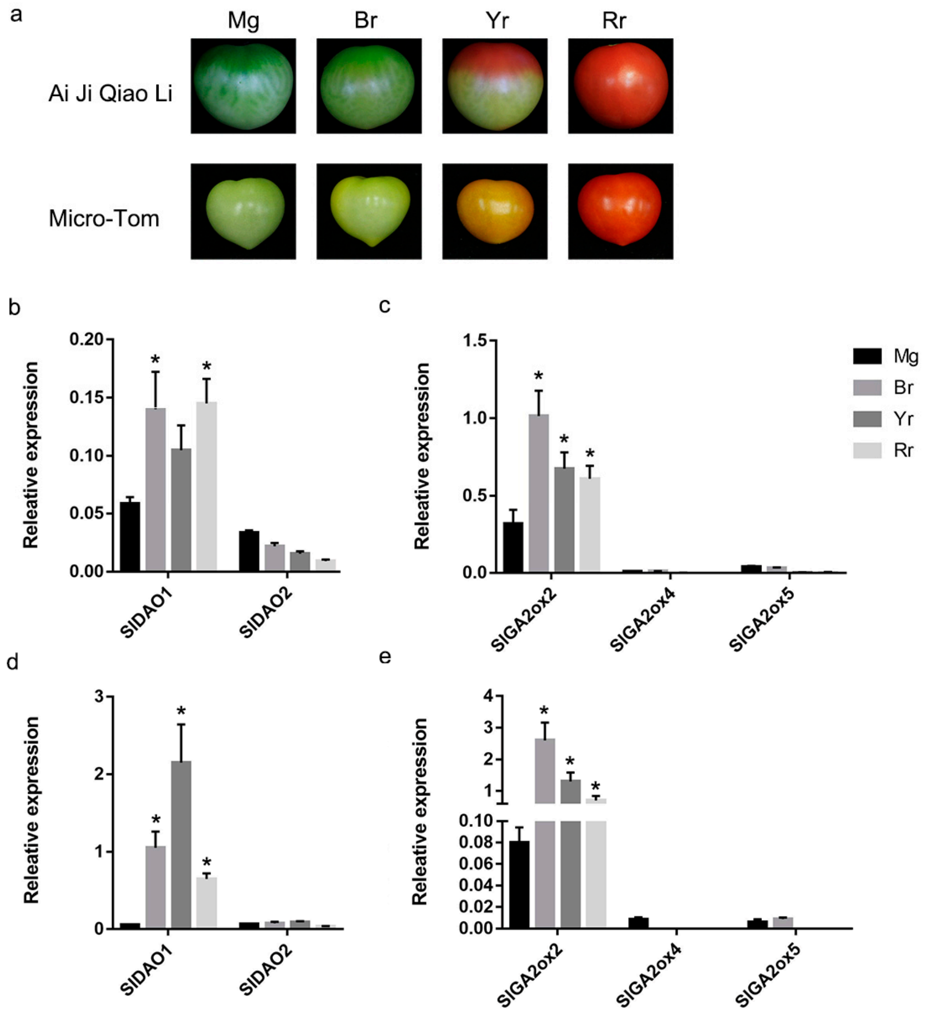


Figure 5. Expression analysis of *SlDAOs* and *SlGA2oxs* genes during tomato fruit ripening in the pericarp. (**a**) Different ripening stages of Ai Ji Qiao Li and Micro-Tom. (**b**) Expression levels of *SlDAOs* in Ai Ji Qiao Li. (**c**) Expression levels of *SlGA2ox* genes in Ai Ji Qi Li. (**d**) Expression levels of *SlDAOs* in Micro-Tom. (**e**) Expression levels of *SlGA2ox* genes in Micro-Tom. Mg: mature-green; Br: breaker; Yr: yellow-ripening; Rr: red-ripening. * The asterisk at the top of each column indicates a significant difference compared to Mg fruits at $p < 0.05$ ($n = 3$) by students t-test.

2.6. Effects of Auxin, GA$_3$, and Ethylene on the Expression of SlDAO1, SlDAO2, and SlGA2ox2

To study the response of *SlDAO1*, *SlDAO2*, and *SlGA2ox2* to auxin, GAs, and ethylene, we treated Micro-Tom mature-green fruits with IAA, GA$_3$, and ethylene, and analyzed their expression after 2 and 4 days. Consistent with previous reports, IAA and GA$_3$ delayed tomato fruit ripening (Figure 6a). Further, the expression of *SlDAO1* was significantly induced by IAA, but was unaffected by GA$_3$ and ethylene at 2 and 4 days, while *SlDAO2* expression was not significantly affected in auxin-, GA-, or ethylene-treated mature-green fruits (Figure 6b). In addition, *SlGA2ox2* showed higher expression in GA$_3$-treated fruits, but similar expression in IAA- and ethylene-treated fruits, compared to the control (Figure 6b). In conclusion, the expression of *SlDAO1* and *SlGA2ox2* was induced by auxin and GAs, respectively, suggesting that *SlDAO1* and *SlGA2ox2* are responsible for regulating auxin and GA catabolism during tomato fruit ripening.

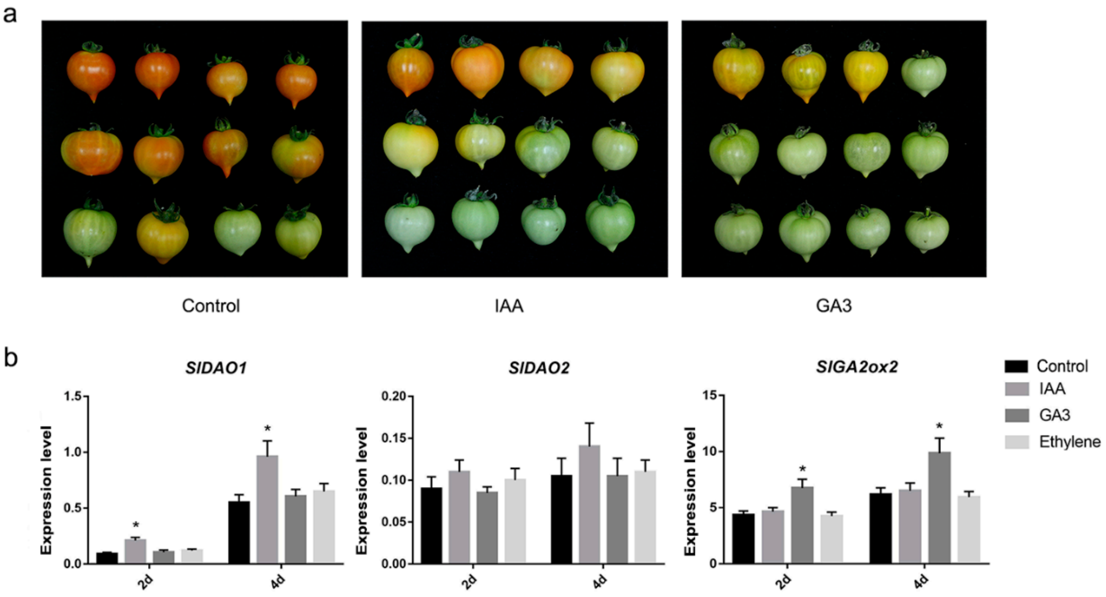


Figure 6. Expression analysis of *SlDAO1*, *SlDAO2*, and *SlGA2ox2* after auxin, GA_3, and ethylene treatments. (**a**) Photos of mature-green fruits after indole-3-acetic acid (IAA) and gibberellin 3 (GA_3) treatment, respectively. (**b**) Expression analysis of *SlDAO1, SlDAO2*, and *SlGA2ox2* after auxin, GA_3, and ethylene treatments. * The asterisk at the top of each column indicates a significant difference at $p < 0.05$ ($n = 3$) by students t-test.

3. Discussion

3.1. Identification of Hormone Biosynthetic and Metabolism Genes from 2OGD Family

The 2OGD superfamily is widespread in microorganisms, fungi, mammals, and plants. In plants, 2OGD proteins are classified as DOXA, DOXB, and DOXC [14]. DOXA proteins are involved in the oxidative demethylation of alkylated nucleic acids and histones, while DOXB proteins are involved in proline 4-hydroxylation in cell-wall protein synthesis, and DOXC proteins in the metabolism of various phytochemicals, such as phytohormones and flavonoids. The number of 2OGDs of the DOXA and DOXB classes is constant across plant species, whereas that of the DOXC class is extremely variable, suggesting that the latter has diversified during the evolution of land plants. The vast majority of 2OGDs from land plants are of the DOXC class, including all hormone biosynthesis- and metabolism-related proteins of the 2OGD family. In this study, the number and classifications of DOXC hormone biosynthesis- and metabolism-related proteins were consistent with the report by Kawal et al. [14]. DOXC proteins are involved in the biosynthesis and metabolism of the phytohormones auxin, GAs, ethylene, JA, SA, and SLs, which play important roles in plant growth and development. Furthermore, the number of DOXC hormone biosynthetic and metabolism genes increases from ancient lower land plants to higher plants, consistent with the high complexity and diversity—and specialized metabolism—of higher plants.

Although the 2OGD superfamily is highly diverse, structural studies suggest that its members have a highly conserved Fe(II) binding HxD/E … H triad motif and a less conserved 2OG C5 carboxy group binding motif (RxS/T) [13]. In this study, forty-three hormone biosynthetic and metabolism proteins of the DOXC family were identified in tomato, but five SlGA20ox7, SlGA20ox8, SlGA20ox9, SlGA20ox10 and SlGA2ox12) lacked the HxD/E … H or RxS/T motif (Figures S3 and S6), suggesting a lack of 2OGD activity. In addition, we identified family-specific conserved motifs in DAOs, GA20oxs, GA3oxs, C19-GA2oxs, C20-GA2oxs, ACOs, and JOXs (Figure 3b); however, their function was unclear. A MdACO1 protein with mutated conserved Lys296 and Arg299 residues in the C-terminal helix retained only 15–30% of the activity of the wild-type, possibly because these two residues are important for ACO activity and may be involved in binding bicarbonate, the unique activator of ACOs [18].

Notably, these two amino acids are located in the ACO-specific conserved motif identified in this study (Figure S7). Therefore, the subfamily-specific conserved motifs may play important roles in the functional differentiation of 2OGD subfamilies.

3.2. Functional Analysis of Hormone Biosynthetic and Metabolism Genes in 2OGD Family

GAs, ethylene, auxin, JA, SA, and SLs regulate many aspects of plant growth and development, and the response to stresses. Several 2OGD genes involved in hormone biosynthesis and metabolism have been functionally analyzed in *Arabidopsis* and rice, and these genes participate in the development of roots, stems, flowers, fruits, and seeds. In tomato, the *SlGA20oxs* GA-biosynthetic genes, particularly *SlGA3oxs*, which function in the final step of GA biosynthesis, were mainly expressed in tomato roots, leaves, flowers, and early developing fruits, suggesting that GAs play a role in the development of these tissues/organs (Figure 4a,b). Consistently, RNAi-mediated silencing of *SlGA20ox1*, *SlGA20ox2*, or *SlGA20ox3* affected the development of tomato stems, leaves, fruit, and seeds [19], and inhibitors of GA biosynthesis decreased tomato fruit growth and fruit set; also, exogenous GA_3 induced parthenocarpic fruits [20,21]. The *SlGA2oxs* GA-metabolism proteins also play key roles in regulating endogenous GA levels. The silencing of *SlGA2ox1*-*SlGA2ox5* increased the active GA_4 content, induced parthenocarpic fruits, and inhibited lateral branching in tomato plants [22]. In this study, the newly identified genes *SlGA2ox7* and *SlGA2ox10*, mainly expressed in roots, leaves, flowers, and early developing fruits (Figure 4c), had the same conserved motif as *SlGA2ox1* to *SlGA2ox5* (Figure S5), suggesting a role for *SlGA2ox7* and *SlGA2ox10* in the metabolism of GAs during the development of these tissues/organs.

Although auxin regulates the growth and development of various plant tissues and organs, studies of auxin in tomato have focused on fruit set and development. Exogenous auxin treatment could induce parthenocarpic fruits, and altering the expression of auxin response genes also affected tomato fruit set and development [21,23]. *DAO*-family proteins irreversibly degrade auxin, and a *dao* mutant in rice displayed defective pollen fertility and seed development [7]; meanwhile, a *dao1* mutant in *Arabidopsis* displayed larger cotyledons, increased lateral root density, and elongated pistils [24]. *DAO* has three homologs in tomato; the expression of *SlDAO2* was higher in flowers and early developing fruits compared to *SlDAO1* and *SlDAO3*, suggesting a role in regulating the auxin level for fruit set and development (Figure 4d). Ethylene plays important roles in fruit set and development [25], especially fruit ripening, likely due to high expression of the ethylene-biosynthetic genes *SlACO1*, *SlACO3*, and *SlACO6* in flower, early developing fruits, and ripening fruits (Figure 4f). Other *ACO* genes (*SlACO2* and *SlACO4*) may contribute to ethylene production for root and flower development. In addition, three JA-metabolism *SlJOX* genes showed high expression in tomato flowers (Figure 4e), indicating roles in regulating JA homeostasis for flowering [26]. *AtDMR6,* the product of which degrades salicylic acid, was involved in plant growth and resistance to pathogens, and the *dmr6* mutant displayed smaller size, early senescence, and a loss of susceptibility to *Pseudomonas syringae* pv tomato DC3000 [10]. In tomato, the homolog *SlDLO1* was highly expressed in roots, leaves, flowers, and fruits (Figure 4g), and CRISPR-Cas9 mediated the mutagenesis of *SlDLO1* in tomato conferred broad-spectrum disease resistance; however, vegetative growth and development were not significantly affected, and its role in reproductive organs was not investigated [27]. *SlDLO2* is highly expressed only in flowers and fruits, suggesting roles in regulating the SA level in reproductive organs. SLs are plant hormones that regulate plant root and branch development, as well as stress tolerance [28,29]. High expression of SL biosynthetic and signaling genes in tomato or strawberry fruit indicated roles in fruit development [30]. *LBO* acts in the final stages of SL biosynthesis to produce active SLs in *Arabidopsis*, and its homolog *SlLBO1* is only expressed in roots and flowers (Figure 4h). This suggests that SLs are synthesized in tomato roots and flowers, but does not mean that SLs have no effect on fruit development; they could be transported to fruit from other organs or tissues.

3.3. SlGA2ox2 and SlDAO1 May Play a Role in GA and Auxin Metabolism for Normal Ripening of Tomato Fruits

Tomato is a model plant for studying the ripening of climacteric fruits, and ethylene regulates tomato fruit ripening. In this study, exogenous GA_3 treatment of tomato fruits at the mature-green stage delayed fruit ripening, while overexpression of the GA catabolism gene *SlGA2ox1* specifically in tomato fruits led to early ripening [4]. We have previously shown that GAs play negative roles in the ethylene pathway by inhibiting the expression of ethylene biosynthetic genes (*SlACS2*, *SlACS4*, and *SlACO1*) and signaling genes (*SlETRs* and *SlEINs*) [4]. Therefore, the concentration of GAs in fruits influences fruit ripening in tomato. In plants, the GA level is regulated by the balance between biosynthesis and metabolism. GA20oxs and GA3oxs catalyze the rate-limiting step of active GA biosynthesis, and GA2oxs converts bioactive GAs or their immediate precursors into inactive forms. In this study, although the expression of one *GA20ox* gene (*SlGA20ox3*) increased from the mature-green to the breaker stage (Figure 4b), no *GA3ox* genes, which encode enzymes that catalyze the last step of GA biosynthesis, were expressed (Figure 4a), suggesting the absence of GA biosynthesis in mature-green and breaker fruits. Further, the expression of three GA-metabolism genes (*SlGA2ox2*, *SlGA2ox4*, and *SlGA2ox5*) was increased, and that of *SlGA2ox2* was highest, and dramatically increased, from the mature-green to the breaker stage (Figure 4c). It has been reported that the concentrations of endogenous active GAs (GA_1 and GA_4) in the fruit pericarp of tomato decrease significantly from the mature -green to the breaker stage (Figure S12) [4]. Therefore, we speculate that *SlGA2ox2* may be vital for GA metabolism from the mature-green to the breaker stage, and the reduced GA level caused by the increase in *SlGA2ox2* expression promotes tomato fruit ripening.

Auxin also negatively regulates tomato fruit ripening. Exogenous applications of IAA reduced expression of ethylene biosynthetic and consequently reduced ethylene production, and also the ethylene signaling genes, resulting in delayed tomato fruit ripening [1,2]. The concentration of endogenous auxin in tomato fruit pericarps is reduced from the mature-green to the breaker stage (Figure S12) [3]. In plants, auxin is synthesized by tryptophan (Trp)-dependent and -independent pathways [31]. Our knowledge of the genes and intermediates of the Trp-independent pathway is limited, but the complete Trp-dependent pathway has been established. YUCCA (YUC) family proteins function in the final step of Trp-dependent auxin biosynthesis, and play a crucial role in auxin biosynthesis in various plant species. In tomato, six *YUC* genes were identified, the transcript levels of five of which were negligible, whereas one *YUC* gene (*ToFZY4*) displayed high expression during ripening of tomato fruit [32]. It is not clear why the auxin concentration was decreased, but the expression of a key gene in auxin biosynthesis was increased in ripening tomato fruit. One explanation for this is that there is a change from the Trp-dependent to the Trp-independent pathway for auxin biosynthesis between the mature and red-ripe stages of tomato fruits [33], and *ToFZY4* may have a novel function related to tomato fruit ripening rather than auxin biosynthesis. Auxin can be deactivated by conjugation to amino acids, or by chemical oxidation. Conjugation of IAA to amino acids is catalyzed by *GH3*-family proteins and yields, for instance, indole-3-acetic acid aspartic acid (IAA-Asp) and indole-3-acetic acid glutamic acid (IAA-Glu). The chemical oxidation of auxin is catalyzed by DAO-family proteins to produce oxIAA. In tomato, 24 *GH3* genes were identified, only 4 (*SlGH3-1*, *SlGH3-2*, *SlGH3-5*, and *SlGH3-24*) of which showed high expression during fruit ripening [3]. Silencing of *SlGH3-2* in tomato increased the auxin level and reduced lycopene accumulation in ripening fruit, suggesting that *SlGH3-2* plays a role in deactivating free auxin to maintaining normal ripening of tomato fruit [3]. However, oxIAA is a major IAA catabolite, where up to 10–100 folds more oxIAA than the major IAA conjugates IAA-Glu and IAA-Asp was detected in *Arabidopsis* [34,35]. More importantly, oxIAA oxidized by DAO is biologically inactive, and is formed rapidly and irreversibly in plant tissues [34–36]. *DAO* is likely involved in maintaining the basal level of active auxin under normal growth conditions, while *GH3* functions in the response to various environmental factors [37]. In this study, we identified three DAO genes in tomato. *SlDAO3* had lost some sequences in the N-terminal (Figure S8), suggesting that it may be not involved in IAA degradation. *SlDAO2* expression was negligible, but that of *SlDAO1* was high and increased from mature-green to breaker fruits (Figure 4d); moreover, it was significantly induced by auxin in mature-green fruits (Figure 6b). These results implicate *SlDAO1*, rather than *SlDAO2* and *SlDAO3*, in auxin metabolism from the mature-green to the breaker stage during tomato ripening. In addition, the reduction in auxin level caused by the increase in *SlDAO1* expression may play an important role in maintaining normal ripening of tomato fruit.

4. Materials and Methods

4.1. Identification and Phylogenetic Analysis of Hormone Biosynthesis and Metabolism Related DOXC Proteins

To find proteins belonging to DOXC family, we used 2OG-FeII_Oxy (PF03171) domain as query in hmmsearch BLAST of *Arabidopsis*, rice, and tomato protein databases downloaded from JGI [38]. All sequences (length ≥ 100 aa) with an E-value cutoff 1×10^{-4} were retrieved. The obtained sequences were submitted to Pfam [39] and SMART [40] to verify the existence of 2OG-FeII_Oxy domain. In order to better understand the relationship among all members of the DOXC and identify proteins involved in hormone biosynthesis and metabolism, we then used all verified protein sequences to construct a phylogenetic tree by MEGA6 with Maximum likelihood. The best model JTT + F was selected by Model Generator software. According to hormone biosynthesis and metabolism related genes with known function in *Arabidopsis* and rice, all proteins which clustered into hormone biosynthesis and metabolism related protein subfamilies were selected to construct a new phylogenetic tree.

4.2. Chromosomal Location and Synteny Analysis

Genome annotation files were downloaded from the *Arabidopsis*, rice, and tomato databases to obtain chromosomal location information of these hormone biosynthetic and metabolism genes, then the Circos software was used to draw location pictures. A method similar to that developed for the Plant Genome Duplication Database (PGDD) [41] was used to identify syntenic blocks in *Arabidopsis*, rice, and tomato. Potential homologous sequences were initially identified by BLASTP (E-value $< 1 \times 10^{-5}$, top 5 matches). MCScanX was used for synteny analysis [42]. Additionally, MCScanX was further used to detect duplicate types of these biosynthetic and metabolism genes in tomato.

4.3. Multiple Sequence Alignment and Motif Composition Analysis

To detect the HxD/E … H and RxS/T motifs, multiple sequence alignments were performed by submitting protein sequences to ClustalW with the default parameters in BioEdit software. Motif composition analysis was performed by submitting protein sequences to MEME [43] with the following parameters: the maximum number of motifs was 50 and the maximum motif length was 15 amino acids.

4.4. Expression Analysis

Transcriptome datasets of different tomato organs were downloaded from Tomato Functional Genomics Database [44]. RPKM values of related genes were transformed in $\log_2$ level, and a heatmap was shown using MeV4.8 software (Dana-Farber Cancer Institute, Boston, MA, USA).

4.5. Plant Materials and Hormone Treatments

Two tomato cultivars Ai Ji Qiao Li grown in greenhouse and Micro-Tom grown in climate chamber were chosen as plant materials. The fruit was collected at four different ripening stages: mature-green (Mg), breaker (Br), yellow-ripening (Yr), and red-ripening (Rr). The fruit pericarp sample without placenta and seeds was collected and then immediately frozen in liquid nitrogen prior to storage at −80 °C until RNA extraction.

Tomato cultivars Micro-Tom grown in climate chamber was used for hormone treatments of fruits. Flowers were tagged at the date of pollination. After 36 days, mature-green fruits on the plants were injected with 0.1 mM IAA, 0.1 mM GA_3, and 0.1 mM ethephon, respectively, distilled water was used as the control. The amount of injection was about 50 µL per fruit. Twelve fruits for each treatment were performed. The fruit pericarp without placenta and seeds was collected at two days and four days after treatments, and were immediately frozen in liquid nitrogen, and then stored at −80 °C. Plant growth conditions was: 16-h light (25 °C)/8-h dark (18 °C) photoperiod cycle and 65% relative humidity. In addition, detached mature-green fruits were injected with 0.1 mM IAA and 0.1 mM GA_3, respectively, distilled water was used as the control. Then the fruit was placed under dark at 25 °C and 90% relative humidity, photos were taken after eight days.

4.6. RNA Extraction and qPCR Analysis of Selected Genes

Total RNA was extracted with a modified CTAB method [4]. cDNA library was generated by Primerscript RT reagent Kit with gDNA Erase (Takara, Beijing, China) according to the manufacturer's protocol. qPCR was carried out using SYBR Premix Ex Taq II (Takara, Beijing, China). Primer sequences were listed in Table S6. Three biologicals with triplicates were performed and results were analyzed using the $2^{-\Delta CT}$ method. *Actin* gene (gene ID: Solyc11g005330) was used as the reference.

5. Conclusions

We have identified 43 hormone biosynthetic and metabolism genes of nine subfamilies of the 2OGD family, which were related to GAs, ethylene, auxin, JA, SA, and SLs in tomato. The subfamily-specific conserved motifs identified in this study might play roles in the functional differentiation of 2OGD subfamilies, and the different expression profiles suggest that these genes play diverse roles in tomato organ growth and development. Especially, the expression levels of the auxin-degradation gene *SlDAO1* and the GA-degradation gene *SlGA2ox2* were significantly increased from the mature-green to the breaker stage during tomato fruit ripening, accompanied by decreased endogenous IAA and GAs levels. In addition, the expression of *SlDAO1* and *SlGA2ox2* was increased by IAA and GA_3, respectively, indicating that *SlDAO1* and *SlGA2ox2* may be responsible for reducing IAA and GA concentrations to maintain normal ripening of tomato fruit.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/21/15/5344/s1, Figure S1: Phylogenetic tree of DOXC family proteins identified in *Arabidopsis*, rice, and tomato, Figure S2: Chromosomal location and duplication analysis of hormone biosynthetic and metabolism genes in tomato, Figure S3: Sequence alignment of GA20ox group proteins, Figure S4: Sequence alignment of GA3ox group proteins, Figure S5: Sequence alignment of GA2ox (C19) group proteins, Figure S6: Sequence alignment of GA2ox (C20) group proteins, Figure S7: Sequence alignment of ACO group proteins, Figure S8: Sequence alignment of DAO group proteins, Figure S9: Sequence alignment of Jox group proteins, Figure S10: Sequence alignment of DMR6 group proteins, Figure S11: Sequence alignment of LBO group proteins, Figure S12: The concentrations of GA_1, GA_4, and IAA decrease from the mature-green (Mg) to the breaker (Br) stage, Table S1: Identification and characterization of hormone biosynthetic and metabolism proteins in *Arabidopsis*, rice, and tomato, Table S2: Microsynteny relationships of hormone biosynthetic and metabolism genes in *Arabidopsis*, rice, and tomato, Table S3: Duplication modes of hormone biosynthetic and metabolism genes in tomato, Table S4: Motif composition in DOXC class of 2OGD family in *Arabidopsis*, rice, and tomato, Table S5: Motif sequence identified by MEME tools in *Arabidopsis*, rice, and tomato DOXC class of 2OGDs, Table S6: Primers used in this study.

Author Contributions: Data curation, X.Y. and H.W.; formal analysis, Q.D.; funding acquisition, H.W.; investigation, Q.D. and F.W.; methodology, Q.D. and J.X.; software, F.W. and J.F.; supervision, H.W.; validation, H.W.; visualization, H.C.; writing—original draft, Q.D. and H.W.; writing—review and editing, Y.L. and H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (2016YFD0100506); The Special Financial Grant Program from the China Postdoctoral Science Foundation [2016T90471]; The General Financial Grant Program from the China Postdoctoral Science Foundation [2015M581812]; The Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

Acknowledgments: We thank Helin Tan for his support for the experiment in this study.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

GAs	Gibberellins
JA	Jasmonate
SA	Salicylic acid
SL	Strigolactone
GA20ox	GA20-oxidase
GA3ox	GA3-oxidase
GA2ox	GA2-oxidase
DAO	Dioxygenase for Auxin Oxidation
ACO	1-aminocyclopropane-1-carboxylic acid oxidase
JOX	JASMONATE-INDUCED OXYGENASE
DMR6	Downy Mildew Resistant6
DLO	DMR6-LIKE OXYGENASE
LBO	LATERAL BRANCHING OXIDOREDUCTASE
FPKM	Fragments Per Kilobase Per Million

References

1. Su, L.; Diretto, G.; Purgatto, E.; Danoun, S.; Zouine, M.; Li, Z.; Roustan, J.P.; Bouzayen, M.; Giuliano, G.; Chervin, C. Carotenoid accumulation during tomato fruit ripening is modulated by the auxin-ethylene balance. *BMC Plant Biol.* **2015**, *15*, 114–125. [CrossRef]
2. Li, J.Y.; Tao, X.Y.; Li, L.; Mao, L.; Luo, Z.S.; Khan, Z.U.; Ying, T. Comprehensive RNA-Seq analysis on the regulation of tomato ripening by exogenous auxin. *PLoS ONE* **2016**, *11*, e0156453. [CrossRef]
3. Sravankumar, T.; Akash; Naik, N.; Kumar, R. A ripening-induced *SlGH3-2* gene regulates fruit ripening via adjusting auxin-ethylene levels in tomato (*Solanum lycopersicum* L.). *Plant Mol. Biol.* **2018**, *98*, 455–469. [CrossRef]
4. Li, H.; Wu, H.; Qi, Q.; Li, H.; Li, Z.; Chen, S.; Ding, Q.; Wang, Q.; Yan, Z.; Gai, Y.; et al. Gibberellins play a role in regulating tomato fruit ripening. *Plant Cell Physiol.* **2019**, *60*, 1619–1629. [CrossRef]
5. Farrow, S.C.; Facchini, P.J. Functional diversity of 2-oxoglutarate/Fe(II)-dependent dioxygenases in plant metabolism. *Front Plant Sci.* **2014**, *5*, 1–15. [CrossRef] [PubMed]
6. Hedden, P.; Thomas, S.G. Gibberellin biosynthesis and its regulation. *Biochem. J.* **2012**, *444*, 11–25. [CrossRef]
7. Zhao, Z.; Zhang, Y.; Liu, X.; Zhang, X.; Liu, S.; Yu, X.; Ren, Y.; Zheng, X.; Zhou, K.; Jiang, L.; et al. A role for a dioxygenase in auxin metabolism and reproductive development in rice. *Dev. Cell.* **2013**, *27*, 113–122. [CrossRef] [PubMed]
8. Argueso, C.T.; Hansen, M.; Kieber, J.J. Regulation of ethylene biosynthesis. *J. Plant Growth Regul.* **2007**, *26*, 92–105. [CrossRef]
9. Caarls, L.; Elberse, J.; Awwanah, M.; Ludwig, N.R.; Vries, D.; Zeilmaker, T.; Wees, S.C.M.V.; Schuurink, R.C.; Ackerveken, G.V. Arabidopsis JASMONATE-INDUCED OXYGENASES down-regulate plant immunity by hydroxylation and inactivation of the hormone jasmonic acid. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 6388–6393. [CrossRef]
10. Zhang, Y.; Zhao, L.; Zhao, J.; Li, Y.; Wang, J.; Guo, R.; Gan, S.; Liu, C.; Zhang, W. S5H/DMR6 encodes a salicylic acid 5-hydroxylase that fine-tunes salicylic acid homeostasis. *Plant Physiol.* **2017**, *175*, 1082–1093. [CrossRef]
11. Brewer, P.B.; Kaori, Y.; Fiona, F.; Emma, M.; Adrian, S.; Tancred, F.; Kohki, A.; Yoshiya, S.; Elizabeth, A.D.; Julia, E.C.; et al. Later BRANCHING OXIDOREDUCTASE acts in the final stages of strigolactone biosynthesis in Arabidopsis. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 6301–6306. [CrossRef] [PubMed]
12. Aik, W.; McDonough, M.A.; Thalhammer, A.; Chowdhury, A.; Schofield, C.J. Role of the jelly-roll fold in substrate binding by 2-oxoglutarate oxygenases. *Curr. Opin. Struc. Biol.* **2012**, *22*, 691–700. [CrossRef]
13. Clifton, I.J.; McDonough, M.A.; Ehrismann, D.; Kershaw, N.J.; Granatino, N.; Schofield, C.J. Structural studies on 2-oxoglutarate oxygenases and related double-stranded beta-helix fold proteins. *J. Inorg. Biochem.* **2006**, *100*, 644–669. [CrossRef]
14. Kawal, Y.; Ono, E.; Mizutani, M. Evolution and diversity of the 2–oxoglutarate-dependent dioxygenase superfamily in plants. *Plant J.* **2014**, *78*, 328–343.

15. Kataoka, H.; Yamamoto, Y.; Sekiguchi, M. A new gene (alkB) of Escherichia coli that controls sensitivity to methyl methane sulfonate. *J. Bacteriol.* **1983**, *153*, 1301–1307. [CrossRef] [PubMed]
16. Keskiaho, K.; Hieta, R.; Sormunen, R.; Myllyharju, J. Chlamydomonasreinhardtii has multiple prolyl 4–hydroxylases, one of which is essential for proper cell wall assembly. *Plant Cell* **2007**, *19*, 256–269. [CrossRef]
17. Takehara, S.; Sakuraba, S.; Mikami, B.; Yoshida, H.; Yoshimura, H.; Itoh, A.; Endo, M.; Watanabe, N.; Nagae, T.; Matsuoka, M.; et al. A common allosteric mechanism regulates homeostatic inactivation of auxin and gibberellin. *Nat. Commun.* **2020**, *11*, 2143. [CrossRef]
18. Yoo, A.; Seo, Y.S.; Jung, J.W.; Sung, S.K.; Kim, W.T.; Lee, W.; Yang, D.R. Lys296 and Arg299 residues in the C-terminus of MD-ACO1 are essential for a 1-aminocyclopropane-1-carboxylate oxidase enzyme activity. *J. Struct. Biol.* **2006**, *156*, 407–420. [CrossRef]
19. Xiao, J.; Li, H.; Zhang, J.; Chen, R.; Zhang, Y.; Ouyang, B.; Wang, T.; Ye, Z. Dissection of GA 20-oxisdase members affecting tomato morphology by RNAi-mediated silencing. *Plant Growth Regul.* **2006**, *50*, 179–189. [CrossRef]
20. Chen, S.; Wang, X.; Zhang, L.; Lin, S.; Liu, D.; Wang, Q.; Cai, S.; El-Tanbouly, R.; Gan, L.; Wu, H.; et al. Identification and characterization of tomato gibberellin 2-oxidases (GA2oxs) and effects of fruit-specific *SlGA2ox1* overexpression on fruit and seed growth and development. *Hortic. Res.* **2016**, *3*. [CrossRef]
21. Serrani, J.C.; Fos, M.; Atares, A.; Garcia-Martinez, J.L. Effect of gibberellin and auxin on parthenocarpic fruit growth induction in the cv Micro-Tom of tomato. *J. Plant Growth Regul.* **2007**, *26*, 211–221. [CrossRef]
22. Martinez-Bello, L.; Moritz, T.; Lopez-Diaz, L. Silencing C19-GA 2-oxidases induces parthenocarpic development and inhibits lateral branching in tomato plants. *Plant Physiol.* **2015**, *66*, 5897–5910. [CrossRef] [PubMed]
23. Jong, M.; Wolters-Arts, M.; Garcia-Martinez, J.L.; Mariani, C.; Vriezen, W.H. The Solanum lycopersicum AUXIN RESPONSE FACTOR 7 (SlARF7) mediates cross-talk between auxin and gibberellin signaling during tomato fruit set and development. *J. Exp. Bot.* **2011**, *62*, 617–626. [CrossRef]
24. Zhang, J.; Lin, J.E.; Harris, C.; Pereira, F.C.M.; Wu, F.; Blakeslee, J.J.; Peer, A.J. DAO1 catalyzes temporal and tissue-specific oxidative inactivation of auxin in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 11010–11015. [CrossRef] [PubMed]
25. Shinozaki, Y.; Hao, S.; Kojima, M.; Sakakobara, H.; Ozeki-Lida, Y.; Zheng, Y.; Fei, Z.; Zhong, S.; Giovannoni, J.J.; Rose, J.K.; et al. Ethylene suppresses tomato (Solanum lycopersicum) fruit set through modification of gibberellin metabolism. *Plant J.* **2015**, *83*, 237–251. [CrossRef]
26. Zhai, Q.; Zhang, X.; Wu, F.; Feng, H.; Deng, L.; Xu, L.; Zhang, M.; Wang, Q.; Li, C. Transcriptional mechanism of jasmonate receptor COI1-mediated delay of flowering time in Arabidopsis. *Plant Cell* **2015**, *27*, 2814–2828. [CrossRef] [PubMed]
27. Thomazella, D.P.; De, P.; Brail, Q.; Dahlbeck, D.; Staskawicz, B. CRISPR-Cas9 mediated mutagenesis of a DMR6 ortholog in tomato confers broad-spectrum disease resistance. *bioRxiv* **2016**. [CrossRef]
28. Waters, M.; Gutjahr, C.; Bennett, T.; Nelson, D. Strigolactone signaling and evolution. *Annu. Rev. Plant Biol.* **2017**, *68*, 291–322. [CrossRef]
29. Nasir, F.; Tian, L.; Shi, S.; Chang, C.; Ma, L.; Gao, Y.; Tian, C. Strigolactones positively regulate defense against Magnaporthe oryzae in rice (Oryza sativa). *Plant Physiol. Biochem.* **2017**, *142*, 106–116. [CrossRef]
30. Wu, H.; Li, H.; Chen, H.; Qi, Q.; Ding, Q.; Xue, J.; Ding, J.; Jiang, X.; Hou, X.; Li, Y. Identification and expression of strigolactone biosynthetic and signaling genes reveal strigolactones are involved in fruit development of the woodland strawberry (*Fragaria vesca*). *BMC Plant Biol.* **2019**, *19*, 73. [CrossRef]
31. Zhao, Y. Essential roles of local auxin biosynthesis in plant development and in adaptation to environmental changes. *Annu. Rev. Plant Biol.* **2018**, *69*, 417–435. [CrossRef]
32. Exposito, R.M.; Borges, A.A.; Borges-Perez, A.; Perez, J.A. Gene structure and spatiotemporal expression profile of tomato genes encoding YUCCA-like flavin monooxygenases: The ToFZY gene family. *Plant Physiol. Biochem.* **2011**, *49*, 782–791. [CrossRef]
33. Epstein, E.; Cohen, J.D.; Slovin, J.P. The biosynthetic pathway for indole-3-acetic acid changes during tomato fruit development tomato fruit development. *Plant Growth Regul.* **2002**, *38*, 15–20. [CrossRef]
34. Kowalczyk, M.; Sandberg, M. Quantitative analysis of indole-3-acetic acid metabolites in Arabidopsis. *Plant Physiol.* **2001**, *127*, 1845–1853. [CrossRef] [PubMed]

35. Pencik, A.; Simonovik, B.; Petersson, S.V.; Henykova, E.; Simon, S.; Greenham, K.; Zhang, Y.; Kowalczyk, M.; Estelle, M.; Zazimalova, E.; et al. Regulation of auxin homeostasis and gradients in Arabidopsis roots through the formation of the indole-3-acetic acid catabolite 2-oxindole-3-acetic acid. *Plant Cell* **2013**, *25*, 3858–3870. [CrossRef]
36. Ostin, A.; Kowalyczk, M.; Bhalerao, R.P.; Sandberg, G. Metabolism of indole-3-acetic acid in *Arabidopsis*. *Plant Physiol.* **1998**, *118*, 285–296. [CrossRef]
37. Stepanova, A.N.; Alonso, J.M. Auxin catabolism unplugged: Role of IAA oxidation in auxin homeostasis. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 10742–10744. [CrossRef]
38. JGI. Available online: http://phytozome.jgi.doe.gov/ (accessed on 30 January 2020).
39. PFAM. Available online: http://pfam.xfam.org/search (accessed on 1 February 2020).
40. SMART. Available online: http://smart.embl-heidelberg.de/ (accessed on 1 February 2020).
41. Lee, T.H.; Tang, H.; Wang, X.; Paterson, A.H. PGDD: A database of gene and genome duplication in plants. *Nucleic Acids Res.* **2013**, *41*, 1152–1158. [CrossRef] [PubMed]
42. Wang, Y.; Tang, H.; DeBarry, J.D.; Tan, X.; Li, J.; Wang, X.; Lee, T.; Jin, H.; Marler, B.; Guo, H.; et al. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **2012**, *40*. [CrossRef]
43. MEME. Available online: http://memesuite.org/tools/meme (accessed on 5 February 2020).
44. Tomato Functional Genomics Database. Available online: http://ted.bti.cornell.edu/cgi-bin/TFGD/digital/home.cgi (accessed on 10 February 2020).

International Journal of
Molecular Sciences

MDPI

Article

Silencing of *TaCKX1* Mediates Expression of Other *TaCKX* Genes to Increase Yield Parameters in Wheat

Bartosz Jabłoński [1], Hanna Ogonowska [1], Karolina Szala [1], Andrzej Bajguz [2], Wacław Orczyk [3] and Anna Nadolska-Orczyk [1,*]

[1] Department of Functional Genomics, Plant Breeding and Acclimatization Institute—National Research Institute, Radzikow, 05-870 Blonie, Poland; b.jablonski@ihar.edu.pl (B.J.); h.ogonowska@ihar.edu.pl (H.O.); k.szala@ihar.edu.pl (K.S.)

[2] Laboratory of Plant Biochemistry, Faculty of Biology, University of Bialystok, Ciolkowskiego 1J, 15-245 Bialystok, Poland; abajguz@uwb.edu.pl

[3] Department of Genetic Engineering, Plant Breeding and Acclimatization Institute—National Research Institute, Radzikow, 05-870 Blonie, Poland; w.orczyk@ihar.edu.pl

* Correspondence: a.orczyk@ihar.edu.pl

Received: 19 June 2020; Accepted: 3 July 2020; Published: 7 July 2020

Abstract: *TaCKX*, *Triticum aestivum* (cytokinin oxidase/dehydrogenase) family genes influence the development of wheat plants by the specific regulation of cytokinin content in different organs. However, their detailed role is not known. The *TaCKX1*, highly and specifically expressed in developing spikes and in seedling roots, was silenced by RNAi-mediated gene silencing via *Agrobacterium tumefaciens* and the effect of silencing was investigated in 7 DAP (days after pollination) spikes of T_1 and T_2 generations. Various levels of *TaCKX1* silencing in both generations influence different models of co-expression with other *TaCKX* genes and parameters of yield-related traits. Only a high level of silencing in T_2 resulted in strong down-regulation of *TaCKX11 (3)*, up-regulation of *TaCKX2.1*, *2.2*, *5*, and *9* (*10*), and a high yielding phenotype. This phenotype is characterized by a higher spike number, grain number, and grain yield, but lower thousand grain weight (TGW). The content of most of cytokinin forms in 7 DAP spikes of silenced T_2 lines increased from 23% to 76% compared to the non-silenced control. The CKs cross talk with other phytohormones. Each of the tested yield-related traits is regulated by various up- or down-regulated *TaCKX* genes and phytohormones. The coordinated effect of *TaCKX1* silencing on the expression of other *TaCKX* genes, phytohormone levels in 7 DAP spikes, and yield-related traits in silenced T_2 lines is presented.

Keywords: wheat; cereals; *TaCKX1*; *TaCKX* expression; grain yield; cytokinins; phytohormones; gene silencing; RNAi; wheat spikes

1. Introduction

Wheat (*Triticum aestivum* L.) is the third most economically important crop in the world after corn and rice, and probably the most important in moderate climates. It provides approximately 20% of human calories and protein [1]. The large genome of this high-yielding species, composed of three (AABBDD) genomes, has been very challenging for improving traits [2]. However, it might be a great reservoir to sustain a further increase of grain productivity [3]. The continuous increase of wheat production is necessary to feed the rapidly growing world population [4]. Biotechnological tools implemented in the process of increasing wheat productivity are expected to be beneficial.

Cytokinins (CKs) are important regulators of plant growth and development, influencing many agriculturally important processes [5]. This regulation might occur at the posttranscriptional and/or posttranslational level [6,7], or by the modulation of context-dependent chromatin accessibility [8]. CKs modulate the expression of other genes involved in the control of various processes including meristem activity, hormonal cross talk, nutrient acquisition, and various stress responses [9]. There is growing evidence on their key role in seed yield regulation [10]. In cereals and grasses, an increased content of CKs has been reported to positively affect sink potential in developing grains [11] and maintain leaf chlorophyll status during plant senescence [12] and grain filling [13].

The majority of naturally occurring CKs in plants belong to isoprenoid cytokinins grouping N^6-(12-isopentenyl) adenine (iP), *trans*-zeatin (tZ), *cis*-zeatin (cZ), and dihydrozeatin (DZ) derived from tRNA degradation or from isopentenylation of free adenine nucleosides catalysed by isopentenyltransferase (IPT) or tRNA-IPT. The second, smaller group comprise N6-aromatic CKs, represented by benzyladenine (BA) [14]. To better characterize their physiological role, CKs are classified into such -base active forms as tZ, cZ, and iP, translocation forms (nucleosides) as tZ-ribosides (tZR), which exhibit a low level of activity, and sugar conjugates (*O*-glucosides), which are storage and inactivated forms [14,15].

CKs function as local or long-distance regulatory signals, but the mechanisms of their precise spatial and temporal control are still largely unknown [16]. They are produced in roots as well as in various sites of the aerial part of plants [17]. The level of CKs in respective cells and tissues is dependent on many processes, including biosynthesis, metabolism, activation, transport, and signal transduction. Active CKs can be metabolized via oxidation by cytokinin oxidase/dehydrogenase (CKX) or by activity of glycosyltransferases. Many reports have demonstrated that the irreversible degradation step by the CKX enzyme plays an important role in the regulation of cytokinin level in some cereals, namely maize [18], rice [19], barley [20,21], and wheat [22].

The *CKX* gene families in plants show different numbers of genes and various expression patterns, which are tissue- and organ-specific, suggesting gene-specific functions. The specificity of expression of 11 *TaCKX* in developing wheat plants were assigned to four groups: highly specific to leaves, specific to developing spikes and inflorescences, highly specific to roots and expressed through all the organs tested [23]. The *TaCKX* genes co-operated inside and among organs. Their role in plant productivity has been described in many plants including model plants and some cereals. Knock-out mutation or silencing by RNAi of *OsCKX2* in rice significantly increased grain number [19]. The same effect of elevated grain number, spike number, and yield was reported for RNAi-silenced *HvCKX1* in barley [20,21,24] and repeated for the same gene under field conditions [25]. Moreover, significantly increased grain number per spike was found as the effect of the *TaCKX2.4* gene silencing by RNAi [26]. Knock-out mutation of *HvCKX1* by CRISPR/Cas9 editing had a limited effect on yield productivity, however significantly decreased CKX enzyme activity in young spikes and 10-day old roots corresponded to greater root length, numbers of root hairs and increased surface area [27]. In contrast, roots of knock-out mutants of ckx3 were smaller.

The role of other *TaCKX* genes in wheat was analysed based on natural *TaCKX* variation. Haplotype variants of *TaCKX6a02* and *TaCKX6-D1* were related to higher filling rate and grain size [28,29]. Quantitative trait locus (QTL) found in recombinant inbred lines containing a higher copy number of *TaCKX4* was associated with higher chlorophyll content and grain size [30].

To arrange the numbering of *TaCKX* family genes, a new annotation for the first two was suggested by Ogonowska et al. (2019) based on the Ensembl Plants database [31] and phylogenetic analysis. *TaCKX6a02* was annotated as *TaCKX2.1*, *TaCKX6-D1* (JQ797673) was annotated as *TaCKX2.2* and *TaCKX2.4* was annotated as *TaCKX2.2*. Annotations for these genes were maintained in the recently published review on the *TaCKX* [22], however tested in this research *TaCKX10* was renamed as *TaCKX9* and *TaCKX3* was renamed as *TaCKX11*. Newly revised by Chen et al. [22], naming is applied and former names are given in brackets.

Due to the size and complexity of the wheat genomes, the knowledge about the role of *TaCKX* genes, containing homologues from three genomes, is more difficult to obtain, because of the limited number of natural mutants. Most homoeologous genes are expected to have overlapping functions [32], therefore the effect of gene mutations might be masked by the other genomes. One solution to silence all of them is to apply RNAi-mediated gene silencing, which allows silencing of all the homologues.

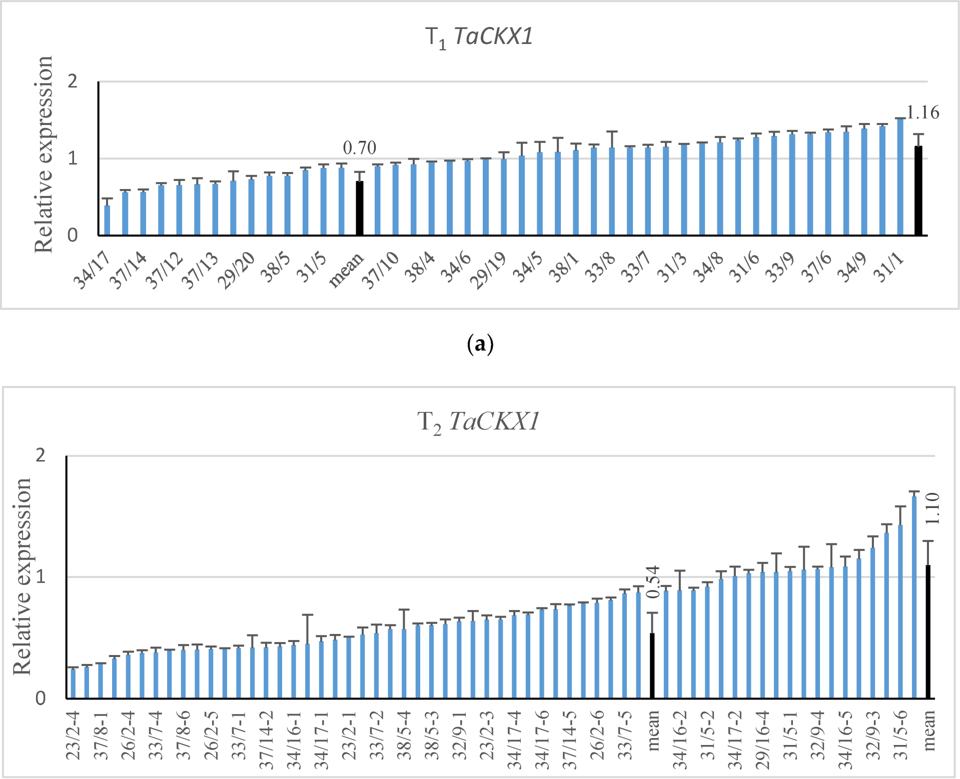
Moreover, this tool made it possible to obtain a number of lines with different levels of silencing, which in the case of genes coding proteins of key importance for life gave a possibility to regenerate plants for analysis [33]. The introduction of a silencing cassette by stable transformation results in a stable, and inherited to T_4, effect of silencing [21,34]. The applicability of *Agrobacterium*-mediated transformation compared to a biolistic one for gene silencing of the developmentally regulated gene *HvCKX10* (2) was proved to be reliable [24].

We present the first report on the role of *TaCKX1* in the co-regulation of expression of other *TaCKX* genes, phytohormone content, and their joint participation in the regulation of yield-related traits in wheat. Various levels of gene silencing in T_1 and T_2 have been related to different patterns of other *TaCKX* expression, strongly influencing yield-related traits. Models of regulation of phytohormone levels and phenotypic traits in non-silenced and highly silenced T_2 plants by the coordinated expression of *TaCKX* genes are proposed.

2. Results

2.1. Expression Levels of Silenced *TaCKX1* in Segregating T_1 and T_2 Plants

Expression levels of *TaCKX1* were measured in 44 segregating T_1 plants from 8 T_0 PCR+ lines. In 14 T_1 plants relative expression (related to the control = 1.00) ranged from 0.39 to 0.88 with the mean of 0.67 (±0.14). In 30 T_1 plants, relative expression ranged from 0.90 to 1.52 with the mean of 1.16 (±0.18) (Figure 1). The proportion of silenced to non-silenced plants changed in the T_2 generation. There were 42 silenced from 0.24 to 0.88 plants with the mean of 0.54 (±0.14) and 20 non-silenced plants. Eight of them, with low relative expression ranging from 0.24 to 0.40 (mean 0.33 ±0.14) and representing different T_1 lines, were selected for further analysis.



(**a**)

(**b**)

Figure 1. Relative expression level of silenced *TaCKX1* in segregating T_1 (**a**) and T_2 (**b**) plants. The level of expression is related to the control set as 1.00.

2.2. Co-Expression of Silenced TaCKX1 with Other TaCKX Genes in T_1 and T_2 and CKX Enzyme Activity

Mean relative expression of *TaCKX1* in the selected 8 lines was 0.67 in T_1 and was decreased to 0.33 in T_2 (Figure 2). Similarly, in the case of *TaCKX11* (*3*) related gene expression was 0.81 in T_1 and was decreased to 0.34 in T_2. Relative expression levels of *TaCKX2.2* and *TaCKX9* (*10*) were decreased in T_1 to 0.51 and 0.39 and increased in T_2 slightly above the control level, to 1.08 and to 1.10 respectively. Mean relative values for *TaCKX2.1* were similar to control in T_1 (1.05) and slightly increased in T_2 (1.17). Relative expression of *TaCKX5*, which was in T_1 below the control level (0.84), was significantly increased to 1.82 in T_2. The relative values of CKX enzyme activity in both generations were around the control, 1.00.

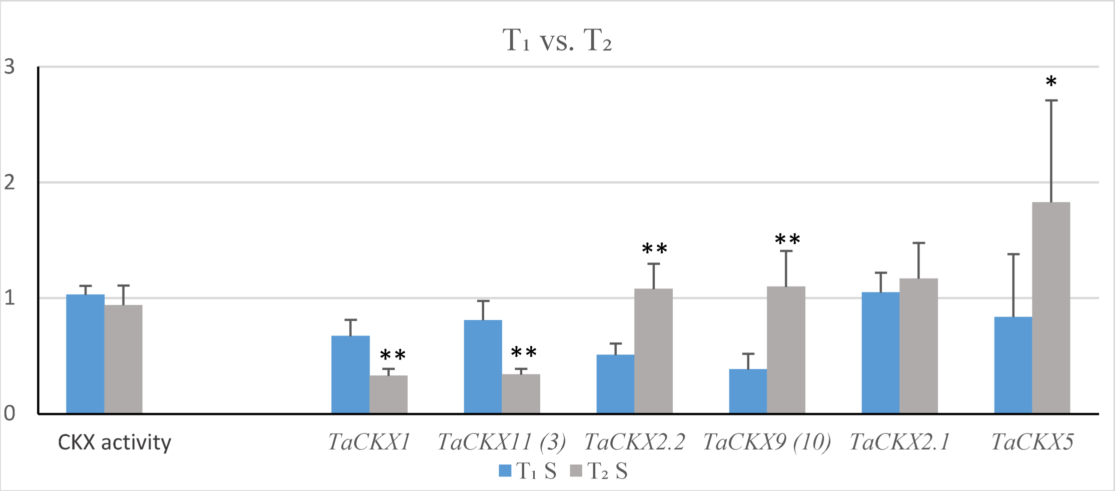


Figure 2. Comparison of means of relative CKX enzyme activity and selected gene expression levels in T_1 (bars) and T_2 (line) generation of silenced lines. *—significant at $p < 0.05$; **—significant at $p < 0.01$.

The effect of *TaCKX1* silencing on the levels of expression of selected *TaCKX* genes is presented by the expression ratio indicator (Table 1), which is a quotient of the mean relative value in silent per mean relative value in non-silent, control plants. In the case of *TaCKX1* and *TaCKX11* (*3*), the ratio indicator, significantly decreased in T_1, was strongly decreased in T_2. The value of the ratio indicator for *TaCKX2.2* was not changed in T_1 compared to the control and was only slightly decreased in T_2. The expression ratio indicator of *TaCKX9 (10)*, strongly decreased to 0.59 in T_1, rose above the control level (1.15) in T_2. Already high in T_1, the expression ratio indicator for *TaCKX2.1* (1.22) increased to 1.32 in T_2. The phenotype ratio indicator for CKX enzyme activity was 1.01 in T_1 and 0.99 in T_2.

Table 1. Effect of *TaCKX1* silencing on expression levels of selected *TaCKX* genes presented by expression ratio indicator (mean value in silent/mean value in non-silent, control plants) in T_1 and T_2 generations.

	T_1 (SD)	T_2 (SD)	Effect of *TaCKX1* Silencing T_1/T_2
TaCKX1 *	0.58 (0.12)	0.28 (0.05)	decreased/strongly decreased
TaCKX11 (3)	0.80 (0.16)	0.36 (0.05)	decreased/strongly decreased
TaCKX2.2	1.08 (0.22)	0.98 (0.18)	slightly increased/similar
TaCKX9 (10)	0.59 (0.20)	1.15 (0.32)	strongly decreased/slightly increased
TaCKX2.1	1.22 (0.19)	1.32 (0.35)	increased/increased
TaCKX5	1.00 (0.65)	1.08 (0.52)	the same/similar
CKX activity	1.01 (0.07)	0.99 (0.18)	the same/the same

*—significant at $p < 0.05$.

In T_1 segregating plants, CKX enzyme activity significantly correlated with spike length (0.51; n = 16) and grain weight (0.50; n = 16), but in T_2 these correlations were not significant.

2.3. Influence of TaCKX1 Silencing on Phenotypic Traits and Chlorophyll Content in Flag Leaves of T_1 and T_2 Plants

The values of phenotypic traits in T_1 plants with slightly decreased relative expression of *TaCKX1* (0.67 ± 0.14) compared to control plants (1.00) were on the same level in the case of plant height and lower for number of spikes, spike length, grain number, and grain yield (Supplementary Table S2). Higher values were obtained for TGW. Data for chlorophyll content measured by SPAD in the flag leaves of first spikes and the next spikes were similar. All these differences were not significant. Opposite results were obtained for some traits in T_2 plants with highly silent *TaCKX1* (0.33 ± 0.06) compared to the control (1.00) (Supplementary Table S3). Silent T_2 plants were substantially smaller, had a higher number of spikes, number of grains, grain yield, seedling root weight, and lower SPAD values for the flag leaves of first spikes. TGW and spike length were significantly lower than in control plants.

These differences between the slightly silenced T_1 and highly silent T_2 generation are expressed by comparison of ratio indicators of phenotypic traits in both generations (Figure 3). There were no changes in plant height, TGW or spike length in T_1 plants compared to the control. However, these values were respectively 7%, 10%, and 25% lower in T_2 plants. Opposite phenotype ratio indicators for number of spikes per plant and number of grains per plant were about 21% and 30% lower in T_1 and 57% and 29% higher in T_2. These differences for spike number, grain number, and TGW were significant.

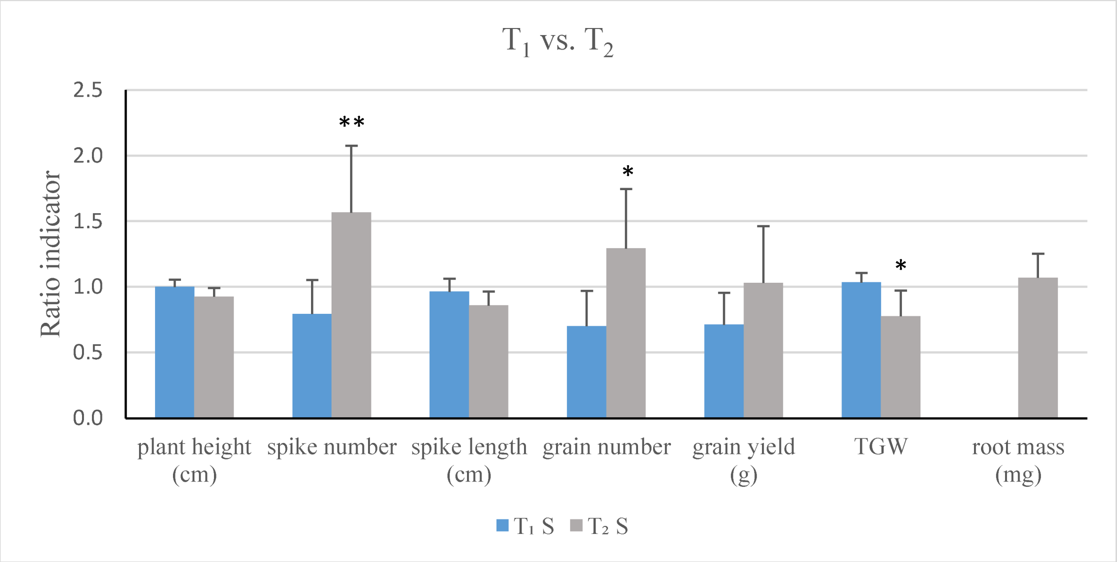


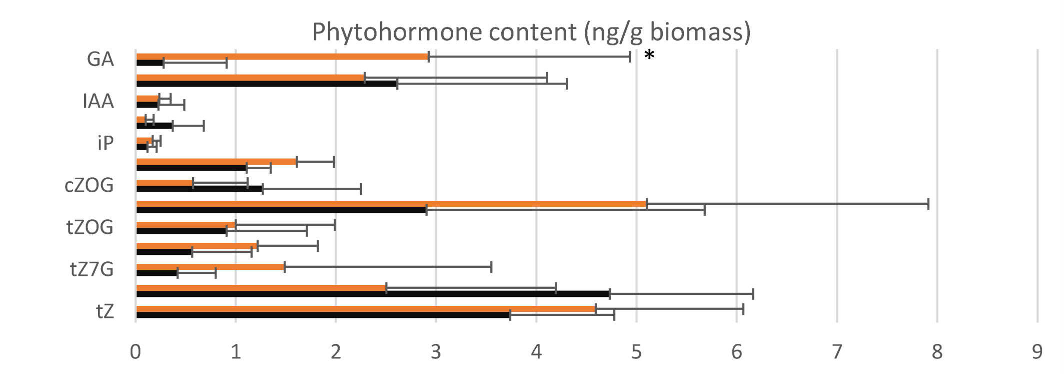
Figure 3. Comparison of phenotypic effect of silencing of *TaCKX1* in T_1 and T_2 generations based on ratio indicators. *—significant at $p < 0.05$; **—significant at $p < 0.01$.

The levels of expression of *TaCKX1* in 7 DAP spikes of all T_1 significantly correlated with number of grains, grain weight, spike length and spike number (0.47, 0.39, 0.42 and 0.33 respectively; n = 42) and grain weight correlated with enzyme activity (0.33; n = 42). The *TaCKX9 (10)* expression level significantly correlated with grain number (0.51; n = 16).

Correlation coefficients among the expression of all tested *TaCKX* genes and enzyme activity, and phenotypic traits in non-silent and highly silent T_2 are included in Supplementary Table S4A,B. All these correlations are graphically presented in Figures and described in Section 2.6.

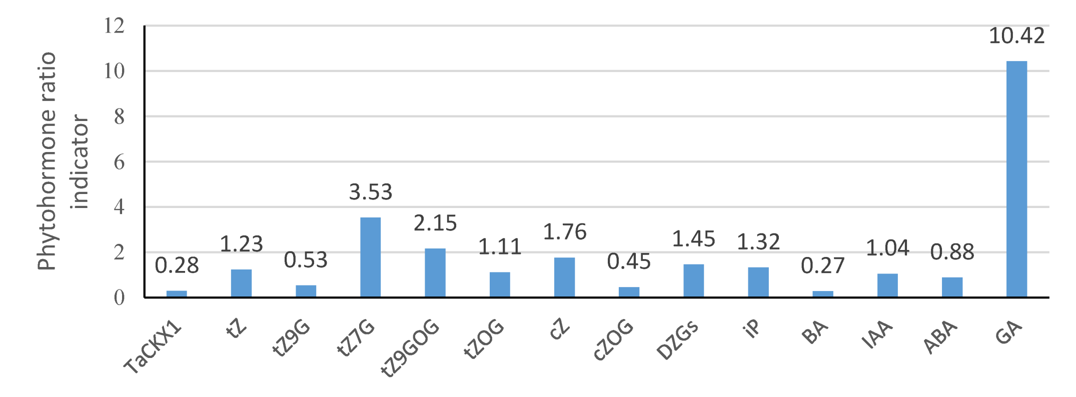
2.4. Phytohormone Content in 7 DAP Spikes of T_2

tZGs, which were mainly composed of tZ9G, tZ7G, tZOG and tZ9GOG, were the most abundant cytokinin group in 7 DAP spikes (Figure 4a). Their mean content in control plants was 6.97 ng/g biomass and in silent T_2 was 6.24 ng/g biomass respectively. The second most abundant was tZ with the level of 3.74 ng/g biomass in the control and 4.59 ng/g biomass in silent T_2. The content of cZ was slightly lower to tZ (2.90 ng/g biomass) in control but higher (5.10 ng/g biomass) in silent plants. cZOG was more abundant in the control than the groups of silent plants, and the content was 1.27 and 0.57 ng/g biomass respectively. The concentration of DZGs (sum of DZ7G, DZOG, DZ9G and DZOGR) was higher in silent (1.61 ng/g biomass) than in control plants (1.11 ng/g biomass). Low concentrations (below 0.5 ng/g biomass) were measured for iP and BA. The concentration of IAA was also low and on a comparable level in control and in silent plants (0.23 and 0.24 ng/g biomass respectively). In the case of ABA, the concentration in the control was slightly decreased in silent plants (2.61 and 2.29 ng/g biomass respectively). The concentration of GA was increased from 0.28 ng/g biomass in the control to 2.93 ng/g biomass in silent plants, which was more than a 10-fold increase.



	tZ	tZ9G	tZ7G	tZ9GOG	tZOG	cZ	cZOG	DZGs	iP	BA	IAA	ABA	GA
🟧 Silent	4.59	2.50	1.49	1.22	1.00	5.10	0.57	1.61	0.17	0.10	0.24	2.29	2.93
⬛ Non-silent	3.74	4.73	0.42	0.57	0.91	2.90	1.27	1.11	0.12	0.37	0.23	2.61	0.28

(**a**)



(**b**)

Figure 4. Phytohormone content (ng/g biomass) measured in the group of control and silent T_2 plants (**a**). Phytohormone ratio indicators (mean value in silent per mean value in not silent, control plants) in silent T_2 plants (**b**). *—significant at $p < 0.05$. Small amounts (≤1.00 ng/g biomass): tZR, tZOGR, cZOGR, DZOG, DZ7G, DZ9G, DZOGR, iP, iP7G, BA, IAA. Trace amounts (≤0.05 ng/g biomass) or not detected: cZ9G, cZR, DZ, DZR, iPR, IBA, IPA, NAA, PAA.

Most of the phytohormone ratio indicators in the group of six silent T_2 plants (Figure 4b) were much higher than in control plants. There were the following cytokinins: tZ (1.23), tZ7G (3.53), tZ9GOG (2.15), tZOG (1.11), cZ (1.76), sum of DZGs (1.45) and iP (1.32). The ratio indicators for some of them were significantly lower, as in the case of BA (0.27), cZOG (0.45) and tZ9G (0.53). Similar values were observed for IAA (1.04), and slightly lower for ABA (0.88), but much higher for GA (10.42).

2.5. Coordinated Effect of TaCKX1 Silencing on Expression of Other TaCKX Genes and Phytohormone Level in 7 DAP Spikes as Well as Phenotype in T_2

A graphic presentation of the coordinated effect of *TaCKX1* silencing on expression of other *TaCKX* genes and phytohormone levels in 7 DAP spikes as well as the phenotype of T_2 plants is presented in Figure 5. The significant decrease of expression of *TaCKX1* was coordinated with the significant decrease of *TaCKX11* (*3*), which presumably resulted in a significant increase of most CKs: tZ, tZGs, cZ, DZGs, iP, as well as GA. The increased phytohormone level in the first 7 DAP spikes positively influenced traits such as spike number and grain number, reaching the ratio indicators 1.57 and 1.29, respectively, and negatively influenced TGW (0.78), spike length (0.86), plant height (0.93), and flag leaf senescence (0.95). Opposing data were obtained for *TaCKX2.1* and *TaCKX9* (*10*), which showed increased expression in silenced 7 DAP spikes (1.32 and 1.15 respectively). This might have influenced the decreased ratio indicators for phytohormones—cZOG (0.45), BA (0.27), and ABA content (0.88), and slightly increased ratio indicators for yield-related traits: root weight and grain yield (1.07 and 1.03 respectively). Expression ratio indicators for *TaCKX5* and *TaCKX2.2* were both close to 1.00, but their expression significantly increased compared to T_1 and positively correlated with the expression of *TaCKX2.1* and *TaCKX9* (*10*) respectively.

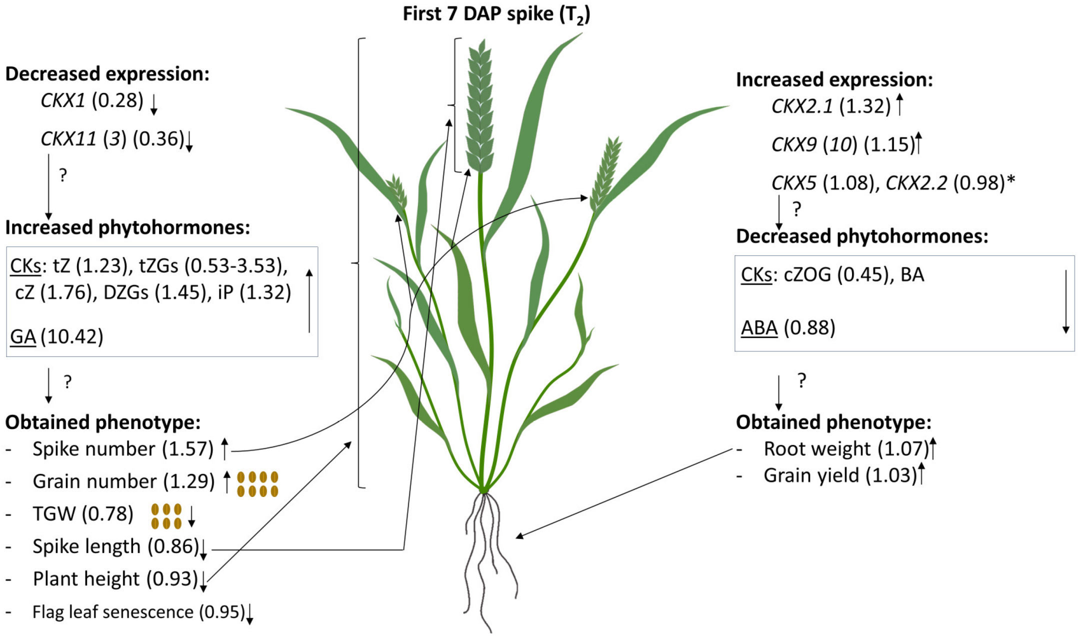
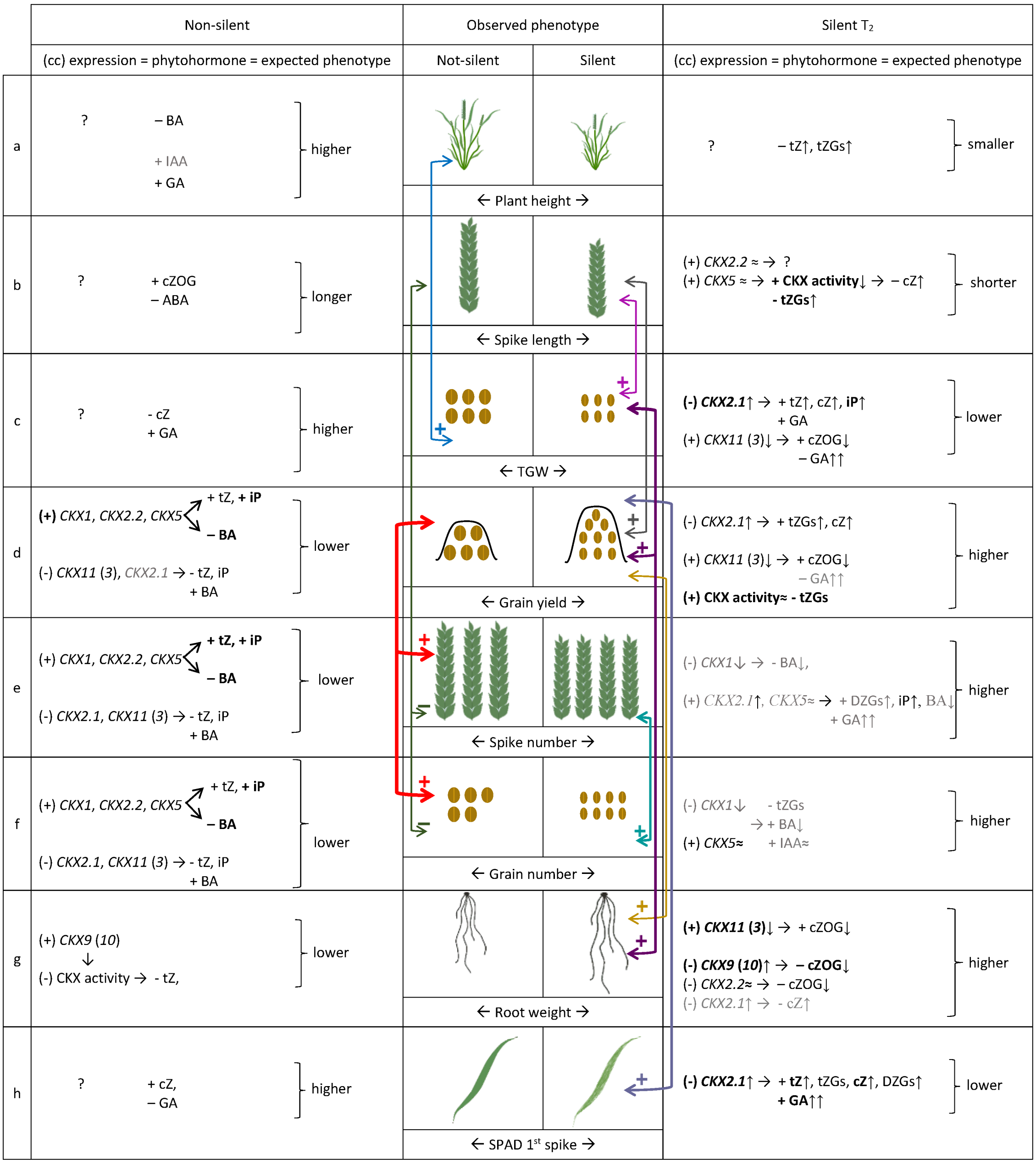


Figure 5. Graphic presentation of coordinated effect of *TaCKX1* silencing on expression of other *TaCKX* genes, phytohormone levels as well as phenotype in 7 DAP spikes of T_2 plants based on ratio indicators. *—significantly increased comparing to T_1; ?—expected changes.

2.6. Models of Co-Regulation of Phytohormone Levels and Phenotype Traits by Coordinated Expression of TaCKX Genes in Non-Silenced and Silenced T_2 Plants

Two different models of co-regulation of *TaCKX* expression, phytohormone levels and phenotypic traits in non- silenced and silenced plants of the T_2 generation are proposed (Figure 6a–h) based on correlation coefficients (Table S4A,B).



	Non-silent	Observed phenotype		Silent T2
	(cc) expression = phytohormone = expected phenotype	Not-silent	Silent	(cc) expression = phytohormone = expected phenotype
a	? – BA + IAA + GA } higher	← Plant height →		? – tZ↑, tZGs↑ } smaller
b	? + cZOG – ABA } longer	← Spike length →		(+) *CKX2.2* ≈ → ? (+) *CKX5* ≈ → **+ CKX activity↓** → – cZ↑ **- tZGs↑** } shorter
c	? - cZ + GA } higher	← TGW →		**(-) *CKX2.1*↑** → + tZ↑, cZ↑, **iP↑** + GA (+) *CKX11 (3)*↓ → + cZOG↓ – GA↑↑ } lower
d	**(+)** *CKX1, CKX2.2, CKX5* → + tZ, **+ iP**; → **– BA** (-) *CKX11* (*3*), *CKX2.1* → - tZ, iP + BA } lower	← Grain yield →		(-) *CKX2.1*↑ → + tZGs↑, cZ↑ (+) *CKX11* (*3*)↓ → + cZOG↓ – GA↑↑ **(+) CKX activity≈ - tZGs** } higher
e	(+) *CKX1, CKX2.2, CKX5* → **+ tZ, + iP**; → **– BA** (-) *CKX2.1, CKX11* (*3*) → - tZ, iP + BA } lower	← Spike number →		(-) *CKX1*↓ → - BA↓, (+) *CKX2.1*↑, *CKX5*≈ → + DZGs↑, iP↑, BA↓ + GA↑↑ } higher
f	(+) *CKX1, CKX2.2, CKX5* → + tZ, **+ iP**; → **– BA** (-) *CKX2.1, CKX11* (*3*) → - tZ, iP + BA } lower	← Grain number →		(-) *CKX1*↓ - tZGs → + BA↓ (+) *CKX5*≈ + IAA≈ } higher
g	(+) *CKX9* (*10*) ↓ (-) CKX activity → - tZ, } lower	← Root weight →		**(+) *CKX11* (3)↓** → + cZOG↓ **(-) *CKX9* (10)↑** → **– cZOG↓** (-) *CKX2.2*≈ → – cZOG↓ (-) *CKX2.1*↑ → - cZ↑ } higher
h	? + cZ, – GA } higher	← SPAD 1st spike →		**(-) *CKX2.1*↑** → + **tZ**↑, tZGs, **cZ**↑, DZGs↑ **+ GA↑↑** } lower

Figure 6. Models of regulation of phytohormone levels and phenotypic traits by coordinated expression of *TaCKX* genes based on correlation coefficients (cc) in non-silenced and silenced wheat plants (**a**–**h**). (cc)—correlation coefficient between expression and trait; (?) – lack of correlation with expression of any gene; bold—strong, significant correlations at $p \leq 0.05$ (cc above 0.82); grey—cc from 0.5 to 0.6.

Plant height (Figure 6a). There was no correlation between plant height and expression values of any *TaCKX* expressed in 7 DAP spikes of non-silent as well as silent plants. In the first group of plants this trait negatively correlated with BA and positively with IAA and GA content. By contrast, in silent plants the values of plant height were negatively correlated with growing concentration of tZ and tZGs, which resulted in a smaller plant phenotype.

Spike length (Figure 6b) in non-silent plants was positively correlated with BA, and negatively with cZ and ABA content. These correlations determined longer spikes and the trait negatively correlated with spike number and grain number. A strong positive correlation between CKX activity and spike length was noted in silent plants. The values of enzyme activity correlated positively with slightly increased *TaCKX5* expression, which negatively correlated with increasing content of cZ and tZGs. Spike length in silent plants was positively correlated with grain yield.

TGW (Figure 6c). There was no correlation of TGW with expression of any *TaCKX* expressed in 7 DAP spikes of non-silent plants. However, the trait was strongly negatively correlated with cZ content and positively with GA. The grains in this group of plants were larger and TGW higher. By contrast, in silent plants there was a strong negative correlation of the trait with growing expression of *TaCKX2.1*, which positively regulated tZ, cZ, iP, and GA content. Moreover, the values of expression of down-regulated *TaCKX11* (*3*) positively correlated with decreasing content of cZOG, negatively with highly growing GA and positively with the trait. Altogether it resulted in lower TGW compared to non-silenced plants. The trait in silent plants was strongly and positively correlated with grain yield (0.82) and root weight (0.77).

Grain yield (Figure 6d). Expression levels of *TaCKX1*, *TaCKX2.2* and *TaCKX5* in non-silent plants positively correlated with tZ and iP and negatively with BA content. However, expression of *TaCKX11* (*3*) and *TaCKX2.1* regulates the same CKs in opposite way. Altogether, it resulted in lower grain yield comparing to silenced plants, and the trait was strongly positively correlated with spike number (0.93) and grain number (0.99). The increasing expression of *TaCKX2.1* positively correlated with a growing content of tZGs and cZ and negatively with the trait in silent plants. Decreasing expression of *TaCKX11* (*3*), which was positively correlated with decreased cZOG content and negatively with GA content, positively correlated with the trait. A positive correlation was observed between CKX activity and grain yield in this group of plants, which was higher than in non-silent plants. Moreover, CKX activity negatively correlated with tZGs. The trait was strongly correlated with TGW (0.82) and root weight (0.66).

Spike number (Figure 6e) and grain number (Figure 6f) in non-silenced plants were positively regulated by *TaCKX1*, *TaCKX2.2* and *TaCKX5*, and their expression was positively correlated with tZ, iP and negatively with BA. On the other hand, expression levels of *TaCKX2.1* plus *TaCKX11* (*3*) were negatively correlated with the traits as well as with tZ, iP and positively with BA. Both groups of genes finally affected lower spike and grain numbers in non-silent plants in comparison to silent plants and were strongly and positively correlated with each other (0.91) and grain yield (0.93 and 0.99 respectively). In silent plants decreasing expression of *TaCKX1* is negatively correlated with both spike and grain number and the gene negatively regulates decreasing BA content. In the case of grain number, the main player positively correlated with the trait is *TaCKX5*, increased expression of which was correlated with slightly higher IAA content, which resulted in higher grain number. Spike number is also positively regulated by *TaCKX5* co-expressed with *TaCKX2.1*, and both genes were positively correlated with growing CKs, DZGs and iP as well as GA, determining higher spike number. Both traits are highly correlated (0.88) with each other.

Seedling root weight (Figure 6g). There was strong, positive correlation between *TaCKX9* (*10*) expression in 7 DAP spikes and seedling root weight in non-silenced plants. Moreover, CKX activity negatively correlated with tZ (in spikes) and the trait, which finally resulted in lower root weight. The decreasing expression of *TaCKX11* (*3*) in the case of silent plants was positively correlated with decreasing content of cZOG and strongly positively correlated with the trait. Increasing expression levels of *TaCKX9* (*10*) plus *TaCKX2.2* negatively correlated with decreasing content of cZOG and root weight.

Chlorophyll content measured by SPAD in flag leaves of first spikes (Figure 6h). There was no correlation between expression level of any *TaCKX* measured in 7 DAP spikes of non-silent plants and the trait. The only correlations were between phytohormone content and the trait, positive for cZ and negative for GA, which resulted in higher SPAD values (chlorophyll content). Increasing expression of *TaCKX2.1* was strongly positively correlated with growing values of tZ, tZGs, cZ, and DZGs as well as GA in silent plants. A strong negative correlation was observed between the gene expression and chlorophyll content, which means that increasing expression of *TaCKX2.1* in 7 DAP spikes results in lower chlorophyll content in silent plants.

3. Discussion

First, 7 DAP spike was chosen as a research objective in wheat since decreased *HvCKX1* expression at this stage in barley resulted in higher yield due to the higher spike and grain number [20,21]. The *TaCKX1* gene is an orthologue of *HvCKX1* and both genes are specifically expressed in developing spikes [23], indicating their possibly important role in the regulation of yield-related traits. The samples were taken from the middle part of the spikes, when anthesis starts, in order to ensure a similar developmental stage of spikelets for research. The 7 DAP spikes of wheat represent the middle of cell division/cell expansion stage [35,36].

3.1. Various Levels of TaCKX1 Silencing Influence Different Models of Co-Expression with Other TaCKX Genes and Parameters of Yield-Related Traits

Various levels of silencing of *TaCKX1* in T_1 and T_2 generate different results of co-expression with other *TaCKX* genes and plant phenotype. For example, the expression of *TaCKX9* (*10*) was highly and significantly correlated with *TaCKX1* only in T_1. However, a new and strong positive correlation between *TaCKX9* (*10*) and *TaCKX2.2* in highly silenced T_2 was observed. Slightly decreased co-expression of silenced *TaCKX1* together with *TaCKX11* (*3*) in T_1 was much stronger in T_2, indicating their positive co-regulation. It should be underlined that there is no homology between the sequence of *TaCKX1* used for silencing and sequences of other *TaCKX* genes tested. Therefore, the process of RNAi silencing was specifically addressed to *TaCKX1* silencing. It indicates that the level of silencing of the modified gene affected variable levels of expression of the other *TaCKX* genes in a co-operative process maintaining homeostasis of CKX enzyme in the research object. The models of co-regulation of other *CKX* by highly silenced *TaCKX1* and knock-out *HvCKX1* [27] differ between these species.

The differences in the levels of expression of *TaCKX1* and various co-expression of other *TaCKX* genes in T_1 and T_2 resulted in opposite phenotypic effects. Since spike number, grain number, and grain yield were reduced in T_1, the same yield-related traits were significantly higher in highly silenced T_2 plants. High-yielding phenotype occurred when highly silenced *TaCKX1* co-operated with down-regulated *TaCKX11* (*3*) but up-regulated *TaCKX5*, *TaCKX2.2*, *TaCKX2.1*, and *TaCKX9* (*10*). These differences showed that both levels of silencing might be helpful to better understand the function of developmentally regulated genes. Unexpectedly, changes in the expression levels of co-working *TaCKX* did not result in different enzyme activity, even in highly silenced T_2 plants. This might be explained by the fact that down-regulation of *TaCKX1* and *TaCKX11* (*3*) is compensated for by the up-regulation of *TaCKX2.2*, *TaCKX5*, and *TaCKX9* (*10*), and therefore the contribution of isozymes encoded by the genes in the general pool of CKX enzyme activity is the same. Since CKX enzymes indicate different specificities for the particular cytokinin hormone [37], the cytokinin contribution and phenotypic traits of modified plants were changed accordingly, with consequent differences in the active pool of CKs influencing phenotype.

3.2. Co-Operating Effect of TaCKX on the Level of Active CKs in Silenced Plants

Since CKX isozymes specifically degrade CKs, the highly decreased expression of *TaCKX1* and *TaCKX11* (*3*) in 7 DAP spikes is expected to result in the observed increase of most major forms of CKs: tZGs, tZ, and cZ in silenced plants. We documented that both tZ and cZ, which are isomers of zeatin, together with their derivatives are a major group of isoprenoid CKs in 7 DAP spikes. It has already been shown that trans-zeatin is the predominant form after anthesis [36,38], but comprehensive analysis of cytokinins during spike, spikelet, ovule and grain development has not yet been reported for wheat using LC-MS/MS [22]. The content of DZGs increased by 40% in silent compared to non-silent wheat plants, suggesting that this less known isoprenoid form of CKs might also play an important role in plant productivity. Interestingly, isoprenoid iP was represented in 7 DAP spikes of non-silent plants at very low quantities, but its content in 7 DAP spikes of silent plants was increased by 32%. A similar relationship between the reduced expression of selected *CKX* family genes and cytokinin accumulation in reproductive organs has been observed in other species including *A. thaliana* [39], rice [19], and barley [25], but detailed data are not comparable to our research in wheat.

The physiological significance of these isoprenoid forms is still not very well known. tZ and iP, which are susceptible to CKX, were found the most abundant and bioactive CKs in maize, whereas cZ, which shows low affinity to CKX was reported to have a weak biological impact and unknown biological role [40,41]. However, the cZ concentrations changed significantly during development in maize grain, as well as in shoot and root tissues [42,43]. High levels of cZ at the first developmental stage of barley spike observed by Powell et al. [44] might indicate an important role of this form in early barley embryo development, what is also documented in our results (discussed further below).

The BA is represented in 7-DAP spikes of wheat at trace amounts but their content was significantly decreased in silent plants. However, their correlations with the *TaCKX* genes as well as yield-related traits of non-silenced plants indicate their importance (discussed in more detail below). Interestingly, BA was found to participate in posttranscriptional and/or posttranslational regulation of protein abundance in *Arabidopsis*, showing high specificity to shoots and roots, and affected differential regulation of hormonal homeostasis [45].

3.3. Cross Talk of CKs with Other Phytohormones

Negative correlations between ABA content and *TaCKX2.2* and *TaCKX9* (*10*) expression, and positive with *TaCKX11* (*3*), were associated with a slight decrease of ABA content in 7 DAP spikes of silenced plants. Moreover, ABA was strongly positively correlated with BA. The main auxin, IAA, remained at the same level. A ten-fold increase of GA content in silenced comparing to non-silenced plants was observed. Such cross regulation of CKs and other plant hormones is documented in other species. In maize kernels the *CKX1* gene is up-regulated by cytokinin and ABA, and abiotic stress [18]. In tobacco altered cytokinin metabolism affected cytokinin, auxin, and ABA contents in leaves and chloroplasts [46], which host the highest proportion of CK-regulated proteins [47]. Moreover, auxin, ABA and cytokinin are involved in the hormonal control of nitrogen acquisition and signalling [48], which often limits plant growth and development. All four phytohormones, CKs, GA, IAA, and ABA, were found to be involved in the regulation of grain development in drought conditions [49]. Moreover, in shoots, BA up-regulated the abundance of proteins involved in ABA biosynthesis and the ABA response, whereas in the roots, BA strongly up-regulated the majority of proteins in the ethylene biosynthetic pathway [45]. We proved that IAA, GA, and ABA contents are also co-regulated by CKs in non-silenced and silenced 7 DAP spikes. Up-regulation of major CKs and down-regulation of some minor ones in silent plants influence GA, ABA, and IAA content in a similar manner as in abiotic stress conditions.

3.4. Coordinated Effect of TaCKX Gene Expression on the Content of CKs, Other Phytohormones and Yield-Related Traits

Plant height in non-silenced plants is down-regulated by BA and up-regulated by IAA and GA content in the first 7 DAP spikes, resulting in taller plants. Oppositely, increased content of tZ and tZGs negatively correlated with the trait in silent plants, stimulated plant height. As it was already showed [50,51] and similarly to our results, plant height and root weight are regulated by CKs and IAA in opposite ways. This may be dependent on basipetal auxin flow in the stem, which suppresses axillary bud outgrowth, and similarly as in pea, auxin derived from a shoot apex suppresses the local level of CKs in the nodal stem through the regulation of *CKX* or *IPT* genes [52].

The main role in spike length seemed to be played by cZ and its glucoside. Increased content of cZOG in non-silenced plants negatively correlated with ABA, resulting in longer spikes. In silent plants the trait is positively regulated by *TaCKX2.2* together with *TaCKX5*, and the latter is a positive regulator of enzyme activity and negative of cZ content. Consequently, a higher content of cZ in 7 DAP spikes led to shorter spikes. cZOG, found as a positive regulator of longer spikes, is a sugar conjugate of cZ-0-glucoside, which is the inactivated form of cZ, showing metabolic stability against CKX activity [53]. Moreover, 0-glucosylation of cZ is catalysed by a specific 0-glucosyltransferase, cisZOG1, discovered in maize [54], and this form mainly functions in the early stages of seed development. Knowledge of function of cZ degradation pathways via the CKX enzyme is limited. Interestingly, two *Arabidopsis* genes, *CKX1* and *CKX7*, expressed in stages of active growth, were shown to have high preference for cZ [37]. In our case the *TaCKX5* positively regulated CKX activity and negatively cZ content.

None of the tested individual *TaCKX* genes was involved in high TGW in non-silenced plants, but a negative correlation with cZ and positive with GA was found. Otherwise a significant negative correlation of *TaCKX2.1* and a positive correlation of *TaCKX11* (*3*) in determining low TGW were observed in silenced plants. Unexpectedly increased expression of the first one positively influenced tZ, cZ, and iP content and negatively GA content, and the opposite was true for the second gene, resulting in lower TGW. Therefore both *TaCKX2.1* and *TaCKX11* (*3*), acting in an opposite manner, maintain homeostasis of CKX enzyme activity and co-regulate TGW in silenced plants. A greater concentration of CKs, especially tZ, was observed during the grain filling stage of high-yielding cultivars [44]. We might suppose that the observed higher concentrations of tZ and other CKs at the 7 DAP stage, which originally was a consequence of *TaCKX1* silencing, might accelerate germination of the grains, which resulted in smaller grains/lower TGW than in non-silenced plants. The silenced *TaCKX1* co-work with down-regulated *TaCKX11* (*3*) in increasing CK content as well as up-regulating *TaCKX2.1*, with seems to play a regulatory role. The involvement of GA in TGW and other traits demonstrated by us might be the effect of co-regulation of *CKX* and other gibberellin-responsive genes regulating yield-related traits as well [55,56]. Fahy et al. [57] suggested that final grain weight might be largely determined by developmental processes prior to grain filling. This is in agreement with our observations, in which yield-related traits are differently regulated in two groups of plants, non-silent and silent. Therefore, we might suppose that the coordinated co-regulation of expression of *TaCKX* genes and related CKs takes place during whole plant and spike development and small seeds in silenced plants are determined at earlier stages.

Grain yield, which is very strongly correlated with grain and spike number in non-silent plants but with TGW in silent plants, is a more complex feature. Two groups of genes up-regulating or down-regulating grain yield in non-silent plants have been found. The first one includes *TaCKX1*, *2.2*, and *5* positively regulating iP content but negatively BA. The second comprises *TaCKX11* (*3*) acting in down-regulation of tZGs. Both groups might determine lower grain yield. It is worth to mention that *TaCKX5*, which is highly expressed in inflorescences and leaves might be a main player of this trait. Higher grain yield was positively regulated by enzyme activity and both, down-regulated *TaCKX11* (*3*) as well as up-regulated *TaCKX2.1* in silenced plants. Again, the *TaCKX2.1* positively regulated tZGs and cZ content just like for TGW, which is rather untypical for a gene encoding a CKX enzyme degrading CKs. Therefore, the positive regulation of the main CK content by *TaCKX2.1* observed by us supports its role in regulation of expression of other genes rather than encoding the CKX isozyme.

As observed in barley cultivars, changes in cytokinin form and concentration in developing kernels correspond with variation in yield [44]. Interestingly, the authors observed no peaks and no differences in CKX activity at the particular stages of spike development. This is in agreement with the homeostasis of the pool of isozymes in 7 DAP spikes of wheat, as suggested by us, which is independent of the level of silencing of *TaCKX1* but is rather a consequence of co-regulation of expression of other *TaCKX* genes. A similar effect of increased grain yield, which was a consequence of higher spike and grain number, was obtained in barley with silenced by RNAi *HvCKX1*, an orthologue of *TaCKX1* [20,21,25]. In this research, CKX activity was decreased, however according to Zalewski et al. [20], it was measured not in 7 DAP spikes, but in 0 DAP spikes and seedling roots. Therefore this inconsistency might be result of measurements in various organs/developmental stages. Another explanation is that these two cereal species varied three times in ploidy level, what might influence differences in action of both orthologues. The *TaCKX* homologues located on A, B and D chromosomes might significantly affect homeostasis of pooled CKX isozymes in wheat. Incomparable to the results obtained for RNAi silenced *TaCKX1* and

HvCKX1, no changes in yield parameters were observed in mutant lines with knock-out of *HvCKX1* (Gasparis et al., 2019). These essential phenotypic differences between RNAi-silenced *TaCKX1* and *HvCKX1* or knocked out by CRISPR-Cas9 *HvCKX1* might be the result of different processes involved in inactivation of the gene. The first one is regulated at the posttranscriptional and the second at the transcriptional level. Since CKs might regulate various developmental and physiological processes at the posttranscriptional level [6,7] or by modulation of context-dependent chromatin accessibility [8], the way of deactivating *TaCKX* function seemed to be important.

Spike number and grain number are highly correlated in both non-silent and silent plants and are regulated by the same groups of *TaCKX* genes as well as phytohormones. The first group includes *TaCKX1*, *2.2* and *5* positively regulating iP but negatively BA. The second comprises *TaCKX11* (*3*) and *2.1* acting in the opposite way, and homeostasis of these hormones in non-silenced plants maintains a lower spike number. The main role in controlling higher spike and grain number in silent plants seemed to be played by *TaCKX5*, highly expressed in seedling roots, leaves, inflorescences and 0 DAP spikes. These correlations are not significant because they were measured in a stage of plant development in which the number of spikes and seed number have already been set. As reported, the higher spike number was the consequence of a higher tiller number, which was positively correlated with the content of endogenous zeatin in the field-grown wheat after exogenous hormonal application [58]. Shoot branching might also be dependent on the acropetal transport of cytokinin [52].

Root weight was positively correlated with lower expression of *TaCKX9* (*10*) in 7 DAP spikes of non-silent plants and, negatively with increased expression of this gene in silenced plants. Therefore the gene might determine lower root weight in the first group of plants, but higher in the second. Increased expression of *TaCKX9* (*10*) down-regulated cZOG. The same cZOG was up-regulated by *TaCKX11* (*3*), but expression of this gene in 7 DAP spikes of silent plants is strongly decreased. Both cZ and cZOG are involved in spike length regulation as well as TGW and grain yield in the group of silenced plants. Although both tested organs are in different developmental stages, correlations between *TaCKX9* (*10*) and *TaCKX11* (*3*) expression in 7 DAP spikes and weight of seedling roots are reasonable. The *TaCKX9* (*10*) is mainly expressed in younger organs from seedling roots to 0 DAP spikes and highly expressed in leaves. The *TaCKX11* (*3*) is expressed in all organs tested [23] and both seemed to regulate seedling roots as well, although in the opposite manner. Therefore, we should take into consideration the possible action of cytokinin transport and signalling genes as well as other phytohormones which take part in hormonal crosstalk to control the regulation of root growth [59]. Accordingly, cZ type CKs found as the major forms in phloem are translocated from shoots to roots [60,61]. Some *CKX* genes might be induced by transcription factors [62,63], what is also observed in our unpublished yet data.

The lower plant height and higher root weight observed in the group of silenced plants of wheat is in agreement with opposed regulation of these traits by CKs and IAA mentioned above [64,65]. Up-regulated content of active cZ in 7 DAP spikes, might influence down-regulation of this CK in roots. It has been documented that such suppressing cZ levels mediated by overexpression of *AtCKX7* affected root development in *Arabidopsis* [66]. A higher weight of seedling root was also obtained by silencing via RNAi or knock-out via CRISPR/Cas9 of *HvCKX1* in barley plants, as in wheat, and the trait corresponded with decreased activity of CKX enzyme measured in roots (Zalewski et al., 2010; Gasparis et al., 2019).

Leaf senescence was determined in the flag leaf of the first spike by measuring chlorophyll content. Increased expression of *TaCKX2.1* in silent plants up-regulated tZ, tZGs and cZ content in 7 DAP spikes and down-regulated the trait. The gene functions in a similar way, by up-regulating these CKs in determining lower TGW and higher grain yield in silent plants. A higher content of active CKs as well as GA in 7 DAP spikes of silent plants is expected to down-regulate CKs in the flag leaves, accelerating their senescence, what is documented by the results.

It was previously demonstrated that level of chlorophyll content in flag leaves is associated with the senescence process, in which CKs suppress inhibition of senescence [67]. During this processes, proteins are degraded and nutrients are re-mobilised from senescing leaves especially to the developing grains [68]. We might suppose that slower spike ripening in non-silent plants, which is dependent on lower CK content in the 7 DAP spike, causes a slower flow of micronutrients as well as CKs from flag leaf to spike. Therefore, prolonged chlorophyll content in the flag leaf of the first spike negatively correlated with TGW but positively with plant height. Opposite data were obtained for flag leaves of silent plants, in which higher content of CKs in 7 DAP spikes might be the result of faster flow accelerating leaf senescence. The reduced chlorophyll content in flag leaves of the first spike of silent plants positively correlated with grain yield. The important role of tZ and less active cZ in the suppression of senescence was proven in maize leaves [69] and in an oat-leaf assay [37]. It was also documented that delayed senescence of wheat stay-green mutant, tasg1, at the late filling stage was related to high cytokinin and nitrogen contents [70].

4. Materials and Methods

4.1. Vector Construction

The hpRNA type of silencing cassette was constructed in pBract207 (https://www.jic.ac.uk/technologies/crop-transformation-bract/). It contains the Hpt selection gene under the 35S promoter and cloning sites for the cloning silencing cassette under the Ubi promoter. The vector is compatible with the gateway cloning system. For cloning purposes a coding sequence of *TaCKX1* (NCBI JN128583) 378 codons long was used. In the first step, the cassette was amplified using: EAC11-F: 5′-TTGAATTCGACTTCGACCGCGGCGTTTT-3′ and EAC12-R: 5′-TTGAATTC ATGTCTTGGCCAGGGGAGAG-3′ and cloned into the entry vector pCR8/GW/TOPO (Invitrogen). In the next step, the cassette was cloned to the destination Bract7 vector in the gateway reaction. The presence of the silencing cassette in the vector was verified by restriction analysis and sequencing. The vector was electroporated into the AGL1 strain of *Agrobacterium tumefaciens* and used for transformation.

4.2. Plant Material, Agrobacterium-Mediated Transformation and In-Vitro Culture

The spring cultivar of common wheat (*Triticum aestivum* L.) Kontesa was used as a donor plant for transformation experiments as well as transgenic plants. Seeds were germinated into Petri dishes for one day at 4 °C and then five days at room temperature in the dark. Six out of ten seedlings from each Petri dish were replanted into pots with soil. The plants were grown in a growth chamber under controlled environmental conditions with 20 °C/18 °C day/night temperatures and a 16 h light/8 h dark photoperiod. The light intensity was 350 μmol·s^{-1}·m^{-2}.

Agrobacterium-mediated transformation experiments were performed according to our previously described protocols for wheat [71,72]. Putative transgenic plants were regenerated and selected on modified MS media containing 25 mg·L^{-1} of hygromycin as a selectable agent.

First, 7 days after pollination, (DAP) spikes from T_1, T_2, and control plants were collected for RT-qPCR and phytohormone quantification. Only 1 in 3 of the middle part of each spike was used for experiments (upper and lower parts were removed).

4.3. PCR Analysis

Genomic DNA was isolated from well-developed leaves of 14-day plants according to the modified CTAB procedure [73] or by using the KAPA3G Plant PCR Kit (Roche Sequencing and Life Science, Kapa Biosystems, Wilmington, MA, USA). The PCR for genomic DNA isolated by CTAB was carried out in a 25 mL reaction mixture using Platinum Taq DNA Polymerase (Invitrogen by Thermo Fisher Scientific, Waltham, MA, USA) and 120 ng of template DNA. The reaction was run using the following program: initial denaturation step at 94 °C for 2 min, 35 cycles of amplification at 94 °C for 30 s, 65 °C for 30 s, 72 °C for 30 s with a final extension step at 72 °C for 5 min. The PCR for genomic DNA isolated by KAPA3G was carried out in a 50 μL reaction mixture using 1 U of KAPA3G Plant DNA Polymerase and a 0.5 × 0.5 mm leaf fragment. The reaction was run using the following program:

initial denaturation step at 95 °C for 3 min, 40 cycles of amplification at 95 °C for 20 s, 68 °C for 30 s, 72 °C for 30s with a final extension step at 72 °C for 2 min.

Putative transgenic T_0 and T_1 plants were tested with two pairs of specific primers amplifying a fragment of the *hpt* selection gene. The sequences of the primers for the first pair were: hygF1 5′-ATGACGCACAATCCCACTATCCT-3′ and hygR1 5′-AGTTCGGTTTCAGGCAGGTCTT-3′, and the amplified fragment was 405 bp. The sequences of the primers for the second pair were: hygF2 5′-GACGGCAATTTCGATGATG-3′ and hygR2 5′-CCGGTCGGCATCTACTCTAT-3′, and the amplified fragment was 205 bp.

Non-transgenic null segregants were used as a control.

4.4. RNA Extraction and cDNA Synthesis

Total RNA from 7 DAP spikes was extracted using TRI Reagent (Sigma-Aldrich, Hamburg, Germany) and 1-bromo-3-chloropropane (BCP) (AppliChem GmbH, Darmstadt, Germany) according to the manufacturer’s protocol. The purity and concentration of the isolated RNA were determined using a NanoDrop spectrophotometer (NanoDrop ND-1000) and the integrity was checked by electrophoresis on 1.5% (w/v) agarose gels. To remove the residual DNA the RNA samples were treated with DNase I, RNase-free (Thermo Fisher Scientific, Waltham, MA, USA). Each time 1 µg of good quality RNA was used for cDNA synthesis using the RevertAid First Strand cDNA Synthesis Kit (Thermo Fisher Scientific) following the manufacturer’s instructions. The obtained cDNA was diluted 20 times before use in RT-qPCR assays.

4.5. Quantitative RT-qPCR

RT-qPCR assays were performed for 6 target genes: *TaCKX1* (JN128583), *TaCKX2.1* (JF293079)/*2.2* (FJ648070), *TaCKX11* (*3*) (JN128585), *TaCKX5* (Lei et al., 2008), *TaCKX9* (*10*) (JN128591). Primer sequences designed for each gene as well as for the reference gene are shown in Table S1. All real-time reactions were performed in a Rotor-Gene Q (Qiagen) thermal cycler using 1× HOT FIREPol EvaGreen qPCR Mix Plus (Solis BioDyne), 0.2 µM of each primer, and 4 µL of 20 times diluted cDNA in a total volume of 10 µL. Each reaction was carried out in 3 technical replicates at the following temperature profile: 95 °C—15 min initial denaturation and polymerase activation (95 °C—25 s, 62 °C—25 s, 72 °C—25 s) × 45 cycles, 72 °C—5 min, with the melting curve at 72–99 °C, 5 s per step. The expression of *TaCKX* genes was calculated according to the two standard curves method using ADP-ribosylation factor (*Ref 2*) as a normalizer.

Relative expression/silencing of *TaCKX1* was related to mean expression of the gene in non-silenced control plants set as 1.00. Relative expression of other *TaCKX* genes was related to each tested gene set as 1.00 in non-silenced plants.

Statistical analysis was performed using Statistica v13.3 software (StatSoft, Kraków, Poland). The normality of data distribution was tested using the Shapiro–Wilk test. To determine whether the means of two sets of data of expression levels, phytohormone concentrations, and yield-related traits between non-silenced and silenced lines are significantly different from each other (for *p* value less than $p < 0.05$), either the Student’s *t*-test or the Mann–Whitney test was applied. Correlation coefficients were determined using parametric correlation matrices (Pearson’s test) or a nonparametric correlation (Spearman’s test).

4.6. Quantification of ABA, Auxins, Cytokinins and GA_3

Chemicals used for quantification were: the standard of ABA, five standards of auxins: IAA, indole-3-butyric acid (IBA), indole-3-propionic acid (IPA), 1-naphthaleneacetic acid (NAA), and 2-phenylacetic acid (PAA); twenty-seven standards of CKs: tZ, *trans*-zeatin riboside (tZR), *trans*-zeatin-9-glucoside (tZ9G), *trans*-zeatin-7-glucoside (tZ7G), *trans*-zeatin-*O*-glucoside (tZOG), *trans*-zeatin riboside-*O*-glucoside (tZROG), *trans*-zeatin-9-glucoside-*O*-glucoside (tZ9GOG), *trans*-zeatin-9-glucoside riboside (tZ9GR), *c*Z, *cis*-zeatin-riboside (cZR), *cis*-zeatin *O*-glucoside (cZOG), *cis*-zeatin 9-glucoside (cZ9G), *cis*-zeatin-*O*-glucoside-riboside (cZROG), dihydrozeatin (DZ), dihydrozeatin-riboside (DZR), dihydrozeatin-9-glucoside (DZ9G), dihydrozeatin-7-glucoside (DZ7G), dihydrozeatin-*O*-glucoside (DZOG), dihydrozeatin riboside-*O*-glucoside (DZROG), N^6-isopentenyladenine (iP), N^6-isopentenyladenosine (iPR), N^6-isopentenyladenosine-7-glucoside (iP7G), *para*-topolin (*p*T), *meta*-topolin (*m*T), *ortho*-topolin (*o*T), 6-benzylaminopurine (6-BAP), and standard of GA_3.

For the measurement of phytohormones, 200 mg of plant powders were placed into the 2-mL Eppendorf tubes, suspended in 1 mL of (*v*/*v*) 50% ACN, and homogenized in a bead mill (50 Hz, 5 min) using two 5-mm tungsten balls. Then, samples were homogenized using the ultrasound processor VCX 130 (max. power 130 W, max. frequency 20 kHz, 5 min) equipped with titanium probe and mixed in laboratory shaker (90 rpm, dark, 5 °C, 30 min). Samples were centrifuged (9000× *g*, 5 min) and collected in a glass tube. For the quantification of ABA, AXs, CKs, and GA_3, $[^2H_6]$(+)-*cis*,*trans*-ABA (50 ng), $[^2H_5]$ IAA (15 ng), $[^2H_6]$ iP (50 ng), $[^2H_5]$ *t*Z (30 ng), $[^2H_5]$-*t*ZOG (30 ng), $[^2H_3]$-DZR (30 ng), and $[^2H_2]$ GA_3 (30 ng) were added to samples as internal standards.

Prepared extracts were purged using a Waters SPE Oasis HLB cartridge (Waters Corporation, Milford, MA, USA), previously activated and equilibrated using 1 mL of 100% MeOH, 1 mL water, and 1 mL of (*v*/*v*) 50% ACN [74]. Then, extracts were loaded and collected to the Eppendorf tubes and eluted with 1 mL of 30% ACN (*v*/*v*). Samples were evaporated to dryness by centrifugal vacuum concentrator, dissolved in 50 µL of (*v*/*v*) 30% ACN and transferred into the insert vials. Detection of analyzed phytohormones was performed using an Agilent 1260 Infinity series HPLC system (Agilent Technologies, Santa Clara, CA, USA) including a Q-ToF LC/MS mass spectrometer with Dual AJS ESI source; 10 µL of each sample was injected on the Waters XSelect C_{18} column (250 mm × 3.0 mm, 5 µm), heated up to 50 °C. Mobile phase A was 0.01% (*v*/*v*) FA in ACN and phase B 0.01% (*v*/*v*) FA in water; flow was 0.5 mL min^{-1}. Separation of above hormones was done in ESI-positive mode with the following gradient: 0–8 min flowing increased linearly from 5 to 30% A, 8–25 min 80% A, 25–28 min 100% A, 28–30 min 5% A.

For the optimization of MS/MS conditions, the chemical standards of analyzed phytohormones were directly injected to the MS in positive ($[M + H]^+$) ion scan modes, then areas of detected standard peaks were calculated. $[M + H]^+$ was chosen because of its significantly better signal-to-noise ratios compared to the negative ion scan modes.

Chlorophyll content was measured using an SPAD chlorophyll meter.

5. Conclusions

Based on the 7 DAP spike as a research object, we have documented that silencing of *TaCKX1* by RNAi strongly influenced up- or down-regulation of other *TaCKX* genes, as well as phytohormone levels and consequently phenotype. This co-regulation is dependent on the level of silencing of the gene and is independent of cross-silencing of other *TaCKX* genes. Detailed analysis revealed that each tested yield-related trait is regulated by various up- or down-regulated *TaCKX* genes and phytohormones. Key genes involved in the regulation of grain yield, TGW, or root weight in highly silenced plants are *TaCKX2.1* and *TaCKX11* (*3*) acting antagonistically, and increased expression of the first one determines growth of tZ, tZ derivatives, and cZ, whereas decreased expression of the second down-regulates content of cZOG. A key role in determination of the high-yielding phenotype seemed to be played by the growing content of tZ in 7 DAP spikes, which might accelerate maturation of immature grains by speeding up nutrient flow from flag leaves. This finally led to reduction of TGW but enhancement of grain number and yield. The latter traits are the result of a higher spike number, which is determined in the early stages of plant development.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/21/13/4809/s1. Table S1: Primer sequences designed for reference gene and each of 6 tested *TaCKX* genes and amplicon length. Table S2: Phenotypic traits and ratio indicator in silent T_1 and not silent, control plants. Table S3: Phenotypic traits and ratio indicator in silent T_2 and not silent, control plants. Table S4: A. B. Correlation coefficients among expression of all tested *TaCKX* genes and enzyme activity, and phenotypic traits in not-silent (A) and highly silent T_2 plants (B). * non-parametric analysis; in bold: significant at $p < 0.01$.

Author Contributions: Conceptualization, A.N.-O. and W.O.; methodology, B.J., H.O., K.S., and A.B.; software, H.O. and K.S.; validation, B.J.; formal analysis, B.J.; investigation, B.J., H.O., and K.S.; data curation, B.J. and A.B.; writing—original draft preparation, A.N.-O.; writing—review and editing, A.N.-O.; visualization, A.N.-O. and H.O.; supervision, A.N.-O. and W.O.; project administration, A.N.-O.; funding acquisition, A.N.-O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Centre, grant UMO-2014/13/B/NZ9/02376 and a statutory grant of PBAI-NRI.

Acknowledgments: We thank Malgorzata Wojciechowska, Izabela Skuza, Agnieszka Glowacka, Maja Boczkowska, and Agnieszka Onysk for excellent technical assistance.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Reynolds, M.; Braun, H. Wheat breeding benefits to low-input agriculture. *Nat. Plants* **2019**, *5*, 652–653. [CrossRef] [PubMed]
2. Borrill, P.; Harrington, S.A.; Uauy, C. Applying the latest advances in genomics and phenomics for trait discovery in polyploid wheat. *Plant J.* **2019**, *97*, 56–72. [CrossRef] [PubMed]
3. Nadolska-Orczyk, A.; Rajchel, I.K.; Orczyk, W.; Gasparis, S. Major genes determining yield-related traits in wheat and barley. *Theor. Appl. Genet.* **2017**, *130*, 1081–1098. [CrossRef] [PubMed]
4. Foley, J.A.; Ramankutty, N.; Brauman, K.A.; Cassidy, E.S.; Gerber, J.S.; Johnston, M.; Mueller, N.D.; O'Connell, C.; Ray, D.K.; West, P.C.; et al. Solutions for a cultivated planet. *Nature* **2011**, *478*, 337–342. [CrossRef] [PubMed]
5. Kieber, J.J.; Schaller, G.E. Cytokinin signaling in plant development. *Development* **2018**, *478*, 337–342. [CrossRef] [PubMed]
6. Kim, K.; Ryu, H.; Cho, Y.H.; Scacchi, E.; Sabatini, S.; Hwang, I. Cytokinin-facilitated proteolysis of ARABIDOPSIS RESPONSE REGULATOR 2 attenuates signaling output in two-component circuitry. *Plant J.* **2012**, *69*, 934–945. [CrossRef]
7. Cerny, M.; Dycka, F.; Bobalova, J.; Brzobohaty, B. Early cytokinin response proteins and phosphoproteins of *Arabidopsis thaliana* identified by proteome and phosphoproteome profiling. *J. Exp. Bot.* **2011**, *62*, 921–937. [CrossRef]
8. Potter, K.C.; Wang, J.; Schaller, G.E.; Kieber, J.J. Cytokinin modulates context-dependent chromatin accessibility through the type-B response regulators. *Nat. Plants* **2018**, *4*, 1102–1111. [CrossRef]
9. Brenner, W.G.; Ramireddy, E.; Heyl, A.; Schmulling, T. Gene regulation by cytokinin in Arabidopsis. *Front. Plant Sci.* **2012**, *3*, 8. [CrossRef]
10. Jameson, P.E.; Song, J.C. Cytokinin: A key driver of seed yield. *J. Exp. Bot.* **2016**, *67*, 593–606. [CrossRef]
11. Liu, Z.; Lv, Y.; Zhang, M.; Liu, Y.; Kong, L.; Zou, M.; Lu, G.; Cao, J.; Yu, X. Identification, expression, and comparative genomic analysis of the *IPT* and *CKX* gene families in Chinese cabbage (Brassica rapa ssp. pekinensis). *BMC Genom.* **2013**, *14*, 594. [CrossRef] [PubMed]
12. Zhang, J.; Yu, G.H.; Wen, W.W.; Ma, X.Q.; Xu, B.; Huang, B.R. Functional characterization and hormonal regulation of the PHEOPHYTINASE gene *LpPPH* controlling leaf senescence in perennial ryegrass. *J. Exp. Bot.* **2016**, *67*, 935–945. [CrossRef] [PubMed]
13. Panda, B.B.; Sekhar, S.; Dash, S.K.; Behera, L.; Shaw, B.P. Biochemical and molecular characterisation of exogenous cytokinin application on grain filling in rice. *BMC Plant Biol.* **2018**, *18*, 89. [CrossRef] [PubMed]
14. Sakakibara, H. Cytokinins: Activity, biosynthesis, and translocation. *Annu. Rev. Plant Biol.* **2006**, *57*, 431–449. [CrossRef] [PubMed]
15. Bajguz, A.; Piotrowska, A. Conjugates of auxin and cytokinin. *Phytochemistry* **2009**, *70*, 957–969. [CrossRef]
16. Brandizzi, F. Divide, expand, differentiate—New insights on plant organ growth through cytokinin signaling. *Plant J.* **2019**, *97*, 803–804. [CrossRef]
17. Kudo, T.; Kiba, T.; Sakakibara, H. Metabolism and Long-distance Translocation of Cytokinins. *J. Integr. Plant Biol.* **2010**, *52*, 53–60. [CrossRef]

18. Brugière, N.; Jiao, S.; Hantke, S.; Zinselmeier, C.; Roessler, J.A.; Niu, X.; Jones, R.J.; Habben, J.E. Cytokinin oxidase gene expression in maize is localized to the vasculare, and is induced by cytokinins, abscisic acid, and abiotic stress. *Plant Physiol.* **2003**, *132*, 1228–1240. [CrossRef]
19. Ashikari, M.; Sakakibara, H.; Lin, S.Y.; Yamamoto, T.; Takashi, T.; Nishimura, A.; Angeles, E.R.; Qian, Q.; Kitano, H.; Matsuoka, M. Cytokinin oxidase regulates rice grain production. *Science* **2005**, *309*, 741–745. [CrossRef]
20. Zalewski, W.; Galuszka, P.; Gasparis, S.; Orczyk, W.; Nadolska-Orczyk, A. Silencing of the *HvCKX1* gene decreases the cytokinin oxidase/dehydrogenase level in barley and leads to higher plant productivity. *J. Exp. Bot.* **2010**, *61*, 1839–1851. [CrossRef]
21. Zalewski, W.; Gasparis, S.; Boczkowska, M.; Rajchel, I.K.; Kala, M.; Orczyk, W.; Nadolska-Orczyk, A. Expression patterns of *HvCKX* genes indicate their role in growth and reproductive development of barley. *PLoS ONE* **2014**, *9*, e115729. [CrossRef] [PubMed]
22. Chen, L.; Zhao, J.Q.; Song, J.C.; Jameson, P.E. Cytokinin dehydrogenase: A genetic target for yield improvement in wheat. *Plant Biotechnol. J.* **2020**, *18*, 614–630. [CrossRef]
23. Ogonowska, H.; Barchacka, K.; Gasparis, S.; Jablonski, B.; Orczyk, W.; Dmochowska-Boguta, M.; Nadolska-Orczyk, A. Specificity of expression of *TaCKX* family genes in developing plants of wheat and their co-operation within and among organs. *PLoS ONE* **2019**, *14*, e0214239. [CrossRef]
24. Zalewski, W.; Orczyk, W.; Gasparis, S.; Nadolska-Orczyk, A. *HvCKX2* gene silencing by biolistic or *Agrobacterium*-mediated transformation in barley leads to different phenotypes. *BMC Plant Biol.* **2012**, *12*, 206. [CrossRef] [PubMed]
25. Holubova, K.; Hensel, G.; Vojta, P.; Tarkowski, P.; Bergougnoux, V.; Galuszka, P. Modification of barley plant productivity through regulation of cytokinin content by reverse-genetics approaches. *Front. Plant Sci.* **2018**, *9*, 1676. [CrossRef] [PubMed]
26. Li, Y.L.; Song, G.Q.; Gao, J.; Zhang, S.J.; Zhang, R.Z.; Li, W.; Chen, M.L.; Liu, M.; Xia, X.C.; Risacher, T.; et al. Enhancement of grain number per spike by RNA interference of *cytokinin oxidase 2* gene in bread wheat. *Hereditas* **2018**, *155*, 33. [CrossRef] [PubMed]
27. Gasparis, S.; Przyborowski, M.; Kala, M.; Nadolska-Orczyk, A. Knockout of the *HvCKX1* or *HvCKX3* gene in barley (*Hordeum vulgare* L.) by RNA-Guided Cas9 Nuclease affects the regulation of cytokinin metabolism and root morphology. *Cells* **2019**, *8*, 782. [CrossRef]
28. Lu, J.; Chang, C.; Zhang, H.P.; Wang, S.X.; Sun, G.; Xiao, S.H.; Ma, C.X. Identification of a Novel Allele of *TaCKX6a02* Associated with Grain Size, Filling Rate and Weight of Common Wheat. *PLoS ONE* **2015**, *10*, e0144765. [CrossRef] [PubMed]
29. Zhang, L.; Zhao, Y.L.; Gao, L.F.; Zhao, G.Y.; Zhou, R.H.; Zhang, B.S.; Jia, J.Z. *TaCKX6-D1*, the ortholog of rice *OsCKX2*, is associated with grain weight in hexaploid wheat. *New Phytol.* **2012**, *195*, 574–584. [CrossRef]
30. Chang, C.; Lu, J.; Zhang, H.P.; Ma, C.X.; Sun, G.L. Copy Number Variation of Cytokinin Oxidase Gene *Tackx4* Associated with Grain Weight and Chlorophyll Content of Flag Leaf in Common Wheat. *PLoS ONE* **2015**, *10*, 15. [CrossRef]
31. Kersey, P.J.; Allen, J.E.; Allot, A.; Barba, M.; Boddu, S.; Bolt, B.J.; Carvalho-Silva, D.; Christensen, M.; Davis, P.; Grabmueller, C.; et al. Ensembl Genomes 2018: An integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.* **2018**, *46*, D802–D808. [CrossRef] [PubMed]
32. Uauy, C. The high grain protein content gene *Gpc-B1* accelerates senescence and has pleiotropic effects on protein content in wheat. *J. Exp. Bot.* **2006**, *57*, 2785–2794. [CrossRef] [PubMed]
33. Travella, S.; Klimm, T.E.; Keller, B. RNA interference-based gene silencing as an efficient tool for functional genomics in hexaploid bread wheat. *Plant Physiol.* **2006**, *142*, 6–20. [CrossRef] [PubMed]
34. Gasparis, S.; Orczyk, W.; Zalewski, W.; Nadolska-Orczyk, A. The RNA-mediated silencing of one of the *Pin* genes in allohexaploid wheat simultaneously decreases the expression of the other, and increases grain hardness. *J. Exp. Bot.* **2011**, *62*, 4025–4036. [CrossRef] [PubMed]
35. Gao, X.P.; Francis, D.; Ormrod, J.C.; Bennett, M.D. Changes in cell number and cell-division activity during endosperm development in allohexaploid wheat, Triticum-aestivum L. *J. Exp. Bot.* **1992**, *43*, 1603–1609. [CrossRef]
36. Hess, J.R.; Carman, J.G.; Banowetz, G.M. Hormones in wheat kernels during embryony. *J. Plant Physiol.* **2002**, *159*, 379–386. [CrossRef]

37. Gajdosova, S.; Spichal, L.; Kaminek, M.; Hoyerova, K.; Novak, O.; Dobrev, P.I.; Galuszka, P.; Klima, P.; Gaudinova, A.; Zizkova, E.; et al. Distribution, biological activities, metabolism, and the conceivable function of cis-zeatin-type cytokinins in plants. *J. Exp. Bot.* **2011**, *62*, 2827–2840. [CrossRef]
38. Morris, R.O.; Blevins, D.G.; Dietrich, J.T.; Durley, R.C.; Gelvin, S.B.; Gray, J.; Hommes, N.G.; Kaminek, M.; Mathews, L.J.; Meilan, R.; et al. Cytokinins in plant-pathogenic bacteria and developing cereal-grains. *Funct. Plant Physiol.* **1993**, *20*, 621–637. [CrossRef]
39. Bartrina, I.; Otto, E.; Strnad, M.; Werner, T.; Schmulling, T. Cytokinin regulates the activity of reproductive meristems, flower organ size, ovule formation, and thus seed yield in Arabidopsis thaliana. *Plant Cell* **2011**, *23*, 69–80. [CrossRef]
40. Schafer, M.; Brutting, C.; Meza-Canales, I.D.; Grosskinsky, D.K.; Vankova, R.; Baldwin, I.T.; Meldau, S. The role of cis-zeatin-type cytokinins in plant growth regulation and mediating responses to environmental interactions. *J. Exp. Bot.* **2015**, *66*, 4873–4884. [CrossRef]
41. Bilyeu, K.D.; Cole, J.L.; Laskey, J.G.; Riekhof, W.R.; Esparza, T.J.; Kramer, M.D.; Morris, R.O. Molecular and biochemical characterization of a cytokinin oxidase from maize. *Plant Physiol.* **2001**, *125*, 378–386. [CrossRef] [PubMed]
42. Saleem, M.; Lamkemeyer, T.; Schutzenmeister, A.; Madlung, J.; Sakai, H.; Piepho, H.P.; Nordheim, A.; Hochholdinger, F. Specification of Cortical Parenchyma and Stele of Maize Primary Roots by Asymmetric Levels of Auxin, Cytokinin, and Cytokinin-Regulated Proteins. *Plant Physiol.* **2010**, *152*, 4–18. [CrossRef] [PubMed]
43. Zalabak, D.; Galuszka, P.; Mrizova, K.; Podlesakova, K.; Gu, R.L.; Frebortova, J. Biochemical characterization of the maize cytokinin dehydrogenase family and cytokinin profiling in developing maize plantlets in relation to the expression of cytokinin dehydrogenase genes. *Plant Physiol. Biochem.* **2014**, *74*, 283–293. [CrossRef]
44. Powell, A.E.; Paleczny, A.R.; Olechowski, H.; Emery, R.J.N. Changes in cytokinin form and concentration in developing kernels correspond with variation in yield among field-grown barley cultivars. *Plant Physiol. Biochem.* **2013**, *64*, 33–40. [CrossRef] [PubMed]
45. Zd'arska, M.; Zatloukalova, P.; Benitez, M.; Sedo, O.; Potesil, D.; Novak, O.; Svacinova, J.; Pesek, B.; Malbeck, J.; Vasickova, J.; et al. Proteome analysis in Arabidopsis reveals shoot- and root-specific targets of cytokinin action and differential regulation of hormonal homeostasis. *Plant Physiol.* **2013**, *161*, 918–930. [CrossRef]
46. Polanska, L.; Vicankova, A.; Novakova, M.; Malbeck, J.; Dobrev, P.I.; Brzobohaty, B.; Vankova, R.; Machackova, I. Altered cytokinin metabolism affects cytokinin, auxin, and abscisic acid contents in leaves and chloroplasts, and chloroplast ultrastructure in transgenic tobacco. *J. Exp. Bot.* **2007**, *58*, 637–649. [CrossRef]
47. Cerny, M.; Kuklova, A.; Hoehenwarter, W.; Fragner, L.; Novak, O.; Rotkova, G.; Jedelsky, P.L.; Zakova, K.; Smehilova, M.; Strnad, M.; et al. Proteome and metabolome profiling of cytokinin action in Arabidopsis identifying both distinct and similar responses to cytokinin down- and up-regulation. *J. Exp. Bot.* **2013**, *64*, 4193–4206. [CrossRef]
48. Kiba, T.; Kudo, T.; Kojima, M.; Sakakibara, H. Hormonal control of nitrogen acquisition: Roles of auxin, abscisic acid, and cytokinin. *J. Exp. Bot.* **2011**, *62*, 1399–1409. [CrossRef]
49. Abid, M.; Shao, Y.H.; Liu, S.X.; Wang, F.; Gao, J.W.; Jiang, D.; Tian, Z.W.; Dai, T.B. Pre-drought priming sustains grain development under post-anthesis drought stress by regulating the growth hormones in winter wheat (*Triticum aestivum* L.). *Planta* **2017**, *246*, 509–524. [CrossRef]
50. Perilli, S.; Moubayidin, L.; Sabatini, S. The molecular basis of cytokinin function. *Curr. Opin. Plant Biol.* **2010**, *13*, 21–26. [CrossRef]
51. Miyawaki, K.; Matsumoto-Kitano, M.; Kakimoto, T. Expression of cytokinin biosynthetic isopentenyltransferase genes in Arabidopsis: Tissue specificity and regulation by auxin, cytokinin, and nitrate. *Plant J.* **2004**, *37*, 128–138. [CrossRef] [PubMed]
52. Shimizu-Sato, S.; Tanaka, M.; Mori, H. Auxin-cytokinin interactions in the control of shoot branching. *Plant Mol. Biol.* **2009**, *69*, 429–435. [CrossRef] [PubMed]
53. Osugi, A.; Sakakibara, H. Q&A: How do plants respond to cytokinins and what is their importance? *BMC Biol.* **2015**, *13*, 1–10.

54. Martin, R.C.; Mok, M.C.; Habben, J.E.; Mok, D.W.S. A maize cytokinin gene encoding an O-glucosyltransferase specific to cis-zeatin. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 5922–5926. [CrossRef] [PubMed]
55. Chen, L.; Du, Y.Y.; Lu, Q.M.; Chen, H.; Meng, R.S.; Cui, C.G.; Lu, S.; Yang, Y.; Chai, Y.M.; Li, J.; et al. The Photoperiod-Insensitive Allele *Ppd-D1a* Promotes Earlier Flowering in *Rht12* Dwarf Plants of Bread Wheat. *Front. Plant Sci.* **2018**, *9*, 1312. [CrossRef]
56. Guo, Z.F.; Liu, G.Z.; Roder, M.S.; Reif, J.C.; Ganal, M.W.; Schnurbusch, T. Genome-wide association analyses of plant growth traits during the stem elongation phase in wheat. *Plant Biotechnol. J.* **2018**, *16*, 2042–2052. [CrossRef] [PubMed]
57. Fahy, B.; Siddiqui, H.; David, L.C.; Powers, S.J.; Borrill, P.; Uauy, C.; Smith, A.M. Final grain weight is not limited by the activity of key starch-synthesising enzymes during grain filling in wheat. *J. Exp. Bot.* **2018**, *69*, 5461–5475. [CrossRef] [PubMed]
58. Cai, T.; Meng, X.P.; Liu, X.L.; Liu, T.N.; Wang, H.; Jia, Z.K.; Yang, D.Q.; Ren, X.L. Exogenous hormonal application regulates the occurrence of wheat tillers by changing endogenous hormones. *Front. Plant Sci.* **2018**, *9*, 1886. [CrossRef] [PubMed]
59. Pacifici, E.; Polverari, L.; Sabatini, S. Plant hormone cross-talk: The pivot of root growth. *J. Exp. Bot.* **2015**, *66*, 1113–1121. [CrossRef]
60. Corbesier, L.; Prinsen, E.; Jacqmard, A.; Lejeune, P.; Van Onckelen, H.; Perilleux, C.; Bernier, G. Cytokinin levels in leaves, leaf exudate and shoot apical meristem of Arabidopsis thaliana during floral transition. *J. Exp. Bot.* **2003**, *54*, 2511–2517. [CrossRef]
61. Hirose, N.; Takei, K.; Kuroha, T.; Kamada-Nobusada, T.; Hayashi, H.; Sakakibara, H. Regulation of cytokinin biosynthesis, compartmentalization and translocation. *J. Exp. Bot.* **2008**, *59*, 75–83. [CrossRef] [PubMed]
62. Reid, D.E.; Heckmann, A.B.; Novak, O.; Kelly, S.; Stougaard, J. CYTOKININ OXIDASE/DEHYDROGENASE3 Maintains Cytokinin Homeostasis during Root and Nodule Development in Lotus japonicus. *Plant Physiol.* **2016**, *170*, 1060–1074. [CrossRef] [PubMed]
63. Mao, C.J.; He, J.M.; Liu, L.N.; Deng, Q.M.; Yao, X.F.; Liu, C.M.; Qiao, Y.L.; Li, P.; Ming, F. *OsNAC2* integrates auxin and cytokinin pathways to modulate rice root development. *Plant Biotechnol. J.* **2019**, *18*, 429–442. [CrossRef] [PubMed]
64. Bishopp, A.; Lehesranta, S.; Vaten, A.; Help, H.; El-Showk, S.; Scheres, B.; Helariutta, K.; Mahonen, A.P.; Sakakibara, H.; Helariutta, Y. Phloem-transported cytokinin regulates polar auxin transport and maintains vascular pattern in the root meristem. *Curr. Biol.* **2011**, *21*, 927–932. [CrossRef] [PubMed]
65. Brenner, W.G.; Schmulling, T. Transcript profiling of cytokinin action in Arabidopsis roots and shoots discovers largely similar but also organ-specific responses. *BMC Plant Biol.* **2012**, *12*, 112. [CrossRef]
66. Kollmer, I.; Novak, O.; Strnad, M.; Schmulling, T.; Werner, T. Overexpression of the cytosolic cytokinin oxidase/dehydrogenase (*CKX7*) from Arabidopsis causes specific changes in root growth and xylem differentiation. *Plant J.* **2014**, *78*, 359–371. [CrossRef]
67. Gan, S.S.; Amasino, R.M. Inhibition of Leaf senescence by autoregulated production of cytokinin. *Science* **1995**, *270*, 1986–1988. [CrossRef]
68. Gregersen, P.L.; Holm, P.B.; Krupinska, K. Leaf senescence and nutrient remobilisation in barley and wheat. *Plant Biol.* **2008**, *10*, 37–49. [CrossRef]
69. Behr, M.; Motyka, V.; Weihmann, F.; Malbeck, J.; Deising, H.B.; Wirsel, S.G.R. Remodeling of Cytokinin Metabolism at Infection Sites of Colletotrichum graminicola on Maize Leaves. *Mol. Plant-Microbe Interact.* **2012**, *25*, 1073–1082. [CrossRef]
70. Wang, W.Q.; Hao, Q.Q.; Wang, W.L.; Li, Q.X.; Chen, F.J.; Ni, F.; Wang, Y.; Fu, D.L.; Wu, J.J.; Wang, W. The involvement of cytokinin and nitrogen metabolism in delayed flag leaf senescence in a wheat stay-green mutant, tasg1. *Plant Sci.* **2019**, *278*, 70–79. [CrossRef]
71. Przetakiewicz, A.; Orczyk, W.; Nadolska-Orczyk, A. The effect of auxin on plant regeneration of wheat, barley and triticale. *Plant Cell Tissue Organ Cult.* **2003**, *73*, 245–256. [CrossRef]
72. Przetakiewicz, A.; Karas, A.; Orczyk, W.; Nadolska-Orczyk, A. Agrobacterium-mediated transformation of polyploid cereals. The efficiency of selection and transgene expression in wheat. *Cell. Mol. Biol. Lett.* **2004**, *9*, 903–917. [PubMed]

73. Murray, M.G.; Thompson, W.F. Rapid Isolation of High Molecular-Weight Plant DNA. *Nucleic Acids Res.* **1980**, *8*, 4321–4325. [CrossRef] [PubMed]
74. Simura, J.; Antoniadi, I.; Siroka, J.; Tarkowska, D.; Strnad, M.; Ljung, K.; Novak, O. Plant hormonomics: Multiple phytohormone profiling by targeted metabolomics. *Plant Physiol.* **2018**, *177*, 476–489. [CrossRef] [PubMed]

International Journal of
Molecular Sciences

MDPI

Article

Evolutionary and Predictive Functional Insights into the Aquaporin Gene Family in the Allotetraploid Plant *Nicotiana tabacum*

Jahed Ahmed, Sébastien Mercx, Marc Boutry and François Chaumont *

Louvain Institute of Biomolecular Science and Technology, UCLouvain, Croix du Sud 4-L7.07.14, B-1348 Louvain-la-Neuve, Belgium; jahed.ahmed@uclouvain.be (J.A.); mercxsebastien@gmail.com (S.M.); marc.boutry@uclouvain.be (M.B.)

* Correspondence: francois.chaumont@uclouvain.be

Received: 8 June 2020; Accepted: 1 July 2020; Published: 3 July 2020

Abstract: Aquaporins (AQPs) are a class of integral membrane proteins that facilitate the membrane diffusion of water and other small solutes. *Nicotiana tabacum* is an important model plant, and its allotetraploid genome has recently been released, providing us with the opportunity to analyze the *AQP* gene family and its evolution. A total of 88 full-length *AQP* genes were identified in the *N. tabacum* genome, and the encoding proteins were assigned into five subfamilies: 34 plasma membrane intrinsic proteins (PIPs); 27 tonoplast intrinsic proteins (TIPs); 20 nodulin26-like intrinsic proteins (NIPs); 3 small basic intrinsic proteins (SIPs); 4 uncharacterized X intrinsic proteins (XIPs), including two splice variants. We also analyzed the genomes of two *N. tabacum* ancestors, *Nicotiana tomentosiformis* and *Nicotiana sylvestris,* and identified 49 *AQP* genes in each species. Functional prediction, based on the substrate specificity-determining positions (SDPs), revealed significant differences in substrate specificity among the AQP subfamilies. Analysis of the organ-specific *AQP* expression levels in the *N. tabacum* plant and RNA-seq data of *N. tabacum* bright yellow-2 suspension cells indicated that many AQPs are simultaneously expressed, but differentially, according to the organs or the cells. Altogether, these data constitute an important resource for future investigations of the molecular, evolutionary, and physiological functions of AQPs in *N. tabacum*.

Keywords: aquaporins; bright yellow-2 suspension cells; *Nicotiana tabacum*; substrate specificity; phylogeny

1. Introduction

Aquaporins (AQPs), also known as major intrinsic proteins (MIPs), are small integral membrane proteins present in almost all living organisms [1,2]. Plants maintain a large and diverse AQP family compared to mammals. For instance, the genomes of rice (*Oryza sativa)*, Arabidopsis (*Arabidopsis thaliana)*, maize (*Zea mays*), soybean (*Glycine max*), switchgrass (*Panicum virgatum*), foxtail millet (*Setaria italica*), sorghum (*Sorghum bicolor*), *Brachypodium distachyon,* tomato (*Solanum lycopersicum*), poplar (*Populus trichocarpa*), cotton (*Gossypium hirsutum*), and potato (*Solanum tuberosum*) encode 39, 35, 36, 66, 68, 42, 38, 28, 47, 55, 71, and 41 AQP homologs, respectively [3–12], compared to only 13 *AQP* genes in mammals [13]. Based on phylogenetic analysis and subcellular localization, vascular plant AQPs are categorized into five subfamilies: (1) plasma membrane intrinsic proteins (PIPs); (2) tonoplast intrinsic proteins (TIPs); (3) nodulin-26-like intrinsic proteins (NIPs); (4) small basic intrinsic proteins (SIPs); (5) uncharacterized X intrinsic proteins (XIPs). To date, the latter subfamily has not been found in Brassicaceae and in monocots [14,15].

While many plant AQPs primarily function as water channels, they can also transport a wide range of substrates, such as ammonia (NH_3), antimony (Sb), arsenic (As), boron (B), glycerol, hydrogen peroxide (H_2O_2), silicon (Si), and urea (U) [2,16–19]. Furthermore, some AQPs facilitate gas diffusion, such as carbon dioxide (CO_2) and oxygen (O_2) [20–22]. Recently it was reported that AtPIP2;1 has cations (Na^+ and K^+) channel activity [23]. AQPs from various plants are also involved in transmembrane water conductance in numerous physiological processes, such as cell water homeostasis, root water uptake from the soil, root and leaf hydraulic conductance, lateral root emergence, motor cell movement, rapid internode elongation, the diurnal regulation of leaf movements, and petal development and movement [1,2,24–29].

The AQP structure comprises six transmembrane (TM) α-helices (TM1-TM6), which are linked by five loops (loops A–E) and two highly conserved NPA (Asn-Pro-Ala) motifs. They form homo- or hetero-tetrameric complexes in which each subunit acts as a functional water channel [2,30]. The channel pore contains two constriction regions that contribute to the transport selectivity. The first constriction is formed at the pore center by two highly conserved NPA motifs [31]. The second constriction is the aromatic/arginine (ar/R) filter, formed at the extracellular aperture of the pore by four residues from TM2, TM5, and loop E (LE1 and LE2), respectively [32,33]. Additionally, five amino acid residues known as Froger's positions (FPs) designated P1–P5, are also associated with substrate selectivity [34,35]. More recently, some substrate specificity determining positions (SDPs) have been proposed for B, H_2O_2, CO_2, NH_3, As, Sb, and Si [9,17].

Nicotiana tabacum (tobacco), a perennial herbaceous plant of the Solanaceae family, is an allotetraploid (2n = 4x = 48), which evolved by the natural hybridization of the ancestors of *Nicotiana sylvestris* (2n = 24, maternal donor) and *Nicotiana tomentosiformis* (2n = 24, paternal donor) about 200,000 years ago [36,37]. *N. tabacum* is intensively studied as a versatile model organism for understanding genetics, functional genomics, cellular and molecular biology, biochemistry and physiology [38]. In this study, we identified *AQP* genes in the genomes of *N. tabacum* as well as its two ancestors, *N. tomentosiformis* and *N. sylvestris*, and analyzed the transcriptome data of *N. tabacum* plant and Bright Yellow-2 (BY-2) suspension cells [39]. We investigated the phylogenetic relationships, as well as the structural properties and subcellular localization of AQPs in *N. tabacum*. Comparing the primary selectivity motifs, we further predicted their probable substrate transport activities. Altogether, this study provides new insights into the expression patterns in different organs and suspension cells, as well as the transmembrane transport selectivity of AQPs in *N. tabacum*.

2. Results

2.1. Genome-Wide Identification and Characterization of NtAQP Genes

The whole genome shotgun sequence of *N. tabacum* and its two ancestors, *N. tomentosiformis* and *N. sylvestris*, were searched for *AQP* genes, using pBLAST and AQP sequences from *S. tuberosum* and *S. lycopersicum* as queries. NtAQP protein sequences were analyzed and compared with SlAQP and StAQP for domain identification and functional annotation. Of 101 initial unique hits for *NtAQPs*, 13 were considered *AQP* pseudogenes and discarded after a manual inspection of their nucleotide and amino acid sequences and their TM domains. We finally obtained 88 genes encoding 90 full-length AQP proteins, and *NtXIP1;1* and *NtXIP1;2* genes encoding two splice variants (α and β), as shown in Table 1.

Table 1. Aquaporin genes in the *N. tabacum* genome.

Gene Name	Accession Number	IP [1]/MW (kDa)	Amino Acid Number	Predicted Subcellular Localization [2]
NtPIP1;1	NP_001313131.1	8.83/30.70	286	PM, C
NtPIP1;2	XP_016508253.1	8.30/30.76	285	PM
NtPIP1;3	AAB04757.1	9.08/30.58	287	PM
NtPIP1;4	NP_001312189.1	8.30/30.80	287	PM

Table 1. *Cont.*

Gene Name	Accession Number	IP [1]/MW (kDa)	Amino Acid Number	Predicted Subcellular Localization [2]
NtPIP1;5	CAA04750.1	8.29/30.82	287	PM
NtPIP1;6	XP_016476491.1	8.99/30.90	287	PM
NtPIP1;7	NP_001312824.1	8.61/30.84	287	PM
NtPIP1;8	NP_001312222.1	8.83/30.82	286	PM
NtPIP1;9	NP_001312921.1	8.96/30.49	284	PM
NtPIP1;10	XP_016458231.1	9.10/30.74	288	PM
NtPIP1;11	XP_016515710.1	8.23/27.31	254	PM
NtPIP1;12	NP_001312721.1	8.99/30.77	287	PM
NtPIP1;13	XP_016510215.1	9.00/30.64	285	PM
NtPIP2;1	AAL33586.1	9.05/30.49	268	PM, C
NtPIP2;2	NP_001313091.1	9.05/30.47	268	PM
NtPIP2;3	NP_001312414.1	9.04/30.48	268	PM
NtPIP2;4	NP_001312350.1	8.87/30.41	283	PM
NtPIP2;5	NP_001312874.1	8.89/30.39	283	PM
NtPIP2;6	XP_016477641.1	9.02/28.49	284	PM
NtPIP2;7	NP_001313061.1	8.98/28.63	284	PM
NtPIP2;8	XP_016476355.1	9.17/28.49	284	PM, C
NtPIP2;9	NP_001312511.1	8.84/30.37	283	PM
NtPIP2;10	XP_016494749.1	8.63/30.32	283	PM
NtPIP2;11	NP_001311701.1	8.19/30.49	285	PM
NtPIP2;12	NP_001312276.1	7.62/30.48	285	PM
NtPIP2;13	NP_001312334.1	6.94/31.23	291	PM
NtPIP2;14	XP_016486700.1	6.94/31.21	291	PM
NtPIP2;15	NP_001312333.1	7.62/30.26	283	PM
NtPIP2;16	XP_016513533.1	7.62/30.30	283	PM
NtPIP2;17	NP_001312464.1	8.21/30.73	287	PM
NtPIP2;18	NP_001313066.1	8.20/30.78	287	PM
NtPIP2;19	NP_001313208.1	7.04/30.16	283	PM
NtPIP2;20	NP_001311719.1	7.04/30.73	287	PM
NtPIP2;21	NP_001311765.1	7.69/30.68	287	PM
NtTIP1;1	BAF95576.1	5.55/25.79	252	PM
NtTIP1;2	NP_001312131.1	5.70/25.80	252	PM, V
NtTIP1;3	NP_001312871.1	5.70/25.73	248	PM
NtTIP1;4	XP_016513281.1	5.91/26.19	248	PM
NtTIP1;5	XP_016501711.1	5.37/25.91	248	PM
NtTIP1;6	XP_016487055.1	5.37/25.90	251	PM
NtTIP1;7	XP_016471957.1	6.04/25.56	251	PM
NtTIP1;8	XP_016495978.1	5.62/25.12	251	PM, V
NtTIP1;9	XP_016450483.1	5.89/25.25	251	PM
NtTIP2;1	NP_001312646.1	5.35/24.94	248	PM, V
NtTIP2;2	XP_016495734.1	5.35/24.99	248	PM
NtTIP2;3	XP_016503582.1	6.00/25.07	248	PM

Table 1. *Cont.*

Gene Name	Accession Number	IP [1]/MW (kDa)	Amino Acid Number	Predicted Subcellular Localization [2]
NtTIP2;4	XP_016480756.1	5.67/25.01	248	PM
NtTIP2;5	XP_016515893.1	5.67/25.02	248	PM
NtTIP2;6	XP_016445220.1	4.85/25.36	250	V
NtTIP2;7	XP_016481958.1	4.85/25.30	250	V
NtTIP2;8	NP_001312940.1	5.66/25.23	250	V
NtTIP2;9	XP_016481922.1	5.66/25.24	250	V
NtTIP2;10	P24422.2	5.32/25.22	250	V
NtTIP3;1	XP_016491554.1	7.07/27.62	260	PM
NtTIP3;2	XP_016491898.1	8.08/27.58	260	PM
NtTIP3;3	XP_016436583.1	7.07/27.41	259	PM
NtTIP3;4	XP_016500896.1	7.07/27.40	259	PM
NtTIP4;1	NP_001311953.1	5.79/25.96	247	V
NtTIP4;2	XP_016441470.1	5.79/25.98	247	V
NtTIP5;1	XP_016462485.1	7.78/25.63	250	PM
NtTIP5;2	XP_016485861.1	7.78/25.59	250	PM
NtNIP1;1	XP_016487110.1	9.08/30.67	288	PM
NtNIP1;2	XP_016445609.1	9.41/32.65	303	PM
NtNIP2;1	XP_016451246.1	8.96/30.49	286	PM
NtNIP3;1	XP_016460638.1	8.29/37.69	337	PM
NtNIP3;2	XP_016515586.1	8.29/37.91	347	PM
NtNIP4;1	XP_016486634.1	8.52/29.73	281	V
NtNIP4;2	XP_016455585.1	8.83/29.12	275	V
NtNIP4;3	XP_016491262.1	7.74/28.43	270	PM
NtNIP4;4	XP_016453373.1	6.89/28.67	271	PM
NtNIP4;5	XP_016456203.1	8.28/29.07	272	PM
NtNIP4;6	XP_016500017.1	7.69/29.16	272	PM
NtNIP5;1	XP_016470302.1	8.63/30.98	297	V
NtNIP5;2	NP_001312819.1	8.87/30.91	297	V
NtNIP5;3	XP_016493176.1	9.86/31.94	304	PM
NtNIP6;1	XP_016435920.1	8.73/34.50	331	V
NtNIP6;2	XP_016438237.1	8.66/32.35	313	PM
NtNIP7;1	XP_016509644.1	7.71/29.58	280	PM
NtNIP7;2	XP_016496646.1	7.78/31.18	293	PM
NtNIP8;1	XP_016468207.1	8.78/29.88	277	V
NtNIP8;2	XP_016451938.1	9.22/34.00	314	PM
NtSIP1;1	XP_016439604.1	9.22/25.06	238	PM
NtSIP1;2	XP_016492107.1	9.55/25.94	242	PM
NtSIP2;1	XP_016496337.1	10.01/26.45	240	PM, C
NtXIP1;1α	NP_001312796	7.70/34.61	325	PM
NtXIP1;1β	Nitab4.5_0000956g0150.1	7.71/34.75	325	PM
NtXIP1;2α	XP_016446694	7.71/34.68	326	PM
NtXIP1;2β	Nitab4.5_0007293g0050.1	7.71/34.54	326	PM
NtXIP2;1	XP_016489264.1	6.05/33.40	313	PM
NtXIP2;2	XP_016488683	8.70/33.07	308	PM

[1] IP = Isoelectric point. [2] PM: plasma membrane, C: chloroplast, V: vacuole.

This represents the greatest *AQP* gene number in a Solanaceae plant genome. We identified 49 *AQP* genes encoding 51 and 50 full-length proteins in two *N. tabacum* ancestors, namely *N. tomentosiformis* and *N. sylvestris*, respectively, as shown in Table S1. The phylogenetic protein analysis showed that NtAQPs cluster into five subfamilies (PIPs, TIPs, NIPs, SIPs, and XIPs) similar to NtoAQPs, NsAQPs, and SlAQP and StAQP, as shown in Figures 1–3. NtAQPs nomenclature was done from protein sequence comparison with the known SlAQP and StAQP, as shown in Figure 1. Sequences belonging to hybrid intrinsic proteins (HIPs) and GlpF-like intrinsic proteins (GIPs) reported in the non-vascular moss *Physcomitrella patens* [14] were not found. In *N. tabacum*, we identified 34 PIPs, 27 TIPs, 20 NIPs, 3 SIPs, and 6 XIPs, including two splice variants. Figure 1 shows that the PIPs cluster either into the PIP1 or PIP2 groups, and the NtTIPs into five groups (TIP1 to TIP5), similar to the potato and tomato TIPs [3,7]. Eight NIP groups were found in *N. tabacum*, contrary to the seven groups in *Arabidopsis* and soybean [5,11], and three to four NIP groups in poplar, rice, and maize [6,10,12]. Similar to *Arabidopsis*, rice, maize, poplar, and soybean, *N. tabacum* had two SIP groups, namely SIP1 and SIP2s, with two and one isoforms, respectively. Two XIP subgroups were observed in *N. tabacum*, and four XIP subgroups in potato [3].

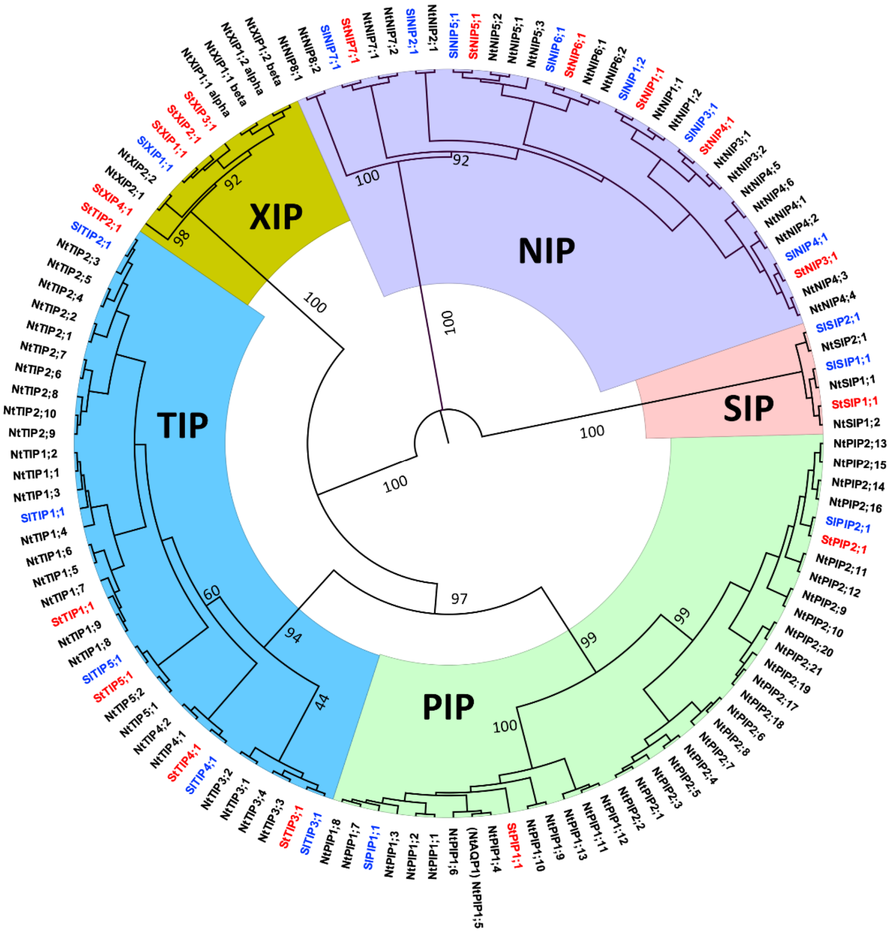


Figure 1. Phylogenetic relationships among *Nicotiana tabacum*, *Solanum tuberosum*, and *Solanum lycopersium* AQPs. For this analysis, 35 selected subgroup representative StAQPs and SlAQPs were aligned with all NtAQPs using the Clustal Omega server (http://www.ebi.ac.uk/Tools/msa/ClustalOmega/) and a phylogenetic tree was constructed using Maximum Likelihood method based on the JTT matrix-based model with 1000 bootstraps. AQPs clustered into five different subfamilies (PIPs, TIPs, NIPs, SIPs, and XIPs). Each AQP subfamily is shown with a specific background color. NtAQPs are indicated in black; StAQPs and SlAQPs are in red and blue, respectively.

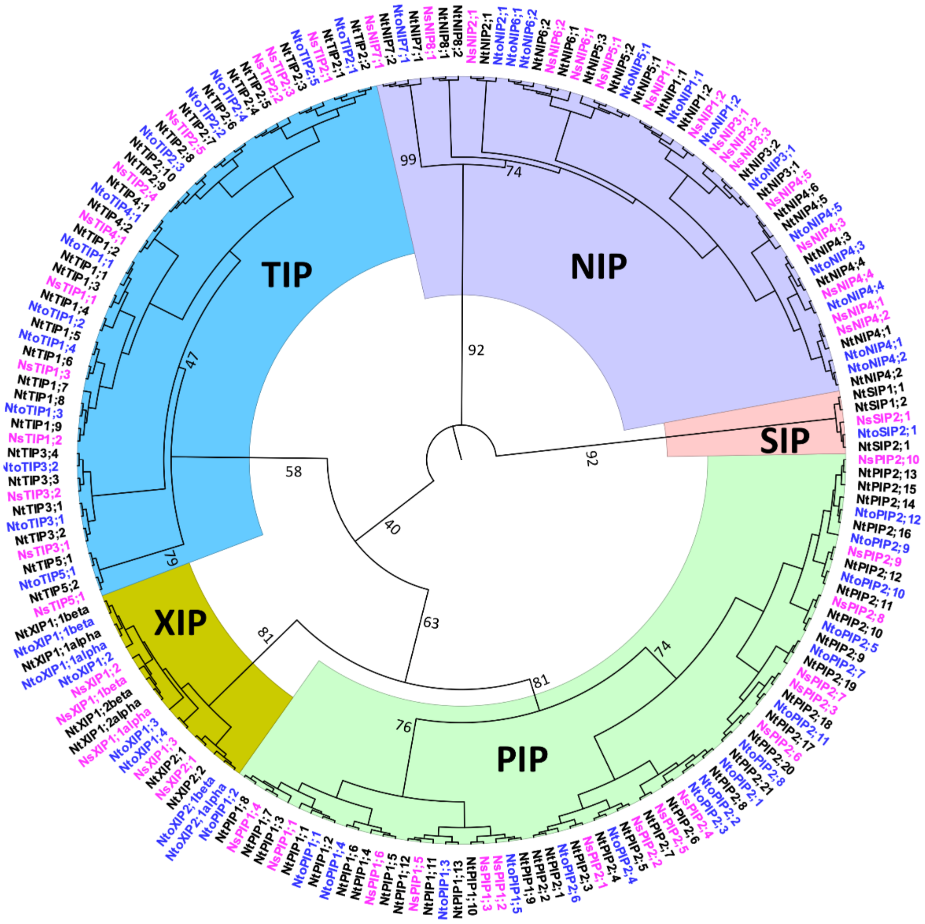


Figure 2. Phylogenetic relationships among *N. tabacum (Nt)* AQPs and its two ancestors, *N. sylvestris (Ns)* and *N. tomentosiformis (Nto)* AQPs. The deduced amino acid sequences of NtAQPs, NtoAQPs, and NsAQPs were aligned using the Clustal Omega server (http://www.ebi.ac.uk/Tools/msa/ClustalOmega/) and a phylogenetic tree was constructed using Maximum Likelihood method based on the JTT matrix-based model with 1000 bootstraps. The NtAQPs clustered into five different subfamilies (PIPs, TIPs, NIPs, SIPs and XIPs), with the corresponding NtoAQP and NsAQP subfamilies. Each AQP subfamily is shown with a specific background color. NtAQPs are indicated in black, NtoAQPs are in blue, and NsAQPs are in magenta.

Subcellular localization prediction was conducted using WoLF PSORT software, and the results were as follows: NtPIPs–plasma membrane (PM) and chloroplast, as shown in Table 1; TIPs–vacuole and PM; NIPs–PM and vacuole; SIPs–PM (SIP2;1 in both the PM and chloroplast); XIPs–PM. These localizations are just predictions and need to be experimentally demonstrated. Part of the predictions are in agreement with the data reported in the literature, but many differences are also observed. For instance, plant PIP2s are not found in the chloroplasts, TIPs are mostly located in the vacuole (and not in the PM, as predicted for many NtTIPs), and NIPs were not identified in the vacuole. SIPs were localized in the PM and/or the ER in Arabidopsis and maize [40] (Lebrun and Chaumont, unpublished data), but never in the chloroplast. The amino acid number, calculated molecular weight (MW), and isoelectric point (pI) of NtAQP homologs are shown in Table 1.

Like their counterparts in other plant species, all PIPs, TIPs, NIP1s, NIP2s, NIP3s, NIP4s, NIP7s, and NIP8s from *N. tabacum,* have two conserved NPA motifs in loops B and E, as shown in Figure 3 and Figures S1–S5. NIP5s and NIP6s have unusual NPA motifs, in which the alanine in loop E is substituted by a valine, and have a characteristic arginine-rich C-terminus, as shown in Figure 3 and Figure S3. In *N. tabacum* SIPs, the alanine in the first NPA motif is substitued by either a threonine (SIP1;1) or a leucine (SIP2s) residue, as shown in Figure 3 and Figure S4. All the SIPs have the conserved NPA motif in loop E with a unique characteristic lysine-rich C-terminus, as shown in Figure S4, which contains an ER retention signal [1,41] (Lebrun and Chaumont, unpublished). In the *N. tabacum* genome, there are four *XIP* genes, including *NtXIP1;1* and *NtXIP1;2*, which encode two splice variants (α and β) [15]. In *N. tabacum* XIPs in the first NPA motif (loop B), alanine is substituted by a valine residue, as shown in Figure 3 and Figure S5.

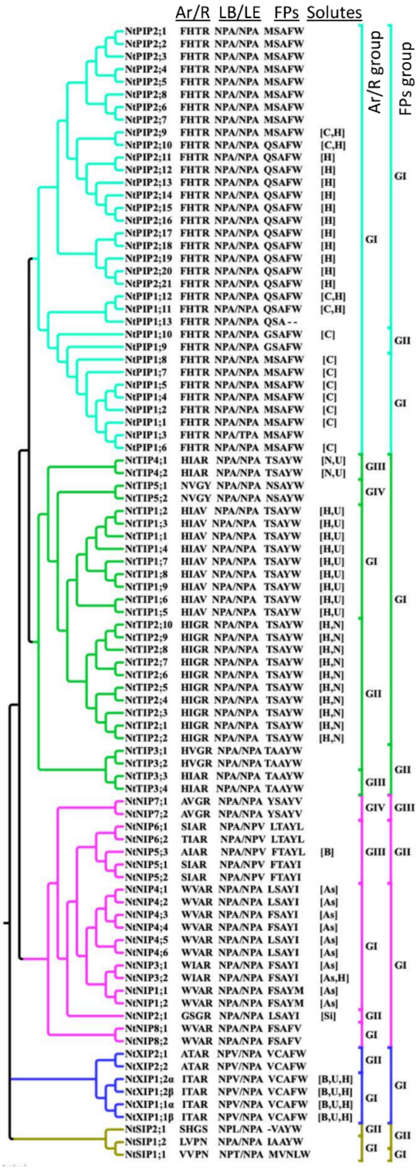

Ar/R LB/LE FPs Solutes
Ar/R group
FPs group
NtPIP2;1 FHTR NPA/NPA MSAFW
NtPIP2;2 FHTR NPA/NPA MSAFW
NtPIP2;3 FHTR NPA/NPA MSAFW
NtPIP2;4 FHTR NPA/NPA MSAFW
NtPIP2;5 FHTR NPA/NPA MSAFW
NtPIP2;8 FHTR NPA/NPA MSAFW
NtPIP2;6 FHTR NPA/NPA MSAFW
NtPIP2;7 FHTR NPA/NPA MSAFW
NtPIP2;9 FHTR NPA/NPA MSAFW [C,H]
NtPIP2;10 FHTR NPA/NPA QSAFW [C,H]
NtPIP2;11 FHTR NPA/NPA QSAFW [H]
NtPIP2;12 FHTR NPA/NPA QSAFW [H]
NtPIP2;13 FHTR NPA/NPA QSAFW [H]
NtPIP2;14 FHTR NPA/NPA QSAFW [H]
NtPIP2;15 FHTR NPA/NPA QSAFW [H]
NtPIP2;16 FHTR NPA/NPA QSAFW [H]
NtPIP2;17 FHTR NPA/NPA QSAFW [H]
NtPIP2;18 FHTR NPA/NPA QSAFW [H]
NtPIP2;19 FHTR NPA/NPA QSAFW [H]
NtPIP2;20 FHTR NPA/NPA QSAFW [H]
NtPIP2;21 FHTR NPA/NPA QSAFW [H]
NtPIP1;12 FHTR NPA/NPA QSAFW [C,H]
NtPIP1;11 FHTR NPA/NPA QSAFW [C,H]
NtPIP1;13 FHTR NPA/NPA QSA - -
NtPIP1;10 FHTR NPA/NPA GSAFW [C]
NtPIP1;9 FHTR NPA/NPA GSAFW
NtPIP1;8 FHTR NPA/NPA MSAFW [C]
NtPIP1;7 FHTR NPA/NPA MSAFW [C]
NtPIP1;5 FHTR NPA/NPA MSAFW [C]
NtPIP1;4 FHTR NPA/NPA MSAFW [C]
NtPIP1;2 FHTR NPA/NPA MSAFW [C]
NtPIP1;1 FHTR NPA/NPA MSAFW [C]
NtPIP1;3 FHTR NPA/TPA MSAFW
NtPIP1;6 FHTR NPA/NPA MSAFW [C]
NtTIP4;1 HIAR NPA/NPA TSAYW [N,U]
NtTIP4;2 HIAR NPA/NPA TSAYW [N,U]
NtTIP5;1 NVGY NPA/NPA NSAYW
NtTIP5;2 NVGY NPA/NPA NSAYW
NtTIP1;2 HIAV NPA/NPA TSAYW [H,U]
NtTIP1;3 HIAV NPA/NPA TSAYW [H,U]
NtTIP1;1 HIAV NPA/NPA TSAYW [H,U]
NtTIP1;4 HIAV NPA/NPA TSAYW [H,U]
NtTIP1;7 HIAV NPA/NPA TSAYW [H,U]
NtTIP1;8 HIAV NPA/NPA TSAYW [H,U]
NtTIP1;9 HIAV NPA/NPA TSAYW [H,U]
NtTIP1;6 HIAV NPA/NPA TSAYW [H,U]
NtTIP1;5 HIAV NPA/NPA TSAYW [H,U]
NtTIP2;10 HIGR NPA/NPA TSAYW [H,N]
NtTIP2;9 HIGR NPA/NPA TSAYW [H,N]
NtTIP2;8 HIGR NPA/NPA TSAYW [H,N]
NtTIP2;7 HIGR NPA/NPA TSAYW [H,N]
NtTIP2;6 HIGR NPA/NPA TSAYW [H,N]
NtTIP2;5 HIGR NPA/NPA TSAYW [H,N]
NtTIP2;4 HIGR NPA/NPA TSAYW [H,N]
NtTIP2;3 HIGR NPA/NPA TSAYW [H,N]
NtTIP2;1 HIGR NPA/NPA TSAYW [H,N]
NtTIP2;2 HIGR NPA/NPA TSAYW [H,N]
NtTIP3;1 HVGR NPA/NPA TAAYW
NtTIP3;2 HVGR NPA/NPA TAAYW
NtTIP3;3 HIAR NPA/NPA TAAYW
NtTIP3;4 HIAR NPA/NPA TAAYW
NtNIP7;1 AVGR NPA/NPA YSAYV
NtNIP7;2 AVGR NPA/NPA YSAYV
NtNIP6;1 SIAR NPA/NPV LTAYL
NtNIP6;2 TIAR NPA/NPV LTAYL
NtNIP5;3 AIAR NPA/NPV FTAYL [B]
NtNIP5;1 SIAR NPA/NPV FTAYI
NtNIP5;2 SIAR NPA/NPV FTAYI
NtNIP4;1 WVAR NPA/NPA LSAYI [As]
NtNIP4;2 WVAR NPA/NPA LSAYI [As]
NtNIP4;3 WVAR NPA/NPA FSAYI [As]
NtNIP4;4 WVAR NPA/NPA FSAYI [As]
NtNIP4;5 WVAR NPA/NPA LSAYI [As]
NtNIP4;6 WVAR NPA/NPA LSAYI [As]
NtNIP3;1 WIAR NPA/NPA FSAYI [As]
NtNIP3;2 WIAR NPA/NPA FSAYI [As,H]
NtNIP1;1 WVAR NPA/NPA FSAYM [As]
NtNIP1;2 WVAR NPA/NPA FSAYM [As]
NtNIP2;1 GSGR NPA/NPA LSAYI [Si]
NtNIP8;1 WVAR NPA/NPA FSAFV
NtNIP8;2 WVAR NPA/NPA FSAFV
NtXIP2;1 ATAR NPV/NPA VCAFW
NtXIP2;2 ATAR NPV/NPA VCAFW
NtXIP1;2α ITAR NPV/NPA VCAFW [B,U,H]
NtXIP1;2β ITAR NPV/NPA VCAFW [B,U,H]
NtXIP1;1α ITAR NPV/NPA VCAFW [B,U,H]
NtXIP1;1β ITAR NPV/NPA VCAFW [B,U,H]
NtSIP2;1 SHGS NPL/NPA -VAYW
NtSIP1;2 LVPN NPA/NPA IAAYW
NtSIP1;1 VVPN NPT/NPA MVNLW
GI
GII
GI
GIII
GIV
GI
GII
GIII
GIV
GIII
GI
GII
GI
GII
GI
GII
GI
GI
GII
GI
GII
GIII
GII
GI
GI
GII
GI

Figure 3. Grouping of *N. tabacum* PIPs, TIPs, NIPs, SIPs, and XIPs based on the ar/R and FPs. The phylogenetic tree was generated as described in Figure 1. The residues in the ar/R selectivity filter and the FPs were identified from the multiple sequence alignment, shown in Figures S1–S5. The ar/R and FP groupings within each subfamily were done based on the corresponding amino acid compositions, which are indicated on the right side of the phylogenetic tree. The solutes predicted, based on substrate specific signature sequences to be transported, are mentioned in square brackets. As, B, C, H, N, Si, Sb, and U indicate arsenic, boron, CO_2, H_2O_2, ammonia, silicon, antimony, and urea, respectively.

2.2. NtAQP Gene Structures

The *N. tabacum AQP* genomic sequences were analyzed for introns and exons, as shown in Figure 4 and Figure S6. Apart from a few inconsistencies, the number and position of introns are conserved within each *AQP* subfamily. *NtPIP* genes have two or three introns, except for *NtPIP2;3*, which has a single intron, and *NtPIP1;5*, *NtPIP1;7*, *NtPIP1;11*, and *NtPIP2;8*, which have no introns, as shown in Figure 4. Among them, *NtPIP2;2* has a very long intron (~15 kb), as shown in Figure S6. The *NtTIP* subfamily exhibits relatively stable gene structure in comparison with other subfamilies. The majority of them have two introns except for *TIP1;2–4* and *TIP1;8–9*, which have a single intron and *NtTIP1;1* with no intron, as shown in Figure 4. The majority of *NtNIPs* have four introns with variable intron-exon organization, as shown in Figure 4 and Figure S6. *NtNIP5;1* has three introns, and *NtNIP3s* and *NtNIP6;1* have five introns, while *NtNIP8;2* possesses a unique gene structure with six introns (the greatest number of introns in an *AQP* gene), one of which is 10 kb long, as shown in Figure S6. The *NtSIP* genes have two introns, except for *NtSIP2;1,* which has no intron. The *NtXIPs* gene structure was very conserved with two introns, except for *NtXIP2;1,* which has a single intron, as shown in Figure 4.

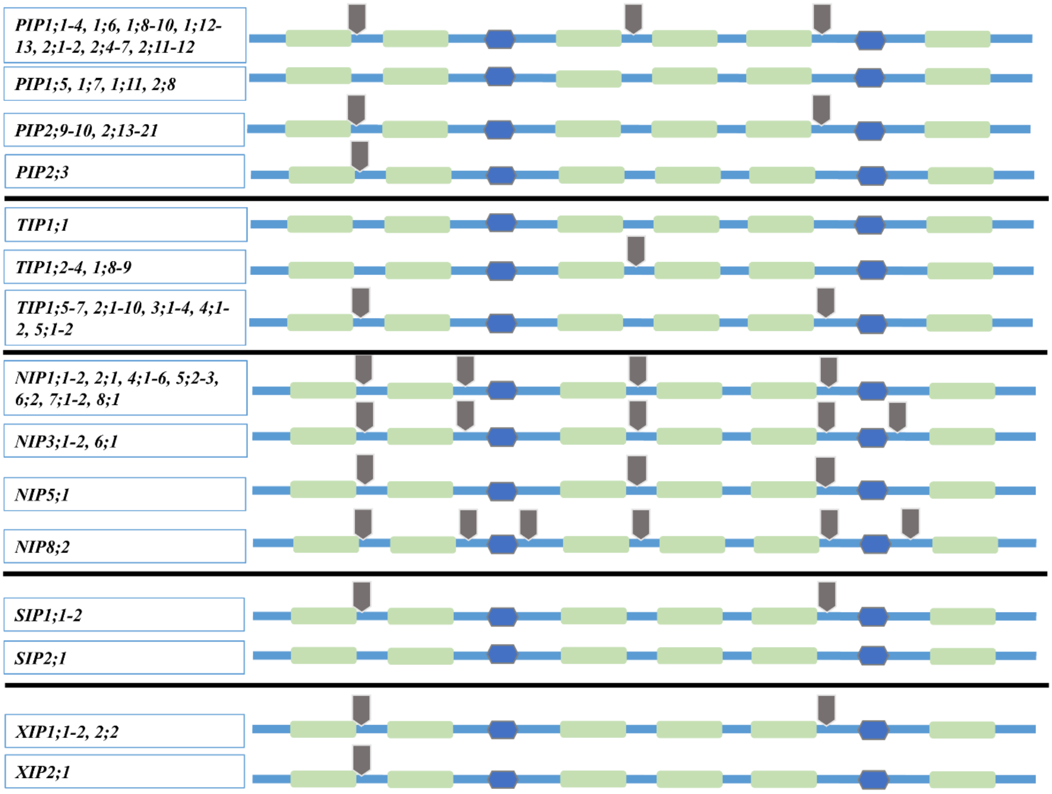

PIP1;1-4, 1;6, 1;8-10, 1;12-13, 2;1-2, 2;4-7, 2;11-12
PIP1;5, 1;7, 1;11, 2;8
PIP2;9-10, 2;13-21
PIP2;3
TIP1;1
TIP1;2-4, 1;8-9
TIP1;5-7, 2;1-10, 3;1-4, 4;1-2, 5;1-2
NIP1;1-2, 2;1, 4;1-6, 5;2-3, 6;2, 7;1-2, 8;1
NIP3;1-2, 6;1
NIP5;1
NIP8;2
SIP1;1-2
SIP2;1
XIP1;1-2, 2;2
XIP2;1

Figure 4. *NtAQP* genes and 2-D protein structure. Introns in the *NtAQP* genes are indicated by gray arrows. The six TM regions are shown in light green bars, and loops B and E are shown in blue hexagons.

2.3. Analysis of NtAQPs Ar/R Selectivity Filter and Froger's Position

We identified the four amino acid residues at the ar/R selectivity filter and the five residues in the FPs using sequence alignments, and used them to group the NtAQPs based on the amino acid residue properties and to compare these groups with those of other species, such as tomato and potato, as shown in Figure 3 [3,7,9]. In addition, all NtAQPs were subjected to the ScanProsite tool (http://prosite.expasy.org/scanprosite/), to identify the substrate specificity-determining positions (SDPs) based on the ar/R, FP, and NPA motifs, and thereby the predicted substrate(s) of each isoform, as shown in Table 2, Figure 3, and Table S2. Water is considered as the universal substrate for AQPs, even though some isoforms were shown not to facilitate its diffusion through the membrane [15].

Table 2. Substrate specificity determining positions (SDPs) in NtAQPs.

Substrates	Ar/R (H2-H5-LE1-LE2)	LB (NPA Region)	LE (NPA Region)	FPs (P1-P5)	Transporters Based on Those SDP Positions
Bo	[AGI][ISV][GA]R	SG[AG]H[ILM]NP[ASV][VLI][TS]	[GS][GA][SG]MNP[AV]R[STC][LF]G	[FIV][TC]A[YF][LFW]	NtNIP5;1, NtXIP1;1–2
CO_2	FHTR	SGGHINPAVT	GTGINPARSLG	[MQ]SAFW	NtPIP1;1–2, NtPIP1;4–8, NtPIP1;10, NtPIP1;12–13, NtPIP2;9–10
H_2O_2	[HFWI][IHV][ATG][VR]	SG[GA]H[VLIF]NP[AV][VI][TS]	G[AGT][SG][MI]NP[AG][VR][ASC][FL]	[TQFV][ASC]A[YF][WI]	NtPIP1;12–13, NtPIP2;9–21, NtTIP1;1–9, NtTIP2;1–10, NtNIP3;2, NtXIP1;1–2
NH_3	[HW][IV][AG]R	SGGH[VLF]NPAVT	G[GA]SMNPARS[FL]G	[FT]SAY[LW]	NtTIP2;1–10, NtTIP4;1–2
Si	GSGR	SGAHMNPAVT	GGSMNPARTL[GA]	[IL]TAYF	NtNIP2;1
U	[HGANI][ISV][AG][RVC]	SG[GA]H[ILVM]NP[AV][VI][TS]	[GS][AG][SG]MNP[AV][RVC][TSC][LF]G	[MTLFVI][SATC]A[YF][WFL]	NtTIP1;1–9, NtTIP4;1–2, NtXIP1;1–2
As	[GAW][VSAI][GA][RV]	SG[AC]H[LIVMF]NP[AS][VI]T	[GS][GA]SMNP[AV]R[ST][LI][AG]	[LIFY][TS]AY[FILM]	NtNIP1;1–2, NtNIP3;1–2, NtNIP4;1–6
Sb	[AGT][IVSA][GA]R	SG[AC]H[LM]NP[SA][VIT][TS]	[GS][GA]SMNP[VA]R[TS]L[GA]	[FYIL][TS]AY[LMF]	-

Bo, Boron; H_2O_2, Hydrogen peroxide; CO_2, Carbon dioxide; U, Urea; NH_3, ammonia; As, arsenic; Sb, antimony; Si, silicon.

The ar/R selectivity filter in all the NtPIPs is composed of F, H, T, and R residues in TM2, TM5, LE1, and LE2, respectively, and is identical to the ar/R filter found in all the plant PIPs, as shown in Figure 3. According to the residues located at the P1 of FPs, M or Q (G), NtPIPs cluster into two groups, I and II, as shown in Figure 3. Twelve PIPs (mainly PIP1s) are predicted CO_2 channels and thirteen PIPs (mainly PIP2s) are predicted H_2O_2 channels, as shown in Figure 3. Based on the ar/R filter, the NtTIPs cluster into four groups (I, II, III, and IV), as shown in Figure 3. The P3–P5 positions in FPs of all NtTIPs are conserved and consist of A, Y, and W residues, respectively, as shown in Figure 3. Based on the disparities in P1 and P2 positions, all TIPs could be divided into two groups. TIP1s and TIP2s are predicted H_2O_2 channels, and TIP1s and TIP4s are predicted urea channels, as shown in Figure 3. TIP2s and TIP4s are also predicted as NH_3 channels, which is in agreement with experimental evidence in other species [18,42]. Based on the ar/R selectivity filters, all NtNIPs are divided into four different groups, as shown in Figure 3. On the other hand, based on the FPs, NtNIPs cluster into three groups, as shown in Figure 3, such as potato and tomato, but unlike other plants (Arabidopsis, maize, etc.) [3,6,7,11]. Our analysis predicted that the As transporters are only distributed among the NtNIPs (10 NIPs belonging to Group I, based on the ar/R filter and FPs), as shown in Figure 3. NIP2;1, NIP5;1, and NIP3;2 are predicted as Si, B, and H_2O_2 channels, respectively. The NtSIPs are grouped into two groups based on both the ar/R selectivity filter and FPs, as shown in Figure 3. Very few studies have examined the channel specificity of SIPs. Two SIPs from Arabidopsis showed some water channel activity when expressed in yeast [40]. The NtXIPs are clustered into two groups based on the ar/R selectivity filter. However, based on FPs, all NtXIPs were grouped in a single group, as shown in Figure 3. XIP1;1 and XIP1;2 are predicted as B, urea, and H_2O_2 channels, as shown in Figure 3. The specificity and function of NtXIP1;1, including its splice variant, were studied in detail and were shown to facilitate the diffusion of B, H_2O_2, NH_3, and urea, but not water [15,43].

2.4. Expression of NtAQP Genes in Roots, Leaves, and Flowers as well as BY-2 Suspension Cells

The heatmap based on FPKM values shows the *NtAQPs* transcript levels in roots, leaves, and flowers, as shown in Figure 5. Among the 88 *NtAQPs* genes, 73, 75, and 71 are expressed in mature flowers, leaves, and roots, respectively, and 68 genes are ubiquitously expressed in all analyzed organs. *PIPs* are expressed in flowers, leaves, and roots but differently according to the isoforms. A greater number of *NtPIP1* genes are expressed in flowers and leaves than in roots—*NtPIP1;1* and *NtPIP1;10* being the most expressed isoforms in flowers and leaves, respectively, and *NtPIP1;3–8* and *NtPIP1;11* not being expressed in roots. A decreased amount of *NtPIP2* transcripts is generally observed, but all *NtPIP2s* are expressed in the three organs with the exception of *NtPIP2;9* and *NtPIP2;18*, which are not expressed or are expressed very little, as shown in Figure 5. *NtTIP* gene expression levels are often greater in the leaves compared with the other organs, even if a greater number of *NtTIP* genes are expressed in roots, as shown in Figure 5. Among the 20 *NtNIP* genes, seven (*NtNIP3;2* and all the *NtNIP4s*) are not or very lowly expressed in the three organs in the tested conditions. The other *NtNIP* genes are relatively less expressed compared to the other *AQP* subfamily members, as shown in Figure 5. All *NtSIP* genes were ubiquitously expressed in flowers, leaves, and roots, *NtSIP1;2* being the most expressed *NtSIP* in the leaf, as shown in Figure 5. Finally, *NtXIP1;1* was the most expressed *NtXIP* in the three organs with the expression of the others being very decreased.

N. tabacum BY-2 suspension cells are widely used to study different physiological processes, the role of specific proteins, or as a heterologous expression system to produce high value pharmaceutical antigens or antibodies [44–48]. We determined which *AQP* genes are expressed in those cells that grow in suspension in an aqueous environment. RNA from wild-type BY-2 cells was extracted and RNA-seq data analyzed for the expression of the 88 *NtAQP* genes. The heatmap based on FPKM values is shown in Figure 6. mRNA of 53 *NtAQP* genes were detected in BY-2 cells growing in a standard MS medium. The most expressed *NtAQP* genes were 11 *PIP1s*, *TIP1;1*, the three *SIPs*, and *XIP1;1*.

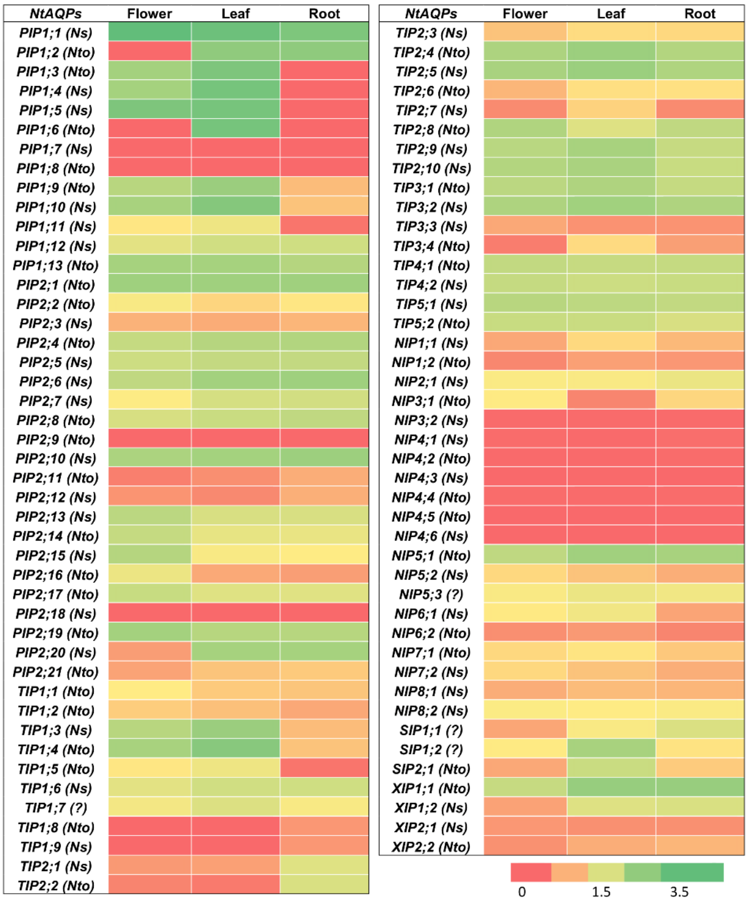


Figure 5. Expression analyses of 88 *NtAQP* genes in root, leaf, and flower. Color scale represents logarithmic FPKM values, where green indicates high expression and red indicates no or very low expression. *Ns* and *Nto* in parentheses indicate that corresponding *NtAQP* gene evolved from *N. sylvestris (Ns)* or *N. tomentosiformis (Nto).* Question mark (?) indicates that *NtAQP* gene origin (*N. sylvestris* or *N. tomentosiformis*) was not identified.

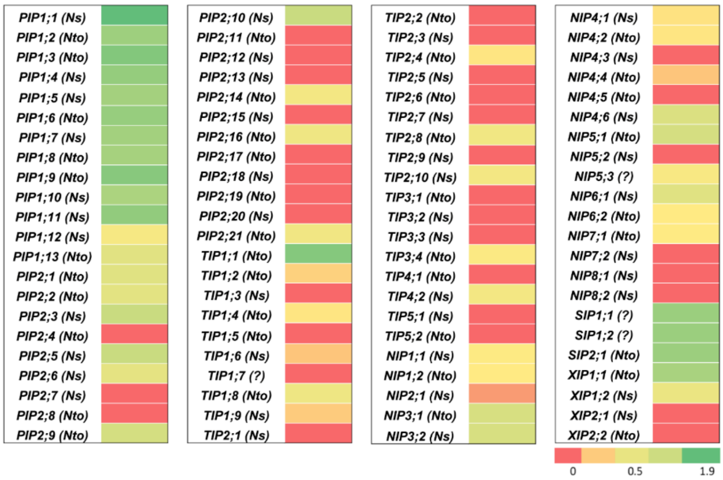


Figure 6. Expression analyses of 88 *NtAQP* genes in *N. tabacum* BY-2 cells. Color scale represents logarithmic FPKM values, where green indicates high expression and red indicates no expression or very low expression. *Ns* and *Nto* in parentheses indicate that corresponding *NtAQP* gene evolved from *N. sylvestris (Ns)* or *N. tomentosiformis (Nto).* Question mark (?) indicates that *NtAQP* gene origin (*N. sylvestris* or *N. tomentosiformis*) was not identified.

3. Discussion

By screening the *N. tabacum* genome databases, we identified 88 complete *AQP* genes, almost twice the number of *AQP* genes identified in tomato and potato [3,7]. The number of *AQP* homologs always varies between plant species, the dicot plant genomes usually encoding more homologs than the monocot plants, except for the 68 full-length *AQP* genes found in *P. virgatum*, a polyploid monocot species [9]. The great number of *AQP* genes in the *N. tabacum* genome arose from an allotetraploidization event that occurred about 200,000 years ago [36,38] between *N. tomentosiformis* and *N. sylvestris*, which each have 49 *AQP* genes. The difference between the identified gene number in *N. tabacum* (88) and the sum of the *N. tomentosiformis* and *N. sylvestris AQP* genes (98) suggests that some were lost after the polyploidization event. In addition, we also could not exclude the recent local duplication events in each species, as deduced by the protein phylogenetic tree, shown in Figure 2, in which two very close isoforms from the same species are found on the same branch (i.e., NtPIP2;1 and 2;2, NtoPIP2;2 and 2;3, NsNIP3;1 and 3;2, NtoNIP6;1 and 6;2, etc.). Models have been proposed to explain duplicated gene fate: pseudogenization, sub-functionalization, and neo-functionalization [49]. Redundancy also allows one of the copies to accumulate mutations without affecting plant fitness, and new allelic variants or changes in the gene expression pattern can be observed [50]. While activity determination of the duplicated isoforms would be required to determine a sub- or neo-functionalization, changes in expression patterns can be deduced from the rough *NtAQP* expression data analysis. For instance, the duplicated *NtPIP2;1* and *NtPIP2;2* showed different expression levels, which can be organ dependent.

We identified five subfamilies (PIP, TIP, NIP, SIP, and XIP) among the three *Nicotiana* species, similar to most other dicots, except for Brassicaceae and monocots, which have no XIP subfamily [15]. Several *N. tabacum* AQPs have been characterized [51–55], and some became paradigms in the plant AQP community [21,22]. NtAQP1, corresponding to NtPIP1;5 in our study, is a PIP1 protein located both in the plasma membrane and the chloroplast envelope, which exhibits water and CO_2 channel permeability [21]. This discovery highlighted the important diverse roles of AQPs in plant physiology and, more particularly, in photosynthesis, through their contribution in facilitating CO_2 membrane diffusion [28]. More recently, the membrane diffusion of another gas, O_2, was reported to be facilitated by NtPIP1;3 when expressed in yeast, and an increased *NtPIP1;3* transcript level was measured in *N. tabacum* roots after a seven day hypoxia treatment [22], suggesting a potential new physiological role of plant AQPs in O_2 membrane permeability. NtXIP1s are the first plant XIP isoforms that have been functionally characterized [15]. NtXIP1;1 is located in the plasma membrane and is shown in a functional assay in heterologous systems to facilitate the membrane diffusion of H_2O_2, glycerol, boron, and urea, but not water [15]. NtXIP1;1 overexpression in *N. tabacum* results in disturbed boron tissue distribution, leading to boron deficient phenotypes in meristems and young leaves [43]. Interestingly, the *NtXIP1;1* gene contains a sequence motif in the first intron that initiates an RNA-processing mechanism that results in two splice variants (α and β), resulting in two amino acid residue differences [15]. We also identified XIP spliced variants for *NtXIP1;2*, *NtoXIP1;1*, *NsXIP1;1*, and *NtoXIP2;1* isoforms, and also XIPs from *S. tuberosum* and *S. lycopersicum* [15], indicating a conservation of this genomic feature in the Solanaceae family.

To elucidate the substrate specificity of NtAQPs, different signature sequences, including SDPs, NPA motifs, ar/R filter, and FPs were identified, as shown in Figure 3 and Table 2. From this multiple analysis, a majority of PIP1s and PIP2s were predicted to facilitate CO_2 and H_2O_2 diffusion, respectively, in addition to water, as shown in Figure 3. This was confirmed in functional assays performed for NtPIP1;5 (NtAQP1) and NtPIP2;1 [21,53,54]. Most TIPs have similar NPA and FPs, suggesting that differences in their substrate transport selectivity might be regulated by the ar/R filter residues. Based on this ar/R filter, Group I and Group II TIPs have a wider pore aperture, which might facilitate the diffusion of relatively larger substrates than water, such as urea, ammonia, and H_2O_2 [56–58]. NtTIP4;1 (NtTIPa) was indeed shown to be permeable to water and urea, but also glycerol [51]. NIPs are most diverse in their NPA motifs, ar/R filter, and FPs, suggesting various substrate transport selectivities for these subfamily members and putatively important physiological roles. NIPs are also known to facilitate the transport of metalloids, such as arsenic and boron, as shown in Figure 3. NtXIPs are predicted to transport H_2O_2, boric acid, and urea, and were confirmed in transport assays performed with NtXIP1;1 [15,43]. In addition, NtXIP1;1 is not a water channel but is able to facilitate glycerol diffusion [15]. Finally, limited information is available for plant SIP specificity. Water channel activity was determined for AtSIP1s, unlike for AtSIP2;1 [40]. This global substrate specificity study, based on prediction is, however, to be taken with caution, as a single amino acid change, even in the transmembrane domains, could affect the channel characteristic or conformation [59]. Therefore functional assays in heterologous or homologous systems will have to be carried out when analyzing the functional role of specific NtAQP.

As expected, *NtAQP* transcript levels are dependent on the plant organs, but it is quite surprising to observe that 68 of 88 *AQP* genes are ubiquitously expressed in roots, young leaves, and flowers. *PIP* and *TIP* transcripts are relatively more abundant than other subfamily mRNAs, as shown in Figure 5. Considering that the main role of these isoforms is the water facilitated permeation through plasma and vacuolar membranes, this observation confirms their primordial role in water movement through plant tissues, in cell expansion, and cell water homeostasis [24,60]. The *NIP* expression level is low, except for *NtNIP5;1*, but due to their metalloid substrate specificity, a more restricted tissue/cell expression pattern in specific physiological conditions might be expected [43,61,62]. mRNA of 53 *NtAQP* genes were also detected in BY-2 suspension cells growing in a standard MS medium, even if the relative expression level between them was different to what was observed in plant organs. This could be due to the dedifferentiated nature of those cells and/or the specific cell environment of the culture medium. The most expressed *NtAQP* genes in BY-2 cells are 11 *NtPIP1s*, *NtTIP1;1*, the three *NtSIPs*, and *NtXIP1;1*. High *NtPIP* gene expression was also reported in maize Black Mexican Sweet (BMS) suspension cells [63], but in this case, the two most expressed genes belonged to the PIP2 group. Plant PIP1s physically interact with PIP2s within heterotetramers, leading to PIP1 relocalization from the endoplasmic reticulum to the plasma membrane [59,64]. We might wonder whether PIP2 abundance in BY-2 cells is sufficient to bring all PIP1s to the plasma membrane. The increased PIP expression in suspension cells suggests that they are important in controlling membrane water permeability during suspension cell growth. In fact, *PIP* expression varies according to BMS cell growth stages, and this is correlated with greater cell water permeability, measured at the end of the log phase and stationary phase [63]. This might be dependent on variations in the medium composition and/or internal osmotic pressure. *PIP* and *TIP* gene expression in BY-2 suspension cells might also be involved in the control of cell expansion. Cauliflower BobTIP26–1 overexpression in suspension cells (*N. tabacum* cv. Wisconsin 38) increases the cell volume [65], cell enlargement being mostly accounted by vacuole swelling. The quite high expression of *NtSIPs* is also intriguing, knowing that SIPs are mostly expressed in the endoplasmic reticulum and their function is still unknown. *ZmSIP1;2* is also expressed in BMS suspension cells, and its expression is not dependent on the growth stage [63]. Suspension cells might be a promising model to investigate the physiological role at the cell level as well as the biochemical properties of this AQP subfamily. Actually, BY-2 suspension cells represent very useful tools to study AQP function, localization regulation, substrate specificity, and structure, as the cells are easily transformed by *Agrobacterium tumefaciens* or biolistics, and great cell amounts could be obtained for protein purification and reconstitution [66].

In this comprehensive analysis, we identified a highly diverse AQP gene family in *N. tabacum* as well as in its two ancestors, *N. tomentosiformis* and *N. sylvestris*. The signature sequence for substrate selectivity and the possible biological function of NtAQPs were predicted. The transcriptomic data of *N. tabacum* and BY-2 suspension cells represent an excellent resource to guide further analysis of the function of any selected AQP isoform.

4. Materials and Methods

4.1. Identification and Sequence Analysis of NtAQPs

The genomes of *N. tabacum*, *N. tomentosiformis*, and *N. sylvestris* available at the Sol Genomics Network (https://solgenomics.net/organism/Nicotiana_tabacum/genome), were searched for AQPs using BLASTp (http//http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Proteins) tools with the protein sequences of 47 AQPs from *S. lycopersium* (tomato) and 41 AQPs from *S. tuberosum* (potato) as queries. Every sequence from each species was individually compared with functional annotations by browsing the *N. tabacum* databases.

4.2. Phylogenetic Analysis of N. Tabacum AQPs (NtAQPs)

NtAQPs amino acid sequences were separately aligned with *S. lycopersium* AQPs (SlAQPs) and *S. tuberosum* AQPs (StAQPs) using the Clustal Omega program (https://www.ebi.ac.uk/Tools/msa/clustalo/) and a phylogenetic tree was built using Molecular Evolution Genetic Analysis (MEGA), version 7.0 [67]. The phylogenetic analysis was conducted using the Maximum Likelihood method, based on the Jones–Taylor–Thornton (JTT) matrix-based model with 1000 bootstraps. The identified NtAQPs were classified into different subfamilies according to the phylogenetic relationships with SlAQPs and StAQPs.

4.3. Identification of NtAQP Gene Structure and Transmembrane Helices

Gene structures were determined by the GSDS 2.0 software (http://gsds.cbi.pku.edu.cn/) using the *NtAQP* gene and CDS sequences as input. The TM α-helices were predicted by TMpred (http://www.ch.embnet.org/software/TMPRED_form.html) and SOSUI (http://bp.nuap.nagoya-u.ac.jp/sosui/).

4.4. Prediction of Subcellular Localization

The subcellular localization of NtAQPs was predicted by using the WoLF PSORT (http://wolfpsort.org/), TargetP (www.cbs.dtu.dk/Services/TargetP), Cello prediction system (http://cello.life.netu.edu.tw/), and MultiLoc2 (www.abi.inf.uni-tuebingen.de/Services/MultiLoc2) tools.

4.5. Identification of Substrate Specificity Determining Positions (SDPs)

The aligned NtAQP sequences were searched manually for SDPs by following the prediction explained previously [9,17] and clustered into different functional groups. The functional group sequences were aligned using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/).

4.6. Expression Profile of NtAQP Genes

Transcript levels as FPKM (Fragments per Kilobase of Transcript per Million Mapped Reads) values of *NtAQP* genes in different organs (mature flowers, leaves and roots) were obtained from the Gene Expression Omnibus (GEO) repository and GenBank Sequence Read Archive (SRA) under the accession code SRP029183 (SRX338104: *N. tabacum* TN90 root; SRX338101: *N. tabacum* TN90 leaf; SRX495520: *N. tabacum* TN90 mature flower). Three biological replicates were obtained from each organ. The FPKM values of the respective *NtAQP* genes were extracted from the databases and transformed into logarithmic ($\log_{10}$) values to generate the heatmap. A heatmap showing the logarithmic *NtAQPs* transcript levels in root, leaf, and flower was generated using Microsoft Excel conditional formatting, based on the normalized FPKM values. In our analysis, a logarithmic FPKM value > 0 was used as a threshold to consider whether a gene is expressed.

4.7. RNA-Seq Experiment

N. tabacum cv. BY-2 suspension cells were grown in the dark at 25 °C with agitation on a rotary shaker (90 rpm) in liquid MS medium (4.4 g/L Murashige and Skoog salts (MP BIOMEDICALS, Solon, OH), 30 g/L sucrose, 0.2 g/L KH_2PO_4, 2.5 mg/L thiamine, 50 mg/mL myo-inositol, and 0.2 mg/L 2,4-D, pH 5.8 (KOH)). Cultures were grown in 50 mL of medium in a 250 mL Erlenmeyer flask and a 5% inoculum was transferred each week into fresh medium. BY-2 cells (100 mg) were collected three days after inoculation (exponential phase) and the total RNA was extracted from three biological replicates and sent to the Macrogen Company, which performed the library preparation, RNA sequencing, and data analysis. For the library preparation, the mRNA was purified from total RNA and transformed into a template molecule library, appropriate for subsequent cluster generation using the Illumina® TruSeq™ RNA Sample Preparation Kit. The first step in the workflow encompassed purifying the poly-A-containing mRNA molecules using poly-T oligo-attached magnetic beads. After purification, the mRNA was split into small pieces using divalent cations under high temperature. The cleaved RNA fragments were copied into first strand cDNA using reverse transcriptase and random primers. This was followed by the second strand cDNA synthesis using DNA polymerase I and RNase H. These cDNA fragments then went through an end repair process, the addition of a single "A" base, and then the ligation of adapters. Finally, the products were purified and enriched with PCR to generate the final cDNA library. The library was then submitted for paired-end 2 × 100 bp sequencing in Illumina HiSeq2000. Sequencing data were analyzed through the Trinity pipeline, which permitted de novo transcriptome reconstruction. The transcript abundances were calculated using RSEM (1.2.15) software [68]. Blast-X (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastx&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) was used to compare the six-frame translation products of a nucleotide query sequence against a protein sequence database (go_v20150407). Finally, the FPKM values for the respective *AQP* genes were identified from the annotated BY-2 cell transcriptomic data. A heatmap was generated based on the transformed logarithmic ($\log_{10}$) FPKM values. Similar to organ specific expression data, FPKM values > 0 were used as a threshold to consider whether a gene is expressed.

Supplementary Materials: Supplementary Materials can be found at http://www.mdpi.com/1422-0067/21/13/4743/s1.

Author Contributions: All authors designed the experiments. J.A. and S.M. performed the experiments. All the authors analyzed the data. J.A. and F.C. wrote the manuscript. J.A., M.B., and F.C. contributed to the discussion and revision of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Belgian National Fund for Scientific Research (FNRS, FRFC 2.4.501.06F), the "Communauté française de Belgique-Actions de Recherches Concertées" (grants ARC16/21-075). J.A. and S.M. were supported by a research fellow at the Fonds de Formation à la Recherche dans l'Industrie et l'Agriculture.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

AQPs	Aquaporins
ar/R	Aromatic/arginine
As	Arsenic
B	Boron
BY-2	Bright Yellow-2
FPKM	Fragments per Kilobase of Transcript per Million Mapped Reads
FPs(P1–P5)	Froger's positions
GEO	Gene Expression Omnibus
GIPs	GlpF-like intrinsic proteins
H_2O_2	Hydrogen peroxide
HIPs	Hybrid intrinsic proteins
JTT	Jones–Taylor–Thornton
LE	Loop E
MEGA	Molecular Evolution Genetic Analysis
MIPs	Major intrinsic proteins
MW	Molecular weight
NH_3	Ammonia

NIPs	Nodulin-26-like intrinsic proteins
NPA	Asn-Pro-Alanine
pI	Isoelectric point
PIPs	Plasma membrane intrinsic proteins
Sb	Antimony
SDPs	Substrate specificity-determining positions
Si	Silicon
SIPs	Small basic intrinsic proteins
SRA	Sequence Read Archive
TIPs	Tonoplast intrinsic proteins
TM	Transmembrane
U	Urea
XIPs	Uncharacterized X intrinsic proteins

References

1. Gomes, D.; Agasse, A.; Thiébaud, P.; Delrot, S.; Gerós, H.; Chaumont, F. Aquaporins are multifunctional water and solute transporters highly divergent in living organisms. *Biochim. Biophys. Acta -Biomembr.* **2009**, *1788*, 1213–1228. [CrossRef] [PubMed]
2. Maurel, C.; Verdoucq, L.; Luu, D.-T.; Santoni, V. Plant aquaporins: Membrane channels with multiple integrated functions. *Annu. Rev. Plant Biol.* **2008**, *59*, 595–624. [CrossRef] [PubMed]
3. Venkatesh, J.; Yu, J.-W.; Park, S.W. Genome-wide analysis and expression profiling of the *Solanum tuberosum* aquaporins. *Plant Physiol. Biochem.* **2013**, *73*, 392–404. [CrossRef] [PubMed]
4. Bansal, A.; Sankararamakrishnan, R. Homology modeling of major intrinsic proteins in rice, maize and Arabidopsis: Comparative analysis of transmembrane helix association and aromatic/arginine selectivity filters. *BMC Struct. Biol.* **2007**, *7*, 27. [CrossRef]
5. Zhang, D.Y.; Ali, Z.; Wang, C.B.; Xu, L.; Yi, J.X.; Xu, Z.L.; Liu, X.Q.; He, X.L.; Huang, Y.H.; Khan, I.A. Genome-wide sequence characterization and expression analysis of major intrinsic proteins in soybean (*Glycine max* L.). *PLoS ONE* **2013**, *8*, e56312. [CrossRef]
6. Chaumont, F.; Barrieu, F.; Wojcik, E.; Chrispeels, M.J.; Jung, R. Aquaporins constitute a large and highly divergent protein family in maize. *Plant Physiol.* **2001**, *125*, 1206–1215. [CrossRef]
7. Reuscher, S.; Akiyama, M.; Mori, C.; Aoki, K.; Shibata, D.; Shiratake, K. Genome-wide identification and expression analysis of aquaporins in tomato. *PLoS ONE* **2013**, *8*, e79052. [CrossRef]
8. Park, W.; Scheffler, B.E.; Bauer, P.J.; Campbell, B.T. Identification of the family of aquaporin genes and their expression in upland cotton (*Gossypium hirsutum* L.). *BMC Plant Biol.* **2010**, *10*, 142. [CrossRef]
9. Azad, A.K.; Ahmed, J.; Alum, M.A.; Hasan, M.M.; Ishikawa, T.; Sawa, Y.; Katsuhara, M. Genome-wide characterization of major intrinsic proteins in four grass plants and their non-aqua transport selectivity profiles with comparative perspective. *PLoS ONE* **2016**, *11*, e0157735. [CrossRef] [PubMed]
10. Gupta, A.B.; Sankararamakrishnan, R. Genome-wide analysis of major intrinsic proteins in the tree plant *Populus trichocarpa*: Characterization of XIP subfamily of aquaporins from evolutionary perspective. *BMC Plant Biol.* **2009**, *9*, 134. [CrossRef]
11. Johanson, U.; Karlsson, M.; Johansson, I.; Gustavsson, S.; Sjövall, S.; Fraysse, L.; Weig, A.R.; Kjellbom, P. The complete set of genes encoding major intrinsic proteins in Arabidopsis provides a framework for a new nomenclature for major intrinsic proteins in plants. *Plant Physiol.* **2001**, *126*, 1358–1369. [CrossRef] [PubMed]
12. Sakurai, J.; Ishikawa, F.; Yamaguchi, T.; Uemura, M.; Maeshima, M. Identification of 33 rice aquaporin genes and analysis of their expression and function. *Plant Cell Physiol.* **2005**, *46*, 1568–1577. [CrossRef] [PubMed]
13. Ishibashi, K.; Hara, S.; Kondo, S. Aquaporin water channels in mammals. *Clin. Exp. Nephrol.* **2009**, *13*, 107–117. [CrossRef] [PubMed]
14. Danielson, J.Å.; Johanson, U. Unexpected complexity of the aquaporin gene family in the moss *Physcomitrella patens*. *BMC Plant Biol.* **2008**, *8*, 45. [CrossRef] [PubMed]
15. Bienert, G.P.; Bienert, M.D.; Jahn, T.P.; Boutry, M.; Chaumont, F. Solanaceae XIPs are plasma membrane aquaporins that facilitate the transport of many uncharged substrates. *Plant J.* **2011**, *66*, 306–317. [CrossRef]
16. Hachez, C.; Chaumont, F. Aquaporins: A family of highly regulated multifunctional channels. *Adv. Exp. Med. Biol.* **2010**, *679*, 1–17.

17. Hove, R.M.; Bhave, M. Plant aquaporins with non-aqua functions: Deciphering the signature sequences. *Plant Mol. Biol.* **2011**, *75*, 413–430. [CrossRef]
18. Di Giorgio, J.P.; Soto, G.; Alleva, K.; Jozefkowicz, C.; Amodeo, G.; Muschietti, J.P.; Ayub, N.D. Prediction of aquaporin function by integrating evolutionary and functional analyses. *J. Membr. Biol.* **2014**, *247*, 107–125. [CrossRef]
19. Mukhopadhyay, R.; Bhattacharjee, H.; Rosen, B.P. Aquaglyceroporins: Generalized metalloid channels. *Biochim. Biophys. Acta -Gen. Subj.* **2014**, *1840*, 1583–1591. [CrossRef]
20. Jahn, T.P.; Møller, A.L.; Zeuthen, T.; Holm, L.M.; Klærke, D.A.; Mohsin, B.; Kühlbrandt, W.; Schjoerring, J.K. Aquaporin homologues in plants and mammals transport ammonia. *FEBS Lett.* **2004**, *574*, 31–36. [CrossRef]
21. Uehlein, N.; Lovisolo, C.; Siefritz, F.; Kaldenhoff, R. The tobacco aquaporin NtAQP1 is a membrane CO_2 pore with physiological functions. *Nature* **2003**, *425*, 734. [CrossRef] [PubMed]
22. Zwiazek, J.J.; Xu, H.; Tan, X.; Navarro-Ródenas, A.; Morte, A. Significance of oxygen transport through aquaporins. *Sci. Rep.* **2017**, *7*, 40411. [CrossRef] [PubMed]
23. Byrt, C.S.; Zhao, M.; Kourghi, M.; Bose, J.; Henderson, S.W.; Qiu, J.; Gilliham, M.; Schultz, C.; Schwarz, M.; Ramesh, S.A. Non-selective cation channel activity of aquaporin AtPIP2;1 regulated by Ca^{2+} and pH. *Plant Cell Environ.* **2017**, *40*, 802–815. [CrossRef] [PubMed]
24. Chaumont, F.; Tyerman, S.D. Aquaporins: Highly regulated channels controlling plant water relations. *Plant Physiol.* **2014**, *164*, 1600–1618. [CrossRef] [PubMed]
25. Azad, A.K.; Sawa, Y.; Ishikawa, T.; Shibata, H. Phosphorylation of plasma membrane aquaporin regulates temperature-dependent opening of tulip petals. *Plant Cell Physiol.* **2004**, *45*, 608–617. [CrossRef]
26. Muto, Y.; Segami, S.; Hayashi, H.; Sakurai, J.; Murai-Hatano, M.; Hattori, Y.; Ashikari, M.; Maeshima, M. Vacuolar proton pumps and aquaporins involved in rapid internode elongation of deep water rice. *Biosci. Biotechnol. Biochem.* **2011**, *75*, 114–122. [CrossRef]
27. Reinhardt, H.; Hachez, C.; Bienert, M.D.; Beebo, A.; Swarup, K.; Voß, U.; Bouhidel, K.; Frigerio, L.; Schjoerring, J.K.; Bennett, M.J. Tonoplast aquaporins facilitate lateral root emergence. *Plant Physiol.* **2016**, *170*, 1640–1654. [CrossRef]
28. Uehlein, N.; Otto, B.; Hanson, D.T.; Fischer, M.; McDowell, N.; Kaldenhoff, R. Function of *Nicotiana tabacum* aquaporins as chloroplast gas pores challenges the concept of membrane CO_2 permeability. *Plant Cell* **2008**, *20*, 648–657. [CrossRef] [PubMed]
29. Maurel, C.; Boursiac, Y.; Luu, D.-T.; Santoni, V.; Shahzad, Z.; Verdoucq, L. Aquaporins in plants. *Physiol. Rev.* **2015**, *95*, 1321–1358. [CrossRef]
30. Chaumont, F.; Moshelion, M.; Daniels, M.J. Regulation of plant aquaporin activity. *Biol. Cell* **2005**, *97*, 749–764. [CrossRef]
31. Murata, K.; Mitsuoka, K.; Hirai, T.; Walz, T.; Agre, P.; Heymann, J.B.; Engel, A.; Fujiyoshi, Y. Structural determinants of water permeation through aquaporin-1. *Nature* **2000**, *407*, 599–605. [CrossRef] [PubMed]
32. Fu, D.; Libson, A.; Miercke, L.J.; Weitzman, C.; Nollert, P.; Krucinski, J.; Stroud, R.M. Structure of a glycerol-conducting channel and the basis for its selectivity. *Science* **2000**, *290*, 481–486. [CrossRef]
33. Sui, H.; Han, B.-G.; Lee, J.K.; Walian, P.; Jap, B.K. Structural basis of water-specific transport through the AQP1 water channel. *Nature* **2001**, *414*, 872–878. [CrossRef] [PubMed]
34. Heymann, J.B.; Engel, A. Structural clues in the sequences of the aquaporins. *J. Mol. Biol.* **2000**, *295*, 1039–1053. [CrossRef] [PubMed]
35. Froger, A.; Thomas, D.; Delamarche, C.; Tallur, B. Prediction of functional residues in water channels and related proteins. *Protein Sci.* **1998**, *7*, 1458–1468. [CrossRef] [PubMed]
36. Clarkson, J.J.; Dodsworth, S.; Chase, M.W. Time-calibrated phylogenetic trees establish a lag between polyploidisation and diversification in *Nicotiana* (Solanaceae). *Plant Syst. Evol.* **2017**, *303*, 1001–1012. [CrossRef]
37. Sierro, N.; Battey, J.N.; Ouadi, S.; Bakaher, N.; Bovet, L.; Willig, A.; Goepfert, S.; Peitsch, M.C.; Ivanov, N.V. The tobacco genome sequence and its comparison with those of tomato and potato. *Nat. Commun.* **2014**, *5*, 3833. [CrossRef]
38. Sierro, N.; Battey, J.N.; Ouadi, S.; Bovet, L.; Goepfert, S.; Bakaher, N.; Peitsch, M.C.; Ivanov, N.V. Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. *Genome Biol.* **2013**, *14*, 2013–2014. [CrossRef]

39. Nagata, T.; Nemoto, Y.; Hasezawa, S. Tobacco BY-2 cell line as the "HeLa" cell in the cell biology of higher plants. *Int. Rev. Cytol.* **1992**, *132*, 1–30.
40. Ishikawa, F.; Suga, S.; Uemura, T.; Sato, M.H.; Maeshima, M. Novel type aquaporin SIPs are mainly localized to the ER membrane and show cell-specific expression in *Arabidopsis thaliana*. *FEBS Lett.* **2005**, *579*, 5814–5820. [CrossRef]
41. Ishibashi, K. Aquaporin subfamily with unusual NPA boxes. *Biochim. Et Biophys. Acta (Bba)-Biomembr.* **2006**, *1758*, 989–993. [CrossRef] [PubMed]
42. Kirscht, A.; Kaptan, S.S.; Bienert, G.P.; Chaumont, F.; Nissen, P.; de Groot, B.L.; Kjellbom, P.; Gourdon, P.; Johanson, U. Crystal structure of an ammonia-permeable aquaporin. *PLOS Biol* **2016**, *14*, e1002411. [CrossRef] [PubMed]
43. Bienert, M.D.; Muries, B.; Crappe, D.; Chaumont, F.; Bienert, G.P. Overexpression of X Intrinsic Protein 1; 1 in *Nicotiana tabacum* and Arabidopsis reduces boron allocation to shoot sink tissues. *Plant Direct* **2019**, *3*, e00143. [CrossRef]
44. Vasilev, N.; Grömping, U.; Lipperts, A.; Raven, N.; Fischer, R.; Schillberg, S. Optimization of BY-2 cell suspension culture medium for the production of a human antibody using a combination of fractional factorial designs and the response surface method. *Plant Biotechnol. J.* **2013**, *11*, 867–874. [CrossRef] [PubMed]
45. Santos, R.B.; Abranches, R.; Fischer, R.; Sack, M.; Holland, T. Putting the spotlight back on plant suspension cultures. *Front. Plant Sci.* **2016**, *7*, 297. [CrossRef] [PubMed]
46. Magy, B.; Tollet, J.; Laterre, R.; Boutry, M.; Navarre, C. Accumulation of secreted antibodies in plant cell cultures varies according to the isotype, host species and culture conditions. *Plant Biotechnol. J.* **2014**, *12*, 457–467. [CrossRef]
47. De Muynck, B.; Navarre, C.; Nizet, Y.; Stadlmann, J.; Boutry, M. Different subcellular localization and glycosylation for a functional antibody expressed in *Nicotiana tabacum* plants and suspension cells. *Transgenic Res.* **2009**, *18*, 467–482. [CrossRef]
48. Navarre, C.; Smargiasso, N.; Duvivier, L.; Nader, J.; Far, J.; De Pauw, E.; Boutry, M. N-Glycosylation of an IgG antibody secreted by *Nicotiana tabacum* BY-2 cells can be modulated through co-expression of human β-1, 4-galactosyltransferase. *Transgenic Res.* **2017**, *26*, 375–384. [CrossRef]
49. Roulin, A.; Auer, P.L.; Libault, M.; Schlueter, J.; Farmer, A.; May, G.; Stacey, G.; Doerge, R.W.; Jackson, S.A. The fate of duplicated genes in a polyploid plant genome. *Plant J.* **2013**, *73*, 143–153. [CrossRef]
50. Levy, A.A.; Feldman, M. Genetic and epigenetic reprogramming of the wheat genome upon allopolyploidization. *Biol. J. Linn. Soc.* **2004**, *82*, 607–613. [CrossRef]
51. Gerbeau, P.; Güçlü, J.; Ripoche, P.; Maurel, C. Aquaporin Nt-TIPa can account for the high permeability of tobacco cell vacuolar membrane to small neutral solutes. *Plant J.* **1999**, *18*, 577–587. [CrossRef] [PubMed]
52. Bots, M.; Vergeldt, F.; Wolters-Arts, M.; Weterings, K.; van As, H.; Mariani, C. Aquaporins of the PIP2 class are required for efficient anther dehiscence in tobacco. *Plant Physiol.* **2005**, *137*, 1049–1056. [CrossRef] [PubMed]
53. Flexas, J.; Ribas-Carbó, M.; Hanson, D.T.; Bota, J.; Otto, B.; Cifre, J.; McDowell, N.; Medrano, H.; Kaldenhoff, R. Tobacco aquaporin NtAQP1 is involved in mesophyll conductance to CO_2 in vivo. *Plant J.* **2006**, *48*, 427–439. [CrossRef] [PubMed]
54. Otto, B.; Uehlein, N.; Sdorra, S.; Fischer, M.; Ayaz, M.; Belastegui-Macadam, X.; Heckwolf, M.; Lachnit, M.; Pede, N.; Priem, N. Aquaporin tetramer composition modifies the function of tobacco aquaporins. *J. Biol. Chem.* **2010**, *285*, 31253–31260. [CrossRef]
55. Siefritz, F.; Biela, A.; Eckert, M.; Otto, B.; Uehlein, N.; Kaldenhoff, R. The tobacco plasma membrane aquaporin NtAQP1. *J. Exp. Bot.* **2001**, *52*, 1953–1957. [CrossRef]
56. Azad, A.K.; Yoshikawa, N.; Ishikawa, T.; Sawa, Y.; Shibata, H. Substitution of a single amino acid residue in the aromatic/arginine selectivity filter alters the transport profiles of tonoplast aquaporin homologs. *Biochim. Biophys. Acta -Biomembr.* **2012**, *1818*, 1–11. [CrossRef]
57. Beitz, E.; Wu, B.; Holm, L.M.; Schultz, J.E.; Zeuthen, T. Point mutations in the aromatic/arginine region in aquaporin 1 allow passage of urea, glycerol, ammonia, and protons. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 269–274. [CrossRef]
58. Soto, G.; Alleva, K.; Mazzella, M.A.; Amodeo, G.; Muschietti, J.P. AtTIP1;3 and AtTIP5;1, the only highly expressed Arabidopsis pollen-specific aquaporins, transport water and urea. *FEBS Lett.* **2008**, *582*, 4077–4082. [CrossRef]

59. Berny, M.C.; Gilis, D.; Rooman, M.; Chaumont, F. Single mutations in the transmembrane domains of maize plasma membrane aquaporins affect the activity of monomers within a heterotetramer. *Mol. Plant* **2016**, *9*, 986–1003. [CrossRef]
60. Maurel, C.; Verdoucq, L.; Rodrigues, O. Aquaporins and plant transpiration. *Plant Cell Environ.* **2016**, *39*, 2580–2587. [CrossRef]
61. Takano, J.; Wada, M.; Ludewig, U.; Schaaf, G.; Von Wirén, N.; Fujiwara, T. The Arabidopsis major intrinsic protein NIP5; 1 is essential for efficient boron uptake and plant development under boron limitation. *Plant Cell Online* **2006**, *18*, 1498–1509. [CrossRef] [PubMed]
62. Ma, J.F. Silicon transporters in higher plants. *Adv. Exp. Med. Biol.* **2010**, *679*, 99–109. [PubMed]
63. Moshelion, M.; Hachez, C.; Ye, Q.; Cavez, D.; Bajji, M.; Jung, R.; Chaumont, F. Membrane water permeability and aquaporin expression increase during growth of maize suspension cultured cells. *Plant Cell Environ.* **2009**, *32*, 1334–1345. [CrossRef] [PubMed]
64. Zelazny, E.; Borst, J.W.; Muylaert, M.; Batoko, H.; Hemminga, M.A.; Chaumont, F. FRET imaging in living maize cells reveals that plasma membrane aquaporins interact to regulate their subcellular localization. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 12359–12364. [CrossRef]
65. Reisen, D.; Leborgne-Castel, N.; Özalp, C.; Chaumont, F.; Marty, F. Expression of a cauliflower tonoplast aquaporin tagged with GFP in tobacco suspension cells correlates with an increase in cell size. *Plant Mol. Biol.* **2003**, *52*, 387–400. [CrossRef]
66. Pierman, B.; Toussaint, F.; Bertin, A.; Lévy, D.; Smargiasso, N.; De Pauw, E.; Boutry, M. Activity of the purified plant ABC transporter NtPDR1 is stimulated by diterpenes and sesquiterpenes involved in constitutive and induced defenses. *J. Biol. Chem.* **2017**, *292*, 19491–19502. [CrossRef]
67. Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **2016**, *33*, 1870–1874. [CrossRef]
68. Li, B.; Dewey, C.N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [CrossRef]

International Journal of
Molecular Sciences

MDPI

Article

Gene Expression Analysis and Metabolite Profiling of Silymarin Biosynthesis during Milk Thistle (*Silybum marianum* (L.) Gaertn.) Fruit Ripening

Samantha Drouet [1,2], Duangjai Tungmunnithum [1,2,3], Éric Lainé [1,2] and Christophe Hano [1,2,*]

[1] Laboratoire de Biologie des Ligneux et des Grandes Cultures (LBLGC), INRAE USC1328, University of Orleans, 21 rue de Loigny la Bataille, F-28000 Chartres, France; samantha.drouet@univ-orleans.fr (S.D.); duangjai.tun@mahidol.ac.th (D.T.); eric.laine@univ-orleans.fr (É.L.)

[2] Bioactifs et Cosmétiques, CNRS GDR3711, CEDEX 2, 45067 Orléans, France

[3] Department of Pharmaceutical Botany, Faculty of Pharmacy, Mahidol University, 447 Sri-Ayuthaya Road, Rajathevi, Bangkok 10400, Thailand

[*] Correspondence: hano@univ-orleans.fr; Tel.: +33-237-309-753; Fax: +33-237-910-863

Received: 14 June 2020; Accepted: 29 June 2020; Published: 2 July 2020

Abstract: Mature fruits (i.e., achenes) of milk thistle (*Silybum marianum* (L.) Gaertn., Asteraceae) accumulate high amounts of silymarin (SILM), a complex mixture of bioactive flavonolignans deriving from taxifolin. Their biological activities in relation with human health promotion and disease prevention are well described. However, the conditions of their biosynthesis in planta are still obscure. To fill this gap, fruit development stages were first precisely defined to study the accumulation kinetics of SILM constituents during fruit ripening. The accumulation profiles of the SILM components during fruit maturation were determined using the LC-MS analysis of these defined developmental phases. The kinetics of phenylalanine ammonia-lyase (PAL), chalcone synthase (CHS) and peroxidase (POX) activities suggest in situ biosynthesis of SILM from L-Phenylalanine during fruit maturation rather than a transport of precursors to the achene. In particular, in contrast to laccase activity, POX activity was associated with the accumulation of silymarin, thus indicating a possible preferential involvement of peroxidase(s) in the oxidative coupling step leading to flavonolignans. Reference genes have been identified, selected and validated to allow accurate gene expression profiling of candidate biosynthetic genes (*PAL*, *CAD*, *CHS*, *F3H*, *F3'H* and *POX*) related to SILM accumulation. Gene expression profiles were correlated with SILM accumulation kinetic and preferential location in pericarp during *S. marianum* fruit maturation, reaching maximum biosynthesis when desiccation occurs, thus reinforcing the hypothesis of an in situ biosynthesis. This observation led us to consider the involvement of abscisic acid (ABA), a key phytohormone in the control of fruit ripening process. ABA accumulation timing and location during milk thistle fruit ripening appeared in line with a potential regulation of the SLIM accumulation. A possible transcriptional regulation of SILM biosynthesis by ABA was supported by the presence of ABA-responsive cis-acting elements in the promoter regions of the SILM biosynthetic genes studied. These results pave the way for a better understanding of the biosynthetic regulation of SILM during the maturation of *S. marianum* fruit and offer important insights to better control the production of these medicinally important compounds.

Keywords: abscisic acid; flavonolignans; fruit development; gene expression; metabolite profiling; *Silybum marianum*; silymarin

1. Introduction

Silybum marianum ((L.) Gaertn.) is an annual or biennial Asteraceae plant native to the Mediterranean region. It is currently widespread in Southern Europe, Western Asia, North Africa, Australia and North America. It is one of the oldest medicinal plants, already used in Greek and Roman medicines to help digestion or to treat liver and/or gallbladder disorders [1]. In the Middle Ages, its preparation extracts have been used to cure melancholy or black bile, both associated with various hepatic dysfunctions [2]. Nowadays, there is still a great deal of interest in this plant extract for medicinal and cosmetic applications, mainly due to the high accumulation in its fruits of the so-called silymarin (SILM), a complex mixture of flavonoids and flavonolignans (Figure 1). Its efficacy in the treatment of several diseases including liver disorders [1,3,4], inflammatory reactions [5,6] and oxidative stress protection [7–10] has been demonstrated. Most of these properties depended on the accumulation of silybins, a mixture of flavonolignan diastereoisomers [11].

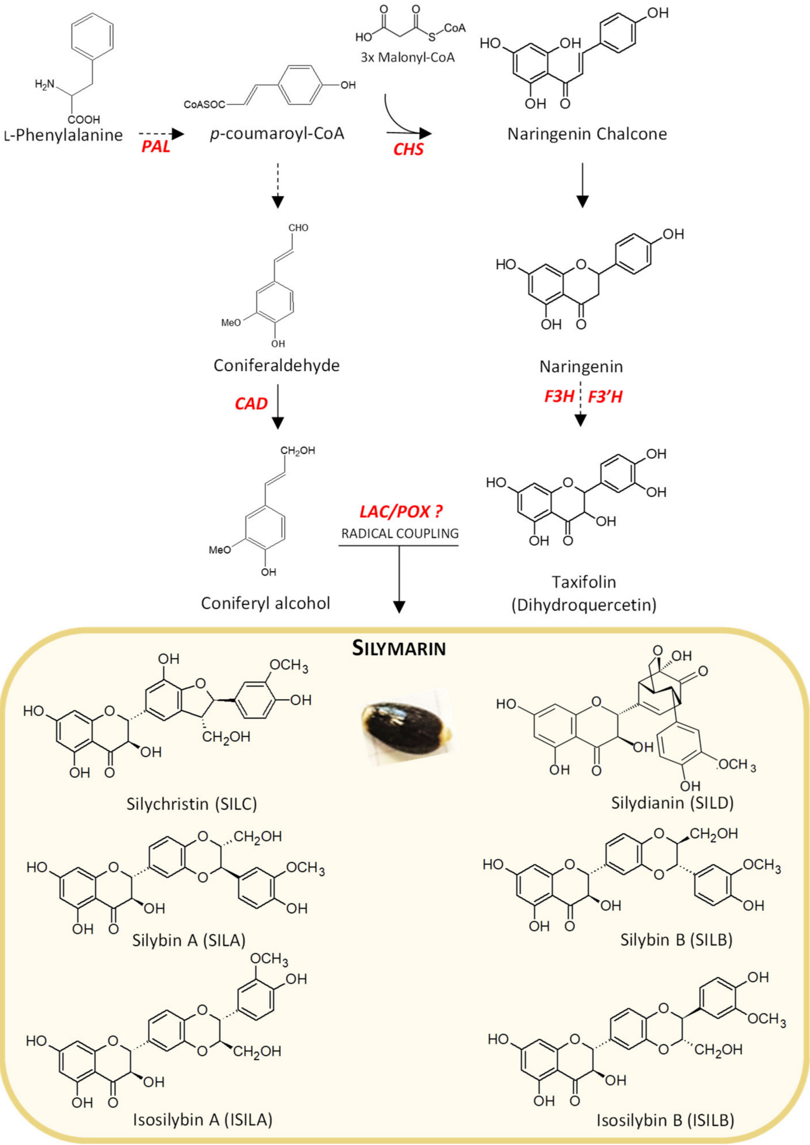


Figure 1. Partial scheme of the silymarin (SILM) flavonolignans biosynthesis pathway in *S. marianum*. Flavonolignans are mainly accumulated in mature achenes (yellow box). In red are presented genes potentially involved in this pathway: *PAL* (*L-phenylalanine ammonia-lyase*), *CAD* (*cinnamyl alcohol dehydrogenase*), *CHS* (*chalcone synthase*), *flavanone 3-dioxygenase* (*F3H*), *flavone 3′-hydroxylase* (*F3′H*), *LAC* (*laccases*) and *POX* (*peroxidases*). The dotted arrows indicate a single step, while dashed arrows indicate several steps in the metabolic pathway while the full arrows indicate direct synthesis of the compound through an enzyme.

The main flavonolignans isolated from milk thistle are derived from flavonoid taxifolin (TAX): silybin A (SILA), silybin B (SILB), isosilybin A (ISILA), isosilybin B (ISILB), silychristin (SILC), and silydianin (SILD) [11–13]. Other types of flavonolignans have been described in many plant species in the literature such as in *Oryza sativa* [14], *Onopordon corymbosum* [15], *Sasa veitchii* [16], *Lepidium meyenii* [17] or *Hydnocarpus anthelmintica* [18]. But milk thistle fruit remains, by far, the primary source of flavonolignans derived from the oxidative coupling of TAX and *E*-coniferyl alcohol and the unique SILM flavonolignans bioactive mixture.

In *S. marianum*, SILM flavonolignans are accumulated in the fruit (i.e., achene, sometimes incorrectly referred to as seed in the literature) [19]. Elucidation of biosynthetic steps and regulatory mechanisms leading to the production of SILM flavonolignans in milk thistle fruit is challenging as few genomic sequences from this species are available. Taking advantage of our current knowledge of the phenylpropanoid biosynthetic pathway in plants, it may be possible to propose a putative biosynthetic sequence leading to SILM flavonolignans and to point some key structural genes (Figure 1). The supposed biosynthetic sequence involved the deamination of L-Phe into trans-cinnamic acid by L-Phe ammonia-lyase (phenylalanine ammonia-lyase (PAL), EC 4.3.1.5), a branch-point enzyme between primary and secondary metabolism in plants. Trans-cinnamic acid is a crucial precursor to several phenylpropanoid derivatives including *E*-coniferyl alcohol in a metabolic sequence involving cinnamyl alcohol dehydrogenase (CAD, EC 1.1.1.195), and TAX in a metabolic sequence involving chalcone synthase (CHS, EC 2.3.1.74), flavanone 3-dioxygenase (F3H, EC 1.14.11.12) and flavone 3′-hydroxylase (F3′H, EC 1.14.13.88). Lastly, oxidative coupling reactions between TAX and E-coniferyl alcohol mediated by laccase(s) (LAC(s), EC 1.10.3.2) and/or peroxidase(s)(POX(s), EC 1.11.1.x) lead to different SILM flavonolignans (Figure 1). It is assumed that their biosynthesis resulting from the oxidative coupling between *E*-coniferyl alcohol and TAX could take place at this site of accumulation. The in vitro biochemical characterization of an ascorbate peroxidase enzyme (APX1) involved in the production of SILA/B and ISILA/B has been reported [20].

To date, however, no information is available on the production of other essential SILM flavonolignans, such as SILD and SILC. Biosynthesis of the latter has been proposed to involve a separate, and possibly more complex, oxidative coupling process that could infer the involvement of other POX or even LAC as well as of dirigent proteins [21,22], already described to direct the stereoselective biosynthesis of lignans in many plant species [23,24]. Besides, with the biosynthesis of *E*-coniferyl alcohol in the pericarp, a complex spatial organization has been suggested, with TAX biosynthesis suggested to be located in the flower [20], thus requiring the transport of this latter to the pericarp. However, this hypothesis is based on an interpretation of gene expression data limited to one single stage of fruit development. Therefore, a full-length analysis of the *S. marianum* fruit development process would be appropriate to check the validity of this hypothesis. To take a critical step towards better monitoring and understanding of SILM biosynthesis, it is therefore important to have a precise description of the various developmental stages during fruit maturation. Little attention has been paid to this point [25–27], and so little is known to date about the precise timing of accumulation of SILM during *S. marianum* fruit development. The lack of validation of reference genes for the study of gene expression in milk thistle is also an obstacle to the understanding of SILM biosynthesis regulation during achene development.

More detailed information on the spatiotemporal production of SILM flavonolignans, as well as on the regulation of their biosynthesis, could provide important information to optimize their accumulation. Both in the model plant *Arabidopsis thaliana* and crops, such as flax (*Linum usitatissimum* L.), it has been shown that abscisic acid (ABA) acts as a key phytohormone involved in the control of many aspects of seed development and phenylpropanoid biosynthesis. ABA has been shown to regulate several genes associated with fruit and seed maturation, and stress response including flavonoid and lignan biosynthesis [28–30]. Therefore, specific interest may be brought to the timing and position of ABA accumulation in *S. marianum* achene.

In the present study, six stages of development of *S. marianum* achenes were first adequately defined, enabling us to establish the precise timing and location SILM biosynthesis during fruit ripening. The enzymatic activities of PAL, CHS, LAC and POX during maturation were also determined in parallel. We identified reference genes from genomic data, selected and validated to follow accurately by RT-qPCR the spatiotemporal gene expression of candidate (i.e., *PAL*, *CAD*, *CHS*, *F3H*, *F3'H* and *POX*) genes potentially involved in flavonolignans biosynthesis. The linkage between gene expression, enzymatic activity, and accumulation of SILM was studied. Finally, ABA's potential role was assessed by determining its accumulation profile and by analyzing the expression of genes involved both in its biosynthesis (*ABA1*) and signaling (*LEC2*).

2. Results and Discussion

2.1. Morphological Characterization of Milk Thistle Achene Development

Fruit production stages have been described as one of the most critical processes in plant life. Such developmental stages may be defined as morphological features of a capitula or achene, or in days after anthesis (DAA) or flowering (DAF) [29]. However, since achene development may be affected by numerous factors including genetic factors (varieties / ecotypes) as well as environmental factors (e.g., light, temperature, and soil properties) and growing conditions (outdoor vs. greenhouse conditions), we decided to identify milk thistle fruit developmental stages according to achene morphological characteristics. Under greenhouse conditions, a complete development cycle leading to fully mature *S. marianum* fruits was achieved in about 50 days after flowering and six different developmental stages were described based on morphological characteristics of achene (Figure 2).

First, at stage 1 (DAF7), the achene was white / cream with no visible seed, the pericarp began to develop and accounted for the total weight of the achene. The fresh weight ratio (FW) to dry weight (DW) was high at this point, with a value around 10 (data not shown).

Subsequently, pericarp continued to develop gradually with stage 2 (DAF10) and stage 3 (DAF15), with a doubling of the FW value between stage 1 and stage 3 of development and the seed at stage 3 of development. At this stage of development, the pericarp color began to become pink. The ratio of FW to DW was still high.

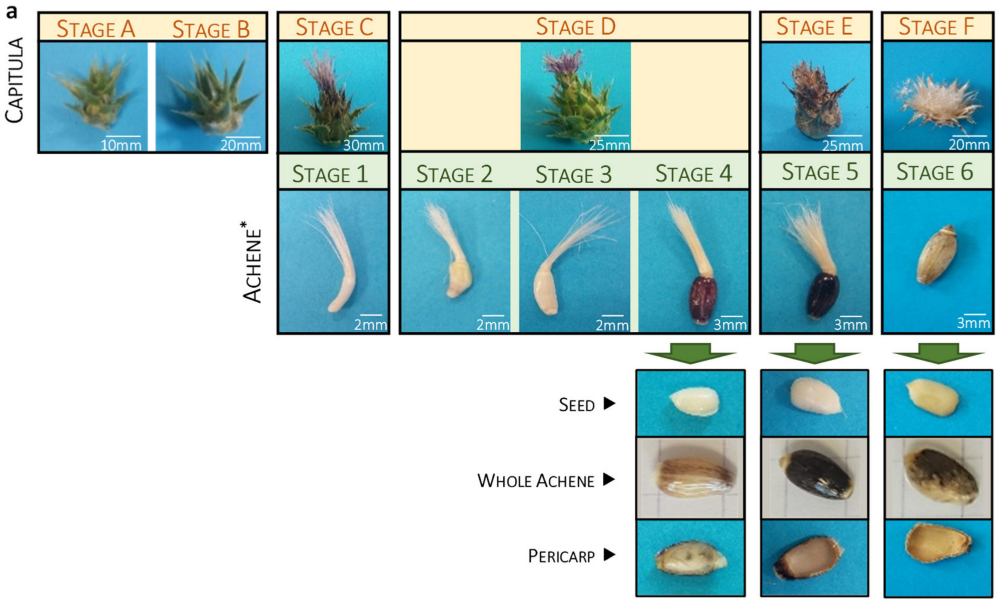
At stage 4 (DAF24) the pericarp became purple and the white seed showed an active growth of around 20% of the achene weight. The ratio of FW to DW began to decrease, thus highlighting the transition from morphogenesis to maturation, with the achene having reached its final length.

At stage 5 (DAF32), the white seed accounted for about 50 % of the achene weight, the FW/DW ratio continued to decrease, confirming the start of fruit desiccation, and the achene color was dark brown.

Finally, stage 6 (DAF50) corresponded to the mature achene, with the light beige seed accounting for more than 55% of the achene weight. The observed decreases in DW up to the minimum values confirmed the completion of the desiccation. We also noted the lightening of pericarp as a result of the appearance of air bubbles in-between the most external cell layers of pericarp (Figure 2a).

The seed showed the presence of two cotyledons and has shown rapid growth from stage 3 to 6 of development, while the pericarp weight has gradually decreased from 30 to 13 mg from stage 4 to 6 of development. Note that the apparent capitula (flower head) morphology was also presented with the corresponding defined developmental stage in Figure 2a to facilitate sampling. For example, a capitula with a pink pappus (stage C) corresponding to the developmental stage 1 of fruit. Note, however, that capitula morphology is not predictive of the developmental stage of the fruit, since it is later centripetal and not completely synchronous. This further confirms the importance of precisely defining the stages of development and of breaking the capitula before the analysis is carried out.

In the literature, in addition to the mature stage, three further stages of maturation have been described in terms of the appearance of capitula (i.e., early flowering, mid-flowering with dry flowers, late flowering with dry flowers and dehiscence of capitula) [25–27], corresponding here to fruit colors ranging from white/cream, purple to (dark) brown. If these stages allow for a general view of maturation, they do not allow for a precise characterization of achene development. Here, six different achene ripening stages were precisely defined (Figure 2), thus simplifying the sampling method and increasing the accuracy of the molecular and (bio)chemical analysis of the tissue.



b

DEVELOPMENT STAGES		DAY AFTER FLOWERING	WHOLE ACHENE DW (mg)	PERICARP DW (mg)	SEED DW (mg)	LENGHT (mm)	MORPHOLOGICAL FEATURES
WA1		7	4.83±0.20 f	4.83±0.20 f	-	3.8	White akene without visible embryo
WA2		10	7.70±0.36 e	7.70±0.36 e	-	5	White akene without visible embryo
WA3		15	9.53±0.09 d	9.53±0.09 d	-	6	White akene without visible embryo
WA4		24	38.10±1.68 b	30.17±1.79 a	7.93±0.27 c	6	Akene color purple/brown
WA5		32	45.50±0.47 a	23.83±1.97 b	21.67±1.65 a	6	Color of akene darker
WA6		50	29.13±1.28 c	13.33±0.84 c	15.80±1.40 b	6.3	Dry mature akene

Figure 2. Development stages of the *S. marianum* achene defined according to their morphological characteristic. (**a**) Six achene developmental stages were defined. For the developmental stages 4, 5 and 6 achenes were manually dissected to allow the visualization of both seed and pericarp. The capitula morphology corresponding to each defined achene developmental stage is presented. Note that the capitula morphology is not predictive of the fruit developmental stage (see text for explanations). (**b**) Morphological features of achene maturation during time such as achene, seed, and pericarp length, dry weight (DW) and day of flowering (DAF). Each value represents means ± SD of $n = 10$ independent sampling. Different letters indicate significant differences at $p < 0.05$.

2.2. Accumulation Kinetic of SILM Constituents during S. marianum Fruit Development

Accumulation kinetics of SILM flavonolignans have been studied along the defined developmental stages of the fruit (Figure 3, Table S1). HPLC chromatograms revealed a major accumulation of flavonolignans during achene maturation starting from stage 4 of development (WA 4 (i.e., whole achene at developmental stage 4); Figure 3a, Table S1).

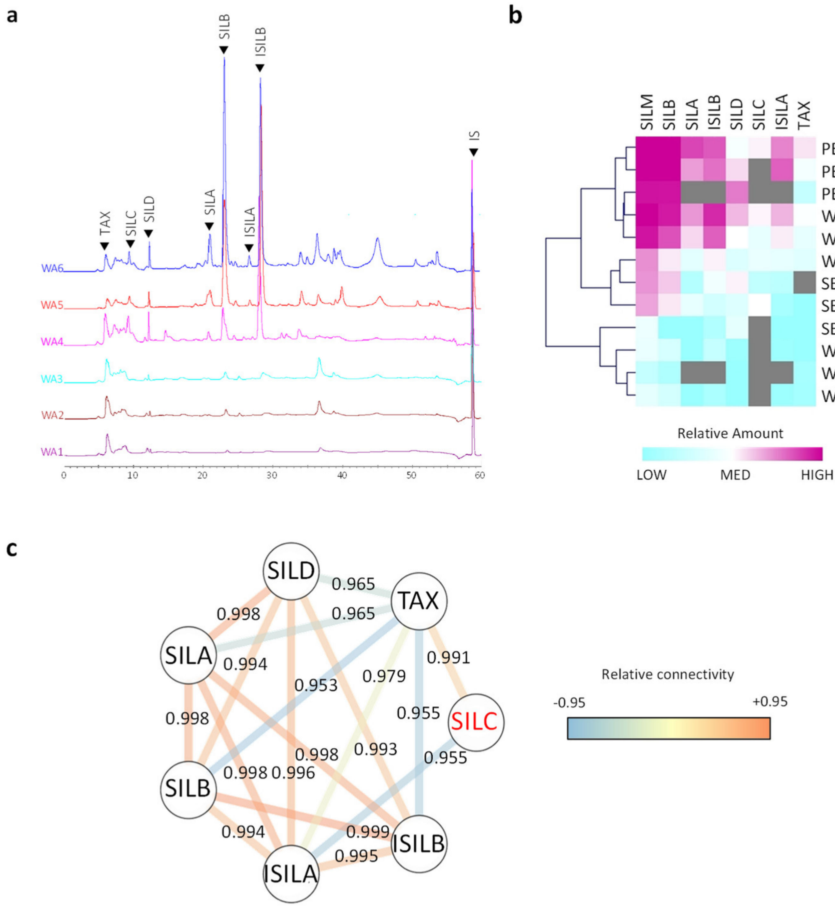


Figure 3. Accumulation of SILM compounds in *S. marianum* during achene development. (**a**) Chromatograms HPLC superposition of all 6 achene developmental stages showing the accumulation of SILM over time by comparison. IS: internal standard (6-methoxyflavone). (**b**) Extraction of compounds in whole achene (WA), pericarp (PER) and seed (SEED) represented in MeV (Multiple Experiment Viewer). Color scale is blue (weak content) to violet (high content) and grey color indicates not detected content. For quantitative values (referred to Table S1 expressed in mg/g DW). Values are means of $n = 3$ independent experiments; (**c**) the metabolite network was constructed using Cytoscape software 3.7, with a 0.95 cut-off value. Color edges from blue, yellow to red indicate increasing strength of the connection between the compounds.

The accumulation of SILM increased dramatically at stage 4 when the desiccation process started. SILM enrichment was observed in pericarp compared to seed (Figure 3B, Table 1, Table S1).

Table 1. Evaluation of SILM and abscisic acid (ABA) contents in whole achenes of *S. marianum* during achene maturation expressed per g DW as well as per achene.

Metabolite	WA1	WA2	WA3	WA4	WA5	WA6
SILM (mg/g DW)	0.14 ± 0.04 ^{e}	0.42 ± 0.03 ^{d}	0.50 ± 0.04 ^{d}	4.58 ± 0.33 ^{c}	24.20 ± 2.12 ^{b}	52.46 ± 2.73 ^{a}
SILM (mg/achene)	0.70 ± 0.21 ^{e}	3.23 ± 0.23 ^{d}	4.78 ± 0.36 ^{d}	174.68 ± 12.46 ^{c}	1100.97 ± 96.55 ^{b}	1528.14 ± 79.59 ^{a}
ABA (ng/g DW)	1.83 ± 0.38 ^{e}	5.57 ± 0.63 ^{d}	13.63 ± 1.42 ^{c}	29.10 ± 1.91 ^{b}	48.57 ± 1.56 ^{a}	46.37 ± 2.15 ^{a}
ABA (ng/achene)	8.86 ± 1.82 ^{f}	42.86 ± 4.88 ^{e}	129.97 ± 13.53 ^{d}	1108.71 ± 72.69 ^{c}	2209.78 ± 70.99 ^{a}	1350.66 ± 62.60 ^{b}

Each value represents means ± SD of at least $n = 3$ independent sampling. Different letters indicate significant differences at $p < 0.05$.

SILM constituents have been quantified in whole achene (WA) for all defined developmental stages of the fruit and in manually separated pericarp and seed from developmental stages 4 to 6 (Figure 3b, Table S1). Accumulation kinetics of each SILM constituent also showed the same spatiotemporal location with accumulation starting from stage 4 of development with maximum values achieved in mature fruit (developmental stage 6) and mainly localized in pericarp. However, SILB, SILD and SILC also accounted for substantial amounts (over 1 mg/g DW) in the seed at the later stage of development. The presence of these three flavonolignans could indicate the potential transport of these molecules from the pericarp to the seed. Since milk thistle oil is rich in polyunsaturated fatty acids [31] that are more prone to oxidation, it can be assumed that the presence of these antioxidant compounds may contribute to the oxidative stability.

SILB and SILD were the first two flavonolignans to be detected in *S. marianum* fruit during its maturation process (Figure 3). In pericarp, these two compounds were also detected in high levels at stage 4 of development, while SILC was detected later (developmental stage 6) in high levels in the same tissue. This difference could suggest the involvement of different enzymes, more complex biosynthesis or a different regulation. Little is known about the biosynthetic sequence leading to the biosynthesis of SILM flavonolignan. The main hypothesis concerning milk thistle flavonolignans biosynthesis suggested an oxidative coupling between TAX and *E*-coniferyl alcohol. In recent years, it has been proposed that the involvement of dirigent proteins should explain the preferential accumulation of some SILM components [22,32]. The Metabolite Network was proposed to evaluate the biochemical connectivity between the various intermediates and/or the branch of a biosynthetic pathway [33]. Except for SILD, the metabolite network showed a strong biochemical connectivity between each compound (Figure 3c). According to this network, SILD only has a high connectivity with TAX (its precursor) and ISILA. It has been proposed that a strong connectivity between the substrate and the product of a considered enzymatic step suggests a weak contribution of this step to the flux control of this biosynthetic pathway and a simpler regulation [33]. These results could, therefore, suggest that there are at least two different regulations for this biosynthetic pathway during *S. marianum* fruit maturation.

2.3. Kinetic Study of Selected Enzymatic Activities Related to SILM Biosynthesis

To gain a deeper insight into the timing of SILM biosynthesis and to discriminate between in situ production or transport, we then determined the activity of PAL, CHS, POX and LAC enzymes during fruit ripening (Figure 4).

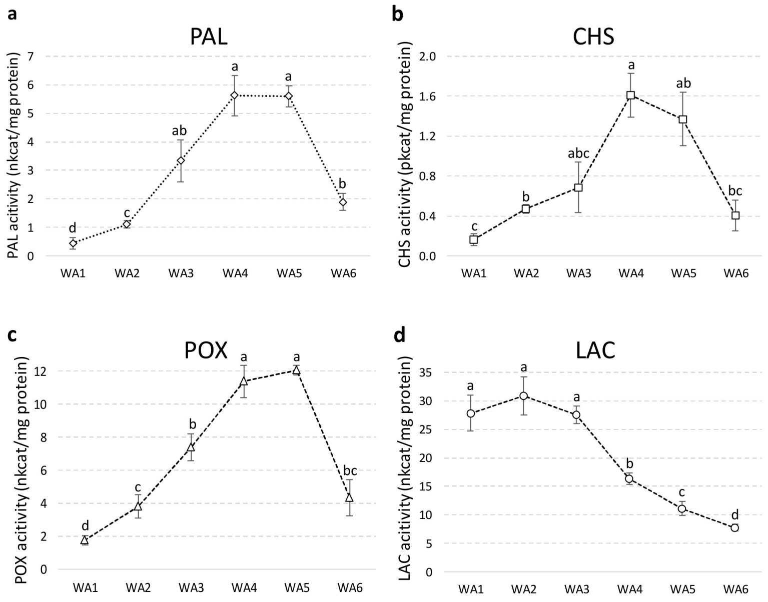


Figure 4. Time course evaluation during fruit development of specific phenylalanine ammonia-lyase (PAL) (**a**), chalcone synthase (CHS) (**b**), peroxidase (POX) (**c**) and laccase (LAC) (**d**) enzymatic activities in the soluble protein fraction from in whole achene of *S. marianum*. Values are the mean ± SD of 3 independent measurements. Different letters indicate significant differences at $p < 0.05$.

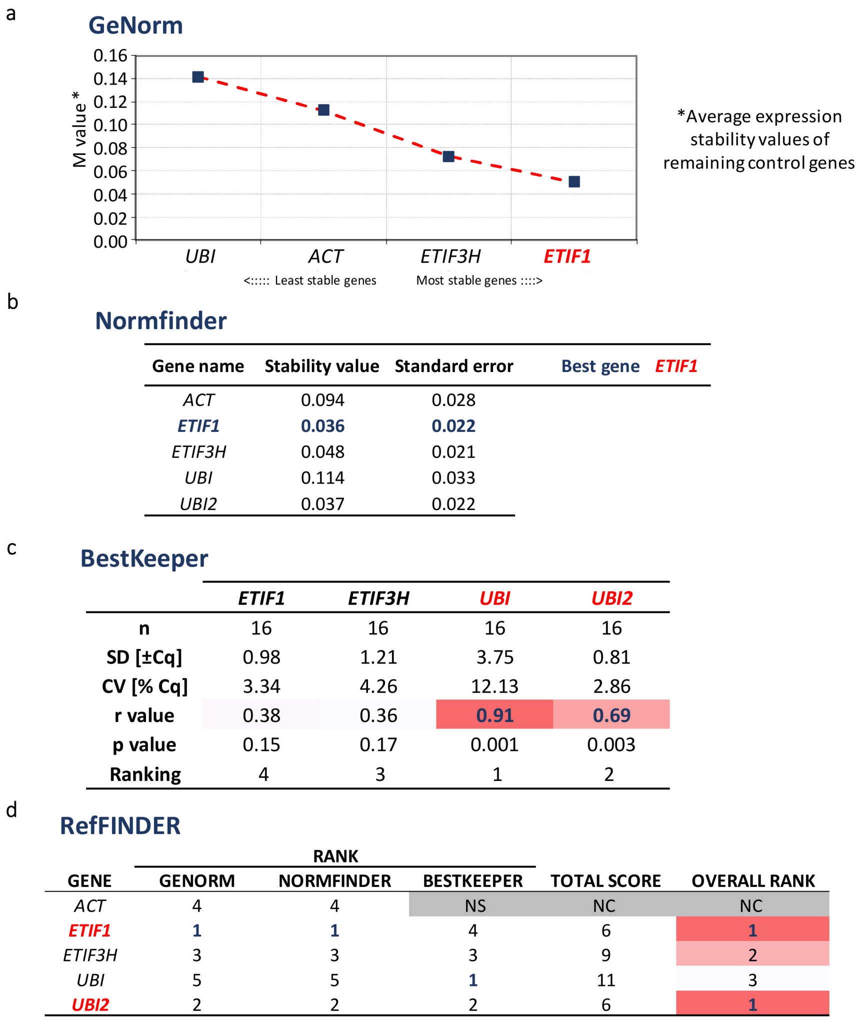
PAL, CHS and POX enzymes showed a similar pattern of activity reaching maximum values at stages 4 and 5 before decreasing in mature fruit (Figure 4a–c). In sharp contrast, LAC activity changed independently of the other enzymes and was high at the early stages of development (Figure 4d). The developmental changes observed in their enzyme activities allow these enzymes to be grouped into two groups that reflect their possible involvement in SILM biosynthesis in *S. marianum*. High and significant correlations between PAL, CHS and POX were calculated, while LAC was not associated with any of these enzymes (Table S2).

The timing of PAL, CHS and POX activities is consistent with SILM accumulation during fruit maturation and could support the involvement of these enzymes in the in situ biosynthesis of SILM flavonolignans. Supportively, the biochemical characterization of one peroxidase active for the formation of SILB, but inactive for the formation of the other *S. marianum* flavonolignans, was presented [20]. Here, the results also favor the involvement of POX in flavonolignan biosynthesis rather than LAC in the final oxidative coupling step. However, the complete sequence of fruit development has been considered, and therefore this hypothesis is further reinforced. *PAL* gene expression has been reported in *S. marianum* fruit at a single stage corresponding to SILM accumulation [20]. Here, the detection of PAL activity indicates that this expression of this gene effectively leads to the production of a functional protein. It also shows that the phenylpropanoid pathway is (at least the first limiting step) active in situ at the time of SILM accumulation. We determined a similar spatio-temporal pattern for CHS activity. Likewise, at three developmental stages based on capitula morphology, Torres and Corchete [25] observed a *CHS* gene expression with a similar timing in *S. marianum*. The two CHS isoforms were also observed in various *S. marianum* organs, including one expressed in the pericarp, following our enzymatic assays. This suggests that the first step of flavonoid biosynthesis, during fruit maturation, is therefore also active in the pericarp. However, Lv et al. [20] suggested the hypothesis for the biosynthesis of the two SILM precursors, *E*-coniferyl alcohol and TAX, of a distinct spatial organization. This hypothesis was based on the analysis of RNAseq data from a single stage of immature fruit development (pericarp vs. seed), collected from outdoor plants 10 days after flowering, and by comparison with root, stem, leaf and flower conditions. First, in situ biosynthesis in the pericarp, at this single stage, of *E*-coniferyl alcohol was supported by the expression detected of several genes involved in its biosynthesis proposed from this study [20]. By contrast, ex situ TAX biosynthesis was proposed because only CHS gene expression was detected in that tissue, while several expressions of several biosynthetic genes (including *F3H* and *F3'H*) were detected in flowers [20]. This has led Lv et al. [20] to propose a separate spatio-temporal organization for the production of the two SILM precursors, including transportation of TAX from petals to pericarp. In sharp contrast, both Torres and Corchete [25] detected both *F3H* and *F3'H* gene expression in immature *S. marianum* fruits at 3 stages of development based on capitula morphology. Moreover, in contrast to this transport hypothesis, previous work observed flavonoid transport in seed was rather limited to intracellular movements between cytoplasm and vacuole, while symplastic interorgan transport was limited to basipetal movement [34,35]. To clarify this discrepancy, taking advantage of these defined developmental stages, our next step was to study the expression time course of SILM biosynthetic genes by RT-qPCR.

2.4. Expression of Genes Involved in Phenolic Compounds Synthesis

2.4.1. Validation of Reference Genes

Before the gene expression analysis of selected biosynthetic genes, validation of reference genes is an essential prerequisite. A preliminary study of the gene expression of "housekeeping" genes should be carried out systematically in the tissues and experimental conditions studied to confirm their stability and to avoid any bias in the results [29,36–38]. Validation of reference genes is a very challenging step in the maturation of seeds and fruits [39]. Therefore, the first consisted in the identification candidate for the selection of reference genes. As a result, we identified 12 candidates in the genome of *S. marianum*. The characteristics of these genes are described in Table S3. The candidate genes not detected under all experimental conditions corresponding to the defined developmental stages have been excluded (Figure S1). Then, we evaluated the remaining selected reference genes using a variety of software (RefFinder, BestKeeper, GeNorm and Normfinder) that allowed us to study and classify their gene expression stability (Figure 5).



Normfinder

Best gene *ETIF1*

Gene name	Stability value	Standard error
ACT	0.094	0.028
ETIF1	**0.036**	**0.022**
ETIF3H	0.048	0.021
UBI	0.114	0.033
UBI2	0.037	0.022

BestKeeper

	ETIF1	*ETIF3H*	*UBI*	*UBI2*
n	16	16	16	16
SD [±Cq]	0.98	1.21	3.75	0.81
CV [% Cq]	3.34	4.26	12.13	2.86
r value	0.38	0.36	**0.91**	**0.69**
p value	0.15	0.17	0.001	0.003
Ranking	4	3	1	2

RefFINDER

	RANK				
GENE	GENORM	NORMFINDER	BESTKEEPER	TOTAL SCORE	OVERALL RANK
ACT	4	4	NS	NC	NC
ETIF1	1	1	4	6	1
ETIF3H	3	3	3	9	2
UBI	5	5	1	11	3
UBI2	2	2	2	6	1

Figure 5. Gene expression stability analysis of the candidate reference genes for RT-qPCR gene expression study in *S. marianum* during fruit maturation according to GeNorm (**a**), NormFinder (**b**), BestKeeper (**c**) as well as RefFINDER ranking result (**d**). NS: not considered as stable enough by the software analysis. NC (Grey): not ranked. Figure 5c,d are represented as heatmap from white (lower stability) to pink (medium stability) and red (higher stability).

From this validation analysis, the two most stable reference genes to normalize the expression SILM candidate genes were *UBI2* and *ETIF1*.

2.4.2. Gene Expression Analysis of Candidate Genes

In *S. marianum*, as in other accumulating plant species, little is known about the regulation of flavonolignans biosynthesis. It is accepted that in *S. marianum*, flavonolignans biosynthesis implies the involvement of different branches of the phenylpropanoid biosynthetic pathway: the general branch leading to *p*-coumaroyl-CoA, from which two specific branches may originate: the monolignol pathway from which the *E*-coniferyl alcohol precursor is produced, and the flavonoid pathway from which the TAX precursor is produced (Figure 1). A final oxidative coupling step occurred between these two precursor moieties, leading to the different SILM flavonolignans. The complete coding sequences of the *PAL*, *CAD*, *CHS*, *F3H*, *F3′H* and *POX* genes were retrieved from the *S. marianum* genomic data to account for each of these metabolic branches leading to SILM flavonolignans. Their characteristics and comparison with the *A. thaliana* orthologous genes [40] are shown in Table S4. Their expression profiles during the development of *S. marianum* fruit established by RT-qPCR are shown in Figure 6.

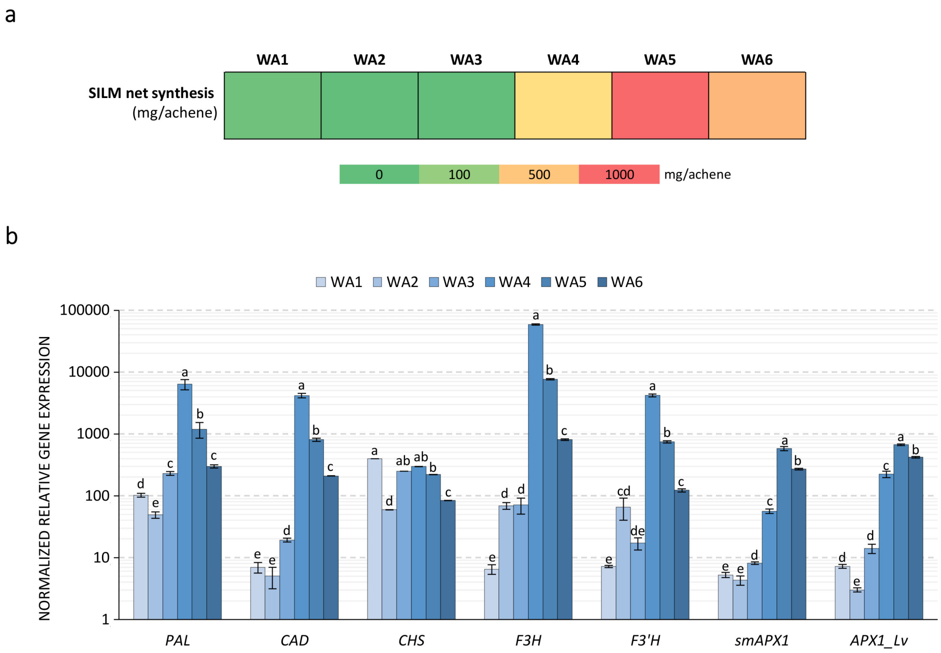
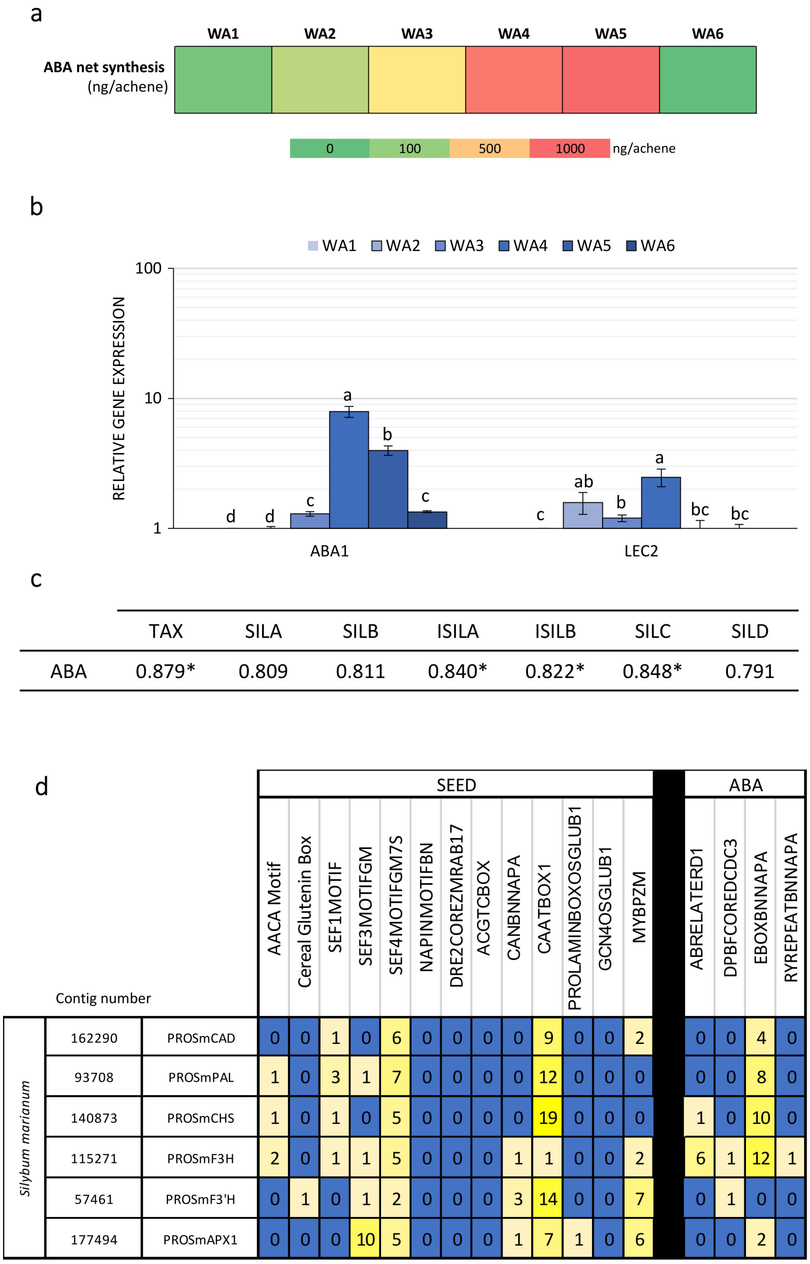


Figure 6. Kinetics of SILM synthetic gene expression during milk thistle achene maturation. (**a**) Kinetics of SILM net production (expressed in mg/achene) at each developmental stage in whole achene (WA) during milk thistle fruit maturation (calculated from Table 1). (**b**) Relative quantification using RT-qPCR of RNA expression of putative genes involved in *S. marianum* flavonolignans biosynthesis (*PAL*, *CAD*, *CHS*, *F3H*, *F3′H*, *smAPX1* and *APX1_Lv*). Values are the mean ± SD of 3 independent measurements. Different letters indicate significant differences at $p < 0.05$.

In agreement with SILM production during fruit maturation (Table 1 and Figure 6a), all genes had a similar pattern of expression with a strong increase in their steady state mRNA levels reaching maximum values at stage 4 before decreasing as with fruit ripening (Figure 6b). These results are consistent with those presented by Torres and Corchete [25]; therefore, they are in favor of the complete in situ biosynthesis of SILM flavonolignans and their precursors. By comparison, Lv et al. [20] did not detect *F3H* and *F3′H* mRNA in immature fruit (10 DAF) using RNAseq as compared to other green vegetative tissue analyzed. This discrepancy may be explained by the difficulty of extracting high-quality RNA and/or proteins from seed tissue as shown by the difficulty of obtaining stable reference genes for this tissue [29,39,41,42]. The expression profile of the ascorbate peroxidase enzyme (APX1) is also consistent with this in vitro biochemical characterization, which showed its ability to synthesize both SILA/B and ISILA/B by Lv et al. [20]. Here, a second *POX* gene, different form the one previously identified and biochemically characterized (Figure S2), was identified from genomic data and its expression profile was consistent with the possible involvement in the biosynthesis of SILC and/or SILD and its derivatives. In addition to the action of oxidase, a more complex stereoselective sequence involving dirigent proteins (DIRs) has been suggested for the biosynthesis of these compounds [21,22]. The presence of different DIRs, for example, is responsible for the stereoselective accumulation of lignans in flax [24,43]. Future work on its biochemical characterization as well as the possible involvement of DIRs to be identified will be undertaken.

2.5. Relationship between Compounds and Kinetics of ABA Content

Many genes associated with the maturation of seeds and fruits or the biosynthesis of phenylpropanoids are regulated by ABA [30]. ABA quantification showed a significant increase during fruit maturation with important accumulation at stage 4 (Figure 7a).



	TAX	SILA	SILB	ISILA	ISILB	SILC	SILD
ABA	0.879*	0.809	0.811	0.840*	0.822*	0.848*	0.791

	Contig number		SEED													ABA			
			AACA Motif	Cereal Glutenin Box	SEF1MOTIF	SEF3MOTIFGM	SEF4MOTIFGM7S	NAPINMOTIFBN	DRE2COREZMRAB17	ACGTCBOX	CANBNNAPA	CAATBOX1	PROLAMINBOXOSGLUB1	GCN4OSGLUB1	MYBPZM	ABRELATERD1	DPBFCOREDCDC3	EBOXBNNAPA	RYREPEATBNNAPA
Silybum marianum	162290	PROSmCAD	0	0	1	0	6	0	0	0	0	9	0	0	2	0	0	4	0
	93708	PROSmPAL	1	0	3	1	7	0	0	0	0	12	0	0	0	0	0	8	0
	140873	PROSmCHS	1	0	1	0	5	0	0	0	0	19	0	0	0	1	0	10	0
	115271	PROSmF3H	2	0	1	1	5	0	0	0	1	1	0	0	2	6	1	12	1
	57461	PROSmF3'H	0	1	0	1	2	0	0	0	3	14	0	0	7	0	1	0	0
	177494	PROSmAPX1	0	0	0	10	5	0	0	0	1	7	1	0	6	0	0	2	0

Figure 7. Linkage between compounds and ABA in whole achene stages. (**a**) Kinetics of abscisic acid (ABA) net production (expressed in ng/achene) at each developmental stage in whole achene (WA) during milk thistle fruit maturation (calculated from Table 1). (**b**) Relative quantification using RT-qPCR of RNA expression of 2 genes implicated in ABA flavonolignans biosynthesis (*ABA1* and *LEC2*). (**c**) Pearson correlation of extraction compounds in whole achene stages were performed by PAST software. (* $p < 0.05$, $n = 3$). (**d**) Identification of ABA-responsive cis-acting elements located in gene promoter gene sequences (from blue to yellow indicating low to high number of each cis-acting element). Values are the mean ± SD of 3 independent measurements. Characteristics of cis-acting elements are provided in Table S5. Different letters indicate significant differences at $p < 0.05$.

Maximum production of ABA coincided with the preceding stage of the embryogenesis, but also observed increases in biosynthetic gene expression and enzyme activity leading to SILM flavonolignan production (Figures 2–4, Figure 6). RT-qPCR analysis of the expression of genes involved in ABA biosynthesis (*ABA1*) and signaling (*LEC2*) confirmed this accumulation time course (Figure 7b). We observed significant associations between ABA accumulation and SILM biosynthesis (Figure 7c). ABA was detected both in the pericarp and in the seed (Table S1). According to these results, it appeared that the biosynthesis of flavonolignans could be regulated by ABA, with the seed appearance as a signal for the start of their biosynthesis. Such a regulation has already been proposed in the regulation of lignan biosynthesis in flax, in which seed development was needed for biosynthesis, and ABA acts as a key regulator [28,41]. Similarly, as observed in flax, in aborted *S. marianum* achenes, no active flavonolignan biosynthesis has been observed in the absence of seed development (data not shown). The potential contribution of ABA in the transcriptional regulation of SILM biosynthesis was further verified by in silico identification of putative ABA-responsive and fruit/seed-specific cis-acting elements (Figure 7d; Table S5). ABA is a central phytohormone involved in seed and fruit maturing regulation in many species of model and crop plants, such as *A. thaliana* and flax [28–30,44]. It has also been related to transcriptional regulation of biosynthesis of phenylpropanoids in seeds and fruits [28–30]. Here, information on the timing and location of ABA accumulation during *S. marianum* fruit maturation and its association with the expression of biosynthetic genes and SILM accumulation is of particular interest to understand how this metabolic pathway is regulated. Future work should be carried out to identify the transcription factors involved in this regulation and to characterize them functionally.

3. Materials and Methods

3.1. Plant Materials

The plants were grown in pots (30 cm in diameter and 30 cm in depth), packed with commercial garden soil (composition: 250 g/m^3 N, 120 g/m^3 P_2O_5, 80 g/m^3 K_2O, dry matter: 37%, organic matter: 65%, pH: 6.2, conductivity: 49 mS/cm, water retention capacity: 70% volume) in a phytotronic room at 25 °C under a 16-h photoperiod (30 µmol/m^2/s total amount of photosynthetically active radiation) and relative humidity (RH) was around 30%. Plants were irrigated once a day using overhead mist irrigation, and one full watering per week until the full development cycle.

3.2. Chemicals

Solvents and reagents used in the present study were all of analytical grade or highest available purity (Fisher Scientific, Illkirch, France). Deionized ultrapure water was produced using a Milli-Q water-purification system (Millipore, Molsheim, France). All analytical solutions were filtered through 0.45 µm nylon syringe membranes prior to use. Commercial standards of TAX, SILC, SILD, SILA, SILB, ISILA and ISILB were purchased from Sigma-Aldrich (Saint-Quentin Fallavier, France).

3.3. Phytochemicals Analysis

Ultrasonic extractions (3 biological and 2 technical replicates) were performed using 60 mg (DW) of achene, pericarp or seed in 1 mL of 50% (*v*/*v*) aqueous ethanol as described by Drouet et al. [45]. For this purpose, an USC1200TH ultrasonic bath with the following inner dimension was used: 300 mm × 240 mm × 200 mm (VWR International, Fontenay-sous-Bois, France). Silymarin composition and quantity were determined by LC-MS using a Water 2695 Alliance (Waters-Micromass, Manchester, UK) coupled with a single quadrupole mass spectrometer ZQ (Waters-Micromass, Manchester, UK) as described previously [22,32].

3.4. Enzymatic Activities

3.4.1. Total Soluble Proteins Extraction and Quantification

From 150 mg of fresh frozen (−80 °C) tissue, total soluble proteins were extracted by homogenization in 3 mL 0.1 M sodium borate buffer (SBB) pH 8.8 containing 10 mM β-mercaptoethanol as described by Hano et al. [46]. Protein concentration was measured with the Quant-iT Protein Assay Kit and Qubit® 3.0 fluorometer according to manufacturer instructions (Thermo Scientific, Courtaboeuf, France).

3.4.2. PAL Activity

PAL activity was spectrometrically determined, monitoring the formation of trans-cinnamate at 290 nm as described by Hano et al. [46].

3.4.3. CHS Activity

CHS activity was determined by HPLC as described by Sun et al. [47], using *p*-coumaroyl-CoA, synthesized according to Beuerle and Pichersky [48], and malonyl-CoA (Sigma-Aldrich, Saint-Quentin Fallavier, France) as substrate, by monitoring at 289 nm the formation of naringenin from the subsequent non-enzymatic conversion of the formed naringenin by CHS activity.

3.4.4. POX Activity

POX (peroxidase) activity was determined spectrometrically using guaiacol (Sigma-Aldrich, Saint-Quentin Fallavier, France) as substrate, and following the absorbance increase at 470 nm as described by Morawski et al. [49].

3.4.5. LAC Activity

LAC (laccase) activity was determined spectrometrically at 415 nm, following the ABTS (2,2′-azino-bis(3-ethylbenzothiazoline-6-sulfonate) (Sigma-Aldrich, Saint-Quentin Fallavier, France) oxidation as described by Wang et al. [50].

3.5. Gene Identification

Gene identification by tBLASTn analysis on NCBI server using publicly available sequence contigs, generated from Illumina Hiseq data of *S. marianum* (NCBI:txid92921, WGS:LMWD01000001:LMWD01258575) using *A. thaliana* orthologs as queries with the comparison matrix BLOSUM62 (at the score value of > 300 and *e*-value < e−100). The results of these searches are presented in Table S3 (reference genes) and Table S4 (biosynthetic genes).

3.6. Gene Promoter Analysis

The corresponding putative promoter sequences were determined as the 1500 base pairs upstream of the predicted starting translation codons. Putative promoter sequences were submitted to PLACE [51] and PlantPAN2.0 [52] analyses to identify putative cis-acting regulatory DNA elements involved in seed expression and/or response to ABA.

3.7. RNA Extraction

The total RNAs of achene, pericarp and seed were extracted from crushed tissue in liquid nitrogen using the GeneJET Plant RNA Purification kit (Thermo Fisher Scientific, Courtaboeuf, France) following manufacturer's recommendations. An additional DNase I treatment (RNase-free DNase, Qiagen, Courtabeauf, France) was applied directly to the column for 15 minutes at 25 °C to remove traces of contaminating DNA. Total RNAs were then quantified using a fluorometer and the QuantiT RNA Assay

Kit (Life Technologies, Courtaboeuf, France) and Qubit fluorometer (Life Technologies, Courtaboeuf, France) according to the manufacturer's instructions. RNA was then stored at −80 °C.

3.8. RT-qPCR Analysis

The first strand of cDNA was retro-synthesized from 50 ng of total RNA using the Maxima Reverse Transcriptase kit (Life Technologies, Courtaboeuf, France) according to manufacturer's instructions and were stored at −25 °C. Quantitative PCRs were realized in 96-well plates using the PikoReal real time PCR system (ThermoFisher, Courtaboeuf, France) and DyNAmoColorFlash SYBR Green qPCR Kit (ThermoFisher, Courtaboeuf, France). Each reaction was performed as described in Corbin et al. [24]. Analysis of the data was performed with Pikoreal software. Three biological replicates and two technical repetitions were realized for each sample. Relative transcript levels were obtained using specific primers (Tables S3 and S4), designed with Primer3 software [53], and normalized using the comparative ΔΔCq method using two validated housekeeping reference genes.

3.9. Validation of Reference Genes

The evaluation of twelve candidate reference genes was performed with RefFinder, a web-based comprehensive tool developed for the evaluation, screening and selection of reference genes from extensive experimental datasets. RefFinder integrates the major available computational programs geNorm [54], Normfinder [38], BestKeeper [55] to compare and rank the tested candidate reference genes. Based on the rankings from each program, it assigns an appropriate weight to an individual gene and calculates the geometric mean for the overall final ranking [56].

3.10. ABA Extraction and Quantification

ABA extraction from developing milk thistle fruit was based on the procedure described by Renouard et al. [28]. Freeze-dried developing achenes (100 mg FW) were extracted for 16 h at 4 °C in the dark with MilliQ water (water/tissue ratio 50:1, *v*/*w*). ABA was quantified by ELISA assay Phytodetek ABA ELISA kit (Agdia EMA, Evry, France) using (±) cis–trans ABA (Sigma, Saint-Quentin Fallavier, France) as a standard. Experiments were realized in triplicates.

3.11. Statistical and Treatment of Data

At least three independent biological repetitions were performed to allow calculation of means and standard deviation. Boxplots were conducted using RStudio. The correlation matrix was obtained with PAST software by performing the Pearson parametric correlation test. Heat maps were produced using the MeV software computed with a hierarchical clustering analysis (HCA) representing the Euclidean distance with a clustering method with a complete linkage clustering. The metabolite network was visualized using the Cytoscape 2.8.3 software by representing only the significant Pearson Correlation Coefficient (PCC) values at $p < 0.05$ with a cut-off value of 0.60 (significant positive (in red) and negative (in blue) correlations). Colors from yellow to red indicate increasing PCC values and the connection size indicates the strength of the connection.

4. Conclusions

Silymarin (SILM) is a complex mixture of bioactive flavonolignans that accumulate milk thistle (*Silybum marianum* (L.) Gaertn., *Asteraceae*) in its mature achene fruits. These compounds are well known for their relationship to promote human health and prevent disease, but the conditions of their biosynthesis in planta remain elusive. Development stages of fruit were precisely described to study the kinetics of accumulation of SILM constituents during fruit ripening. During fruit maturation, the accumulation profiles of the SILM components were evaluated by LC-MS analysis at each of the development stages identified. Reference genes have been identified, selected and validated to allow accurate gene expression profiling of candidate biosynthetic genes. Enzyme activity and biosynthetic gene expression indicated a possible in situ biosynthesis of SILM from L-Phe during fruit ripening. The gene expression profiles were well correlated with SILM kinetic accumulation and preferential location in pericarp during *S. marianum* fruit maturation, reaching maximum biosynthesis when desiccation occurs. This observation led us to consider the possible involvement of abscisic acid (ABA), a key phytohormone in fruit ripening control, for which accumulation timing and location during fruit ripening were consistent with the potential regulation of the SLIM accumulation. This possible transcriptional regulation of SILM biosynthesis by ABA was further supported by the presence of ABA-responsive cis-acting elements in the SILM biosynthetic gene promoter regions studied. These results pave the way for a better understanding of the biosynthetic regulation of SILM during the maturation of *S. marianum* fruit, thereby providing important insights to better control the production of these medicinally important compounds.

Supplementary Materials: The following are available online at http://www.mdpi.com/1422-0067/21/13/4730/s1, Figure S1. a. Variation of Ct values for each of the 12 analyzed potential reference genes in whole achenes, pericarps and seeds during the 6 developmental stages of *S. marianum* maturation. b. Agarose gel electrophoresis analysis of amplified RT-qPCR fragments for each 12 analyzed potential reference genes (here analyzed in whole achenes of *S. marianum* at developmental stage 4); Figure S2. Alignment of APX_1Lv (from Lv., 2017) and smAPX (predicted in Augustus software) in Clustal omega; Table S1: Evolution of the accumulation of SILM and its different constituents during *S. marianum* fruit maturation (in whole achenes (WA), pericarps (P) and seed (S) from stage 1 to stage 6 of maturation); Table S2: Pearson correlation coefficient between PAL, CHS, POX and LAC activities determined during *S. marianum*; Table S2: Primers and characteristics of the genes used for gene references selection for RT-qPCR analysis in maturing fruit of *Silybum marianum*; Table S4: Primers and characteristics of the SILM biosynthetic genes and ABA biosynthetic and signaling genes used for gene expression by RT-qPCR analysis in maturing fruit of *Silybum marianum*; Table S5: List of cis-acting elements located in the SILM biosynthetic gene promoter regions.

Author Contributions: Conceptualization, C.H.; Data curation, S.D. and C.H.; Formal analysis, S.D., É.L. and C.H.; Funding acquisition, C.H.; Investigation, S.D., D.T. and C.H.; Methodology, S.D. and D.T.; Project administration, C.H.; Resources, C.H.; Software, S.D.; Supervision, É.L. and C.H.; Validation, S.D., D.T., É.L. and C.H.; Visualization, S.D. and C.H.; Writing—original draft, S.D. and C.H.; Writing—review and editing, S.D., D.T., É.L. and C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was part of the MariLine project (grant number ICO16029) focusing on the cosmetic application of the milk thistle flavonolignans, supported by Cosmetosciences, a global training and research program dedicated to the cosmetic industry. Located in the heart of the cosmetic valley, this program led by University of Orléans is funded by the Region Centre-Val de Loire.

Acknowledgments: S.D. acknowledges research fellowships of Loire Valley Region. D.T. gratefully acknowledges the support of French government via the French Embassy in Thailand in the form of Junior Research Fellowship Program 2018.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Abenavoli, L.; Capasso, R.; Milic, N.; Capasso, F. Milk Thistle in Liver Diseases: Past, Present, Future. *Phyther. Res.* **2010**, *24*, 1423–1432. [CrossRef] [PubMed]
2. Flora, K.; Martin, H.; Rosen, H.; Benner, K. Clinical reviews Milk Thistle (*Silybum marianum*) for the Theraply of Liver Disease. *Am. J. Gastroenterol.* **1998**, *93*, 139–143. [CrossRef] [PubMed]
3. Federico, A.; Dallio, M.; Loguercio, C. Silymarin/Silybin and chronic liver disease: A marriage of many years. *Molecules* **2017**, *22*, 191. [CrossRef] [PubMed]
4. Shaker, E.; Mahmoud, H.; Mnaa, S. Silymarin, the antioxidant component and *Silybum marianum* extracts prevent liver damage. *Food Chem. Toxicol.* **2010**, *48*, 803–806. [CrossRef] [PubMed]
5. Singh, R.P.; Agarwal, R. Cosmeceuticals and silibinin. *Clin. Dermatol.* **2009**, *27*, 479–484. [CrossRef] [PubMed]
6. Shah, M.; Ullah, M.A.; Drouet, S.; Younas, M.; Tungmunnithum, D.; Giglioli-Guivarc'h, N.; Hano, C.; Abbasi, B.H. Interactive effects of light and melatonin on biosynthesis of silymarin and anti-inflammatory potential in callus cultures of *Silybum marianum* (L.) gaertn. *Molecules* **2019**, *24*, 1207. [CrossRef] [PubMed]

7. Toklu, H.Z.; Tunali-Akbay, T.; Erkanli, G.; Yüksel, M.; Ercan, F.; Şener, G. Silymarin, the antioxidant component of *Silybum marianum*, protects against burn-induced oxidative skin injury. *Burns* **2007**, *33*, 908–916. [CrossRef] [PubMed]
8. Trouillas, P.; Marsal, P.; Svobodova, A.; Vosta, J.; Hrba, J.; Lazzaroni, R.; Duroux, J.; Walterova, D. Mechanism of the Antioxidant Action of Silybin and 2,3-Dehydrosilybin Flavonolignans: A Joint Experimental and Theoretical Study. *J. Phys. Chem. A* **2008**, *112*, 1054–1063. [CrossRef]
9. Katiyar, S.K.; Meleth, S.; Sharma, S.D. Silymarin, a flavonoid from milk thistle (*Silybum marianum* L.), inhibits UV-induced oxidative stress through targeting infiltrating CD11b+cells in mouse skin. *Photochem. Photobiol.* **2008**, *84*, 266–271. [CrossRef]
10. Younas, M.; Drouet, S.; Nadeem, M.; Giglioli-Guivarc'h, N.; Hano, C.; Abbasi, B.H. Differential accumulation of silymarin induced by exposure of *Silybum marianum* L. callus cultures to several spectres of monochromatic lights. *J. Photochem. Photobiol. B Biol.* **2018**, *184*, 61–70. [CrossRef]
11. Pelter, A.; Hansel, R. The structure of silybin (*Silybum* substance E6), the first flavonolignan. *Tetrahedron Lett.* **1968**, *1*, 2911–2916. [CrossRef]
12. Nam-Cheol, K.; Graf, T.N.; Sparacino, C.M.; Wani, M.C.; Wall, M.E. Complete isolation and structure identification of hepatoprotective flavonolignans silybins and isosilybins from the medicinal herb milk thistle (*Silybum marianum*). *Org. Biomol. Chem.* **2003**, *1*, 1684–1689.
13. Kurkin, V.A.; Zapesochnaya, G.G.; Volotsueva, A.V.; Avdeeva, E.V.; Pimenov, K.S. Flavolignans of *Silybum marianum* fruit. *Chem. Nat. Compd.* **2001**, *37*, 315–317. [CrossRef]
14. Jeong, R.; Lee, D.; Cho, J.; Lee, S.; Kang, H.; Seo, W.; Kang, H.; Kim, J.; Baek, N. Article A New Flavonolignan from the Aerial Parts of *Oryza sativa* L. Inhibits Nitric oxide Production in RAW 264.7 Macrophage Cells. *J. Korean Soc. Appl. Biol. Chem.* **2011**, *54*, 865–870. [CrossRef]
15. Cardona, M.L.; Garcia, B.; Pedro, R.; Sinisterra, J.F. Flavonoids, Flavonolignans and a phenylpropanoid from *Onopordon Corymbosum*. *Phytochemistry* **1990**, *29*, 629–631. [CrossRef]
16. Nakajima, Y.; Yun, Y.S.; Kunugi, A. Six new flavonolignans from *Sasa veitchii* (Carr.) Rehder. *Tetrahedron* **2003**, *59*, 8011–8015. [CrossRef]
17. Bai, N.; He, K.; Roller, M.; Lai, C.S.; Bai, L.; Pan, M.H. Flavonolignans and other constituents from *Lepidium meyenii* with activities in anti-inflammation and human cancer cell lines. *J. Agric. Food Chem.* **2015**, *63*, 2458–2463. [CrossRef]
18. Wang, J.F.; Yin, G.F.; Zhou, X.J.; Su, J.; Li, Y.; Zhong, H.M.; Duan, G.; Cheng, Y.X. Anti-inflammatory flavonolignans from *Hydnocarpus anthelminthica* seeds. *J. Asian Nat. Prod. Res.* **2011**, *13*, 80–83. [CrossRef]
19. Karkanis, A.; Bilalis, D.; Efthimiadou, A. Cultivation of milk thistle (*Silybum marianum* L. Gaertn.), a medicinal weed. *Ind. Crops Prod.* **2011**, *34*, 825–830. [CrossRef]
20. Lv, Y.; Gao, S.; Xu, S.; Du, G.; Zhou, J.; Chen, J.; Road, L.; Chen, J. Spatial organization of silybin biosynthesis in milk thistle (*Silybum marianum* (L.) Gaertn.). *Plant J.* **2017**, *92*, 995–1004. [CrossRef]
21. Poppe, L.; Petersen, M. Variation in the flavonolignan composition of fruits from different *Silybum marianum* chemotypes and suspension cultures derived therefrom. *Phytochemistry* **2016**, *131*, 68–75. [CrossRef] [PubMed]
22. Drouet, S.; Abbasi, B.H.; Falguières, A.; Ahmad, W.; Ferroud, C.; Doussot, J.; Vanier, J.R.; Lainé, E.; Hano, C. Single Laboratory Validation of a Quantitative Core Shell-Based LC Separation for the Evaluation of Silymarin Variability and Associated Antioxidant Activity of Pakistani Ecotypes of Milk Thistle (*Silybum Marianum*, L.). *Molecules* **2018**, *23*, 904. [CrossRef]
23. Davin, L.B.; Wang, H.-B.; Crowell, A.L.; Bedgar, D.L.; Martin, D.M.; Sarkanen, S.; Lewis, N.G. Stereoselective bimolecular phenoxy radical coupling by an auxiliary (dirigent) protein without an active center. *Science* **1997**, *275*, 362–367. [CrossRef]
24. Corbin, C.; Drouet, S.; Markulin, L.; Auguin, D.; Lainé, É.; Davin, L.B.; Cort, J.R.; Lewis, N.G.; Hano, C. A genome-wide analysis of the flax (*Linum usitatissimum* L.) dirigent protein family: From gene identification and evolution to differential regulation. *Plant Mol. Biol.* **2018**, *97*, 73–101. [CrossRef] [PubMed]
25. Torres, M.; Corchete, P. Gene expression and flavonolignan production in fruits and cell cultures of *Silybum marianum*. *J. Plant. Physiol.* **2016**, *192*, 111–117. [CrossRef]
26. Martinelli, T.; Andrzejewska, J.; Salis, M.; Sulas, L. Phenological growth stages of *Silybum marianum* according to the extended BBCH scale. *Ann. Appl. Biol.* **2014**, *166*, 53–66. [CrossRef]

27. Carrier, D.J.; Crowe, T.; Sokhansanj, S.; Wahab, J.; Branka, B. Milk Thistle, *Silybum marianum* (L.) Gaertn., flower head development and associated marker compound profile. *J. Herbs Spices Med. Plants* **2003**, *10*, 65–74. [CrossRef]
28. Renouard, S.; Corbin, C.; Lopez, T.; Montguillon, J.; Gutierrez, L.; Lamblin, F.; Lainé, E.; Hano, C. Abscisic acid regulates pinoresinol–lariciresinol reductase gene expression and secoisolariciresinol accumulation in developing flax (*Linum usitatissimum* L.) seeds. *Planta* **2012**, *235*, 85–98. [CrossRef]
29. Gutierrez, L.; Van Wuytswinkel, O.; Castelain, M.; Bellini, C. Combined networks regulating seed maturation. *Trends Plant Sci.* **2007**, *12*, 294–300. [CrossRef]
30. Finkelstein, R.R.; Gampala, S.S.L.; Rock, C.D. Abscisic Acid Signaling in Seeds and Seedlings. *Plant Cell* **2002**, S15–S45. [CrossRef]
31. Ciocarlan, A.; Dragalin, I.; Aricu, A.; Ciocarlan, N. Chromatographic analysis of *Silybum marianum* (L.) gaernt. fatty oil. *Chem. J. Mold.* **2018**, *13*, 63–68.
32. Martinelli, T.; Whittaker, A.; Benedettelli, S.; Carboni, A. The study of flavonolignan association patterns in fruits of diverging *Silybum marianum* (L.) Gaertn. chemotypes provides new insights into the silymarin biosynthetic pathway. *Phytochemistry* **2017**, *144*, 9–18. [CrossRef] [PubMed]
33. Nguyen, T.; Jamali, A.; Grand, E.; Morreel, K.; Marcelo, P.; Gontier, E.; Dauwe, R. Phenylpropanoid profiling reveals a class of hydroxycinnamoyl glucaric acid conjugates in Isatis tinctoria leaves. *Phytochemistry* **2017**, *144*, 127–140. [CrossRef] [PubMed]
34. Buer, C.S.; Muday, G.K.; Djordjevic, M.A. Flavonoids are differentially taken up and transported long distances in Arabidopsis. *Plant Physiol.* **2007**, *145*, 478–490. [CrossRef]
35. Petrussa, E.; Braidot, E.; Zancani, M.; Peresson, C.; Bertolini, A.; Patui, S.; Vianello, A. Plant flavonoids—Biosynthesis, transport and involvement in stress responses. *Int. J. Mol. Sci.* **2013**, *14*, 14950–14973. [CrossRef]
36. Løvdal, T.; Lillo, C. Reference gene selection for quantitative real-time PCR normalization in tomato subjected to nitrogen, cold, and light stress. *Anal. Biochem.* **2009**, *387*, 238–242. [CrossRef] [PubMed]
37. Silver, N.; Best, S.; Jiang, J.; Thein, S.L. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Mol. Biol.* **2006**, *7*, 33. [CrossRef] [PubMed]
38. Andersen, C.L.; Jensen, J.L.; Ørntoft, T.F. Normalization of real-time quantitative reverse transcription-PCR data: A model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res.* **2004**, *64*, 5245–5250. [CrossRef] [PubMed]
39. Gutierrez, L.; Conejero, G.; Castelain, M.; Guénin, S.; Verdeil, J.-L.; Thomasset, B.; Van Wuytswinkel, O. Identification of new gene expression regulators specifically expressed during plant seed maturation. *J. Exp. Bot.* **2006**, *57*, 1919–1932. [CrossRef] [PubMed]
40. Ma, Q.H. Functional analysis of a cinnamyl alcohol dehydrogenase involved in lignin biosynthesis in wheat. *J. Exp. Bot.* **2010**, *61*, 2735–2744. [CrossRef]
41. Hano, C.; Martin, I.; Fliniaux, O.; Legrand, B.; Gutierrez, L.; Arroo, R.R.J.; Mesnard, F.; Lamblin, F.; Lainé, E. Pinoresinol-lariciresinol reductase gene expression and secoisolariciresinol diglucoside accumulation in developing flax (*Linum usitatissimum*) seeds. *Planta* **2006**, *224*, 1291–1301. [CrossRef]
42. Renouard, S.; Corbin, C.; Lopez, T.; Lamblin, F.; Lainé, E.; Hano, C. Isolation of nuclear proteins from flax (*Linum usitatissimum* L.) seed coats for gene expression regulation studies. *BMC Res. Notes* **2012**, *5*. [CrossRef]
43. Davin, L.B.; Lewis, N.G. Dirigent Proteins and Dirigent Sites Explain the Mystery of Specificity of Radical Precursor Coupling in Lignan and Lignin Biosynthesis. *Plant Physiol.* **2000**, *123*, 453–461. [CrossRef]
44. Corbin, C.; Renouard, S.; Lopez, T.; Lamblin, F.; Lainé, E.; Hano, C. Identification and characterization of cis-acting elements involved in the regulation of ABA-and/or GA-mediated LuPLR1 gene expression and lignan biosynthesis in flax (*Linum usitatissimum* L.) cell cultures. *J. Plant Physiol.* **2013**, *170*, 516–522. [CrossRef] [PubMed]
45. Drouet, S.; Leclerc, E.A.; Garros, L.; Tungmunnithum, D.; Kabra, A.; Abbasi, B.H.; Lain, É.; Hano, C. A Green Ultrasound-Assisted Extraction Optimization of the Natural Antioxidant and Anti-Aging Flavonolignans from Milk Thistle *Silybum marianum* (L.) Gaertn. Fruits for Cosmetic Applications. *Antioxidants* **2019**, *8*, 304. [CrossRef] [PubMed]
46. Hano, C.; Addi, M.; Bensaddek, L.; Crônier, D.; Baltora-Rosset, S.; Doussot, J.; Maury, S.; Mesnard, F.; Chabbert, B.; Hawkins, S.; et al. Differential accumulation of monolignol-derived compounds in elicited flax (*Linum usitatissimum*) cell suspension cultures. *Planta* **2006**, *223*, 975–989. [CrossRef]

47. Sun, W.; Meng, X.; Liang, L.; Jiang, W.; Huang, Y.; He, J.; Hu, H.; Almqvist, J.; Gao, X.; Wang, L. Molecular and biochemical analysis of chalcone synthase from Freesia hybrid in flavonoid biosynthetic pathway. *PLoS ONE* **2015**, *10*, e0119054. [CrossRef] [PubMed]
48. Beuerle, T.; Pichersky, E. Enzymatic synthesis and purification of aromatic coenzyme A esters. *Anal. Biochem.* **2002**, *302*, 305–312. [CrossRef]
49. Morawski, B.; Lin, Z.; Cirino, P.; Joo, H.; Bandara, G.; Arnold, F.H. Functional expression of horseradish peroxidase in *Saccharomyces cerevisiae* and *Pichia pastoris*. *Protein Eng.* **2000**, *13*, 377–384. [CrossRef]
50. Wang, F.; Hu, J.-H.; Guo, C.; Liu, C.-Z. Enhanced laccase production by *Trametes versicolor* using corn steep liquor as both nitrogen source and inducer. *Bioresour. Technol.* **2014**, *166*, 602–605. [CrossRef]
51. Higo, K.; Ugawa, Y.; Iwamoto, M.; Korenaga, T. Plant *cis*-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* **1999**, *27*, 297–300. [CrossRef] [PubMed]
52. Chow, C.-N.; Zheng, H.-Q.; Wu, N.-Y.; Chien, C.-H.; Huang, H.-D.; Lee, T.-Y.; Chiang-Hsieh, Y.-F.; Hou, P.-F.; Yang, T.-Y.; Chang, W.-C. PlantPAN 2.0: An update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. *Nucleic Acids Res.* **2016**, *44*, D1154–D1160. [CrossRef] [PubMed]
53. Untergasser, A.; Cutcutache, I.; Koressaar, T.; Ye, J.; Faircloth, B.C.; Remm, M.; Rozen, S.G. Primer3-new capabilities and interfaces. *Nucleic Acids Res.* **2012**, *40*, 1–12. [CrossRef] [PubMed]
54. Vandesompele, J.; De Preter, K.; Pattyn, F.; Poppe, B.; Van Roy, N.; De Paepe, A.; Speleman, F. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **2002**, *3*. [CrossRef] [PubMed]
55. Pfaffl, M.W.; Tichopad, A.; Prgomet, C.; Neuvians, T.P. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper—Excel-based tool using pair-wise correlations. *Biotechnol. Lett.* **2004**, *26*, 509–515. [CrossRef] [PubMed]
56. Xie, F.; Xiao, P.; Chen, D.; Xu, L.; Zhang, B. miRDeepFinder: A miRNA analysis tool for deep sequencing of plant small RNAs. *Plant Mol. Biol.* **2012**, *80*, 75–84. [CrossRef]

International Journal of
Molecular Sciences

MDPI

Article

Red Chinese Cabbage Transcriptome Analysis Reveals Structural Genes and Multiple Transcription Factors Regulating Reddish Purple Color

Jana Jeevan Rameneni [1,†], Su Ryun Choi [1,†], Sushil Satish Chhapekar [1,†], Man-Sun Kim [1], Sonam Singh [1], So Young Yi [1], Sang Heon Oh [1], Hyuna Kim [1], Chang Yeol Lee [1], Man-Ho Oh [2], Jhongchul Lee [3], Oh Ha Kwon [3], Sang Un Park [4], Sun-Ju Kim [5] and Yong Pyo Lim [1,*]

[1] Molecular Genetics and Genomics Laboratory, Department of Horticulture, College of Agriculture and Life Science, Chungnam National University, Daejeon 34134, Korea

[2] Department of Biological Sciences, College of Biological Sciences and Biotechnology, Chungnam National University, Daejeon 34134, Korea

[3] Kwonnong Seed Co., 186 Pungnyeon-ro, Heungdeok-gu, Cheongju 28394, Korea

[4] Department of Crop Science, College of Agriculture and Life Science, Chungnam National University, Daejeon 34134, Korea

[5] Department of Bio-Environmental Chemistry, College of Agriculture and Life Science, Chungnam National University, Daejeon 34134, Korea

* Correspondence: yplim@cnu.ac.kr; Tel.: +82-42-821-5739; Fax: +82-42-821-8847

† The authors contributed equally and shares first authorship.

Received: 1 April 2020; Accepted: 18 April 2020; Published: 21 April 2020

Abstract: Reddish purple Chinese cabbage (RPCC) is a popular variety of *Brassica rapa* (AA = 20). It is rich in anthocyanins, which have many health benefits. We detected novel anthocyanins including cyanidin 3-(feruloyl) diglucoside-5-(malonoyl) glucoside and pelargonidin 3-(caffeoyl) diglucoside-5-(malonoyl) glucoside in RPCC. Analyses of transcriptome data revealed 32,395 genes including 3345 differentially expressed genes (DEGs) between 3-week-old RPCC and green Chinese cabbage (GCC). The DEGs included 218 transcription factor (TF) genes and some functionally uncharacterized genes. Sixty DEGs identified from the transcriptome data were analyzed in 3-, 6- and 9-week old seedlings by RT-qPCR, and 35 of them had higher transcript levels in RPCC than in GCC. We detected *cis*-regulatory motifs of MYB, bHLH, WRKY, bZIP and AP2/ERF TFs in anthocyanin biosynthetic gene promoters. A network analysis revealed that MYB75, MYB90, and MYBL2 strongly interact with anthocyanin biosynthetic genes. Our results show that the late biosynthesis genes *BrDFR, BrLDOX, BrUF3GT, BrUGT75c1-1, Br5MAT, BrAT-1, BrAT-2, BrTT19-1,* and *BrTT19-2* and the regulatory MYB genes *BrMYB90, BrMYB75,* and *BrMYBL2-1* are highly expressed in RPCC, indicative of their important roles in anthocyanin biosynthesis, modification, and accumulation. Finally, we propose a model anthocyanin biosynthesis pathway that includes the unique anthocyanin pigments and genes specific to RPCC.

Keywords: anthocyanins; anthocyanin biosynthetic genes; *cis*-regulatory motifs; DEGs; network analysis; qRT-PCR; reddish purple Chinese cabbage; transcriptome; transcription factors

1. Introduction

Introgression breeding is an important traditional breeding technique for transferring key agronomic traits between two distinct species [1]. Using this technique, improvements have been made to many *Brassica* traits, such as disease resistance, male sterility, seed color, oil quality traits, and other morphological traits [1,2]. Some purple Brassicaceae lines have also been generated [3,4]. Red Chinese cabbage (*Brassica rapa* ssp. *pekinensis* L.) is reddish purple in color and rich in anthocyanins.

This vibrantly colored variety is a popular addition to salads and it has important antioxidant properties [4].

Anthocyanins are a class of secondary metabolites that are synthesized through the phenylpropanoid pathway [5]. These water-soluble compounds with red, purple, or blue colors are synthesized in the cytosol and stored in the vacuole [6]. Anthocyanins play crucial roles in reducing damage from, and in defense responses against, abiotic stresses such as ultraviolet exposure, wounding, high light, chilling, pollution, osmotic stress, and nutrient deficiency, as well as biotic stresses such as pathogen infection [7].

The important structural genes in the anthocyanin biosynthetic pathway are those encoding phenylalanine ammonia-lyase (PAL), cinnamate 4-hydroxylase (C4H), 4-coumaroyl CoA ligase (4CL), chalcone synthase (CHS), chalcone isomerase (CHI), flavanone 3-hydroxylase (F3H), flavonoid 3′-hydroxylase (F3′H), dihydroflavonol 4-reductase (DFR), leucoanthocyanidin dioxygenase (LDOX), UDP-flavonoid glucosyl transferase (UFGT) and glutathione S-transferase (GST) [8]. The late anthocyanin biosynthesis genes are regulated by a complex of transcription factors (TFs), MYB-bHLH-WD40, known as the MBW complex. In vivo assays in *Antirrhinum* revealed that a flower-specific MYB protein activates the transcription of genes involved in phenylpropanoid biosynthesis [9,10]. Four MYB TFs, encoded by *PAP1/MYB75*, *PAP2/MYB90*, *MYB113*, and *MYB114*, have been identified to control anthocyanin biosynthesis in vegetative tissues of *Arabidopsis* [11,12]. Previous biochemical and genetic studies have shown that TTG1 (WD40), GL3/EGL3/TT8 (bHLH) and PAP1/PAP2/MYB113/MYB114 (MYB) are components of potential WBM complexes that activate anthocyanin biosynthesis [13,14]. Interestingly, recent studies on proanthocyanidin and anthocyanin biosynthesis pathways in *Arabidopsis* and *Petunia* suggest that WRKY TFs regulate color accumulation along with the MBW complex [10,15]. Similarly, studies on bZIP-TFs revealed the mechanisms by which they regulate anthocyanin accumulation in apples [16,17].

Secondary metabolites are abundant in Chinese cabbage. Previous studies have shown that some secondary metabolites are specific to particular species, cultivars, or varieties. Such metabolites are usually detected by high performance liquid chromatography (HPLC) coupled with liquid chromatography-tandem mass spectrometry (LC-MS/MS) [4,18–20]. Recent metabolic profiling studies have demonstrated that cyanidin derivatives are highly accumulated in *Brassica* species [18,21–23]. It is difficult to identify all the genes related to specific traits in a plant system, but transcriptome sequencing allows for the prediction of functional genes in the plant genome. Several RNA-seq studies have been conducted for *Brassica* species, and their results have led to the identification of sets of genes that are expressed under various conditions and/or in certain genotypes. This information is very useful for the functional characterization of target genes [24–27].

Reddish purple Chinese cabbage (RPCC) is a variety of *B. rapa*. It is an economically important leafy vegetable that is widely cultivated and consumed in East Asian countries, especially Korea, due to its health promoting properties [28]. To date, the underlying molecular mechanisms and the genes regulating the anthocyanin pigments responsible for its vivid red color are unexplored, and may differ from those reported previously for other *Brassica* lines. In this study, we used a next generation sequencing (NGS) based RNA-sequencing approach to identify a novel set of genes involved in anthocyanin biosynthesis, and the regulation of this pathway, in a red Chinese cabbage variety. A comprehensive analysis of this plant including computational, leaf chemotype, and expressional abundance analyses shows the significance of this variety.

2. Results

2.1. Estimation of Anthocyanin Content in RPCC and GCC Leaf Samples

The red color is a distinguishing feature of some *Brassica* species, and is due to the accumulation of anthocyanins [29]. To determine the types of anthocyanins that confer color in RPCC, we analyzed anthocyanins in the innermost and outermost leaves of 9-week-old GCC and RPCC plants (Table 1 and Table S1) by HPLC-MS/MS. We detected 13 anthocyanin pigments, 12 of which were derivatives of cyanidin. The other one was a pelargonidin derivative. The most abundant pigment in the innermost leaves of RPCC was cyanidin 3-(feruloyl) diglucoside-5-(malonoyl) glucoside (peak 8) with a concentration of 12.63 ± 1.37 mg/g dry weight, followed by pelargonidin 3-(caffeoyl) diglucoside-5-(malonoyl) glucoside (peak 7) with a concentration of 5.66 ± 0.60 mg/g dry weight (Figure 1; Table 1 and Table S1). In the outermost leaves of RPCC, the most abundant pigments were cyanidin 3-(feruloyl) diglucoside-5-(malonoyl) glucoside and cyanidin 3-*O*-(sinapoyl)(feruloyl) diglucoside-5-*O*-(malonyl) glucoside (Figure 1; Table 1 and Table S1). The total amount of anthocyanins was 32.31 ± 2.84 mg/g dry weight in the innermost leaf and 10.17 ± 1.69 mg/g dry weight in the outermost leaf of RPCC. No anthocyanins were detected in the GCC leaves (Table 1). Interestingly, the RPCC pigment peaks 8 and 7 and a few other pigments identified in this study (Table S1) had not been reported previously. This result indicates that cyanidin 3-(feruloyl) diglucoside-5-(malonoyl) glucoside and pelargonidin 3-(caffeoyl) diglucoside-5-(malonoyl) glucoside are specific to this variety of RPCC, and contribute to its color.

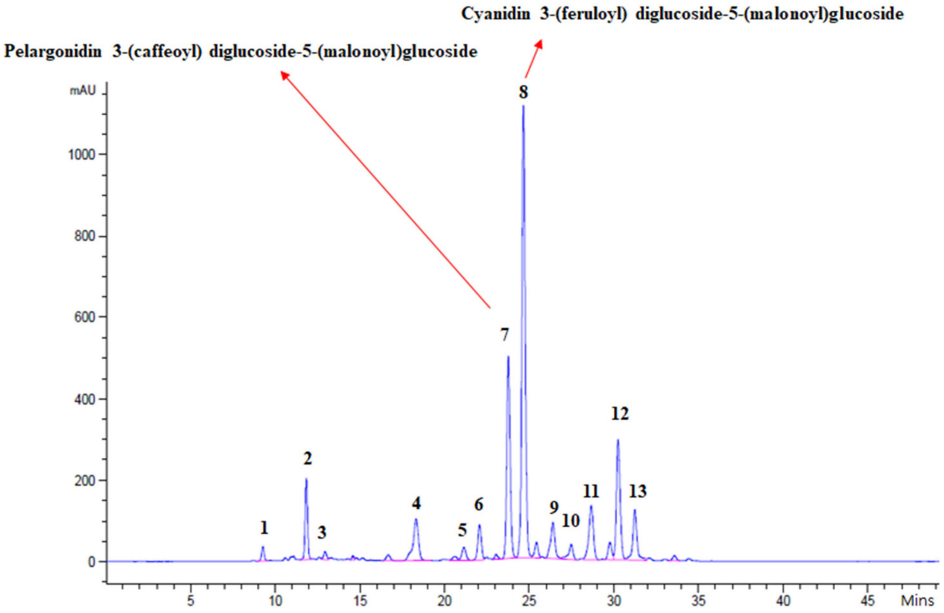


Figure 1. HPLC chromatogram of anthocyanin pigments detected in leaf extract of reddish purple Chinese cabbage at 520 nm. Horizontal axis shows retention time (min); vertical axis indicates strength of the peak (mAU).

Table 1. Total anthocyanin pigments identified in outer and inner leaf tissues of reddish purple and green Chinese cabbage.

No. of Pigments	Trivial Names	RPCC_IL	RPCC_OL	GCC_IL	GCC_OL
1	Cyanidin 3-diglucoside-5-glucoside	0.26 ± 0.00	0.04 ± 0.05	ND	ND
2	Cyanidin 3-diglucoside-5-(malonyl)glucoside	1.74 ± 0.23	0.29 ± 0.23	ND	ND
3	Cyanidin 3-(feruloyl)diglucoside-5-glucoside	0.19 ± 0.01	0.00 ± 0.00	ND	ND
4	Cyanidin 3-(caffeoyl)diglucoside-5-(malonyl)glucoside	1.93 ± 0.23	0.26 ± 0.19	ND	ND
5	Cyanidin 3-(p-coumaroyl)diglucoside-5-glucoside	0.34 ± 0.03	0.13 ± 0.09	ND	ND
6	Cyanidin 3-(feruloyl)diglucoside-5-glucoside	0.76 ± 0.09	0.08 ± 0.06	ND	ND
7	Pelargonidin 3-(caffeoyl)diglucoside-5-(malonoyl)glucoside	5.66 ± 0.60	0.81 ± 0.58	ND	ND
8	Cyanidin 3-(feruloyl)diglucoside-5-(malonoyl)glucoside	12.63 ± 1.37	2.19 ± 1.60	ND	ND
9	Cyanidin 3-(feruloyl)(feruloyl)diglucoside-5-glucoside	1.25 ± 0.09	0.19 ± 0.14	ND	ND
10	Cyanidin 3-*O*-(sinapoyl)(feruloyl)diglucoside-5-*O*-glucoside	0.41 ± 0.06	0.16 ± 0.12	ND	ND
11	Cyanidin 3-*O*-(p-coumaroyl)(sinapoyl)diglucoside-5-*O*-(malonyl)glucoside	1.92 ± 0.12	0.61 ± 0.45	ND	ND
12	Cyanidin 3-*O*-(sinapoyl)(feruloyl)diglucoside-5-*O*-(malonyl)glucoside	3.55 ± 0.26	1.24 ± 0.92	ND	ND
13	Cyanidin 3-*O*-(p-coumaroyl)(sinapoyl)diglucoside-5-*O*-(malonyl)glucoside	1.68 ± 0.12	0.77 ± 0.61	ND	ND
Total		32.31 ± 2.84	10.17 ± 1.69	ND	ND

RPCC-reddish purple Chinese cabbage; GCC- green Chinese cabbage; IL-inner leaf; OL- outer leaf; anthocyanin content-mg/g dry weight.

2.2. RNA-Sequencing of RPCC and GCC Samples

To identify differences in gene transcription between RPCC and GCC, we obtained and sequenced 53,127,646 (GCC) and 48,179,516 (RPCC) total reads, making up 5.37 Gb (giga bases) and 4.87 Gb, respectively (Table 2) with a GC content of 47.6% and 47.2%, respectively. After filtering, 38,611,432 and 34,924,846 clean reads were obtained for the GCC and RPCC samples, respectively, with Q30 values of 94.4% and 94.1%, respectively. A total 32,395 non-redundant transcripts were identified with varying lengths, most in the range of 501 to 1000 nucleotide bases, followed by 200–500 and 1001–1500 nucleotide bases (Figure S1a). Among the identified transcripts, 90.2% were annotated to the following public databases: NR-Viridiplantae (88.61%)¸ Phytozome (88.05%), UniProtKB–Viridiplantae (85.31%), KOG (84.06%), GO (83.72%), InterProscan (68.43%), and KEGG (18.87%) (Figure S1b).

Table 2. Summary of RNA sequence data.

	Reddish Purple Chinese Cabbage (RPCC)	**Green Chinese Cabbage (GCC)**
Total Reads	48,179,516	53,127,646
Total Bases	4866,131,116	5365,892,246
Total Bases(Gb)	4.87 Gb	5.37 Gb
GC_ Count	2320,659,939	2532,951,037
N_ Zero Reads	48,028,490	52,961,958
N5_ Less Reads	48,091,492	53,032,280
N_ Rate	0.03%	0.03%
Q20_More Bases	4699,696,370	5195,093,244
Q30_ More Bases	4579,319,019	5069,583,507
Clean reads	34,924,846	38,611,432
Clean bases	3073,838,581	3400,979,929

2.3. Identification of DEGs

The cDNA libraries of RPCC and GCC were mapped to the *B. rapa* reference genome with coverage of 95.2% and 94.66%, respectively (Figure 2a). Among the mapped and annotated DEGs, the highest proportion had RPKM (reads per kilobase per million mapped reads) values of >10, followed by RPKM values of 1 to 4 (Figure 2b). Among the predicted 3345 DEGs, 2706 were up-regulated and 639 were down-regulated in RPCC vs. GCC (Figure 2c; Table S2 and Table S3). Of them, 643 of the up-regulated DEGs and 354 of the down-regulated DEGs between RPCC and GCC samples were not functionally characterized (Table S2 and Table S3). The DEGs included many ABGs. A Bland–Altman (MA) plot was constructed to show the differentiation of gene expression between the two samples by plotting the values onto M (log ratio) and A (mean average) scales. The differences in gene expression are shown on the MA plot, where genes with ≥1-fold expression values are shown in red and those with negative log2 fold values are shown in green (Figure S2).

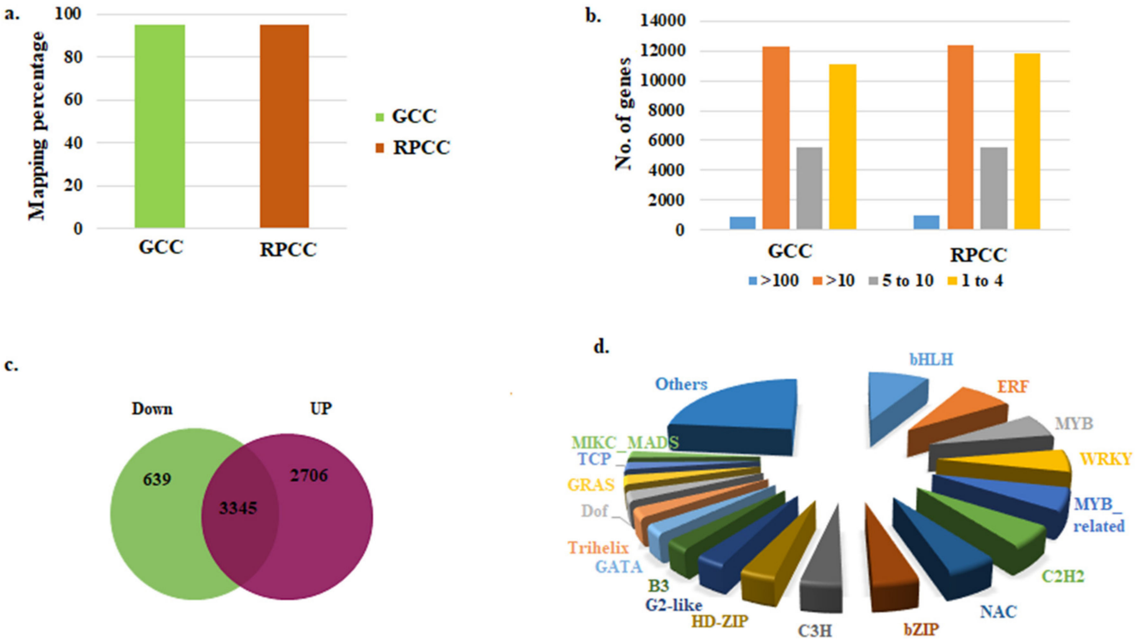


Figure 2. RNA-seq data for reddish purple Chinese cabbage. (**a**) Percentage of transcripts mapped to reference genome; (**b**) Gene expression values (RPKM); (**c**) Differentially expressed genes (DEGs) between two genotypes [green (GCC) and reddish purple (RPCC)]; (**d**) Transcription factor families identified in the transcriptome.

2.4. Identification of Transcription Factor Genes

Transcription factors play crucial roles in regulating gene expression, and many TFs control ABGs [30]. Hence, we searched for TFs that regulate anthocyanin biosynthesis in *B. rapa*. We detected 1625 TF genes in 54 TF families from our transcriptome data (Figure 2d; Table S4 and Table S5). The proportions of TF genes in different TF families were as follows: 8.7% in the basic helix loop helix (bHLH) family, 7.6% in the ethylene response factor (ERF) family, 6.05% in the myeloblastosis (MYB) family, 5.77% in the WRKY family, and 5.72% in the MYB-related family (Figure 2d). Additionally, 218 TF genes with log2fold change expression ≥1 and 37 TFs with log2fold change expression ≤ 1 were identified (Table S4 and Table S5). A few TF genes had very high log2fold change values (5–11.3), including *MYB90 (Bra004162)*, *MYB75 (Bra039763)*, three *RRTF1s (Bra017656*, *Bra011529*, *Bra034624)*, *CBF4 (Bra028290)*, *MYBL2 (Bra016164)*, *TTG2 (Bra023112)*, *DDF1 (Bra019777)* and *ERF-13 (Bra037630)* (Table 3). These DEGs may be involved in various functions related to pigment accumulation in RPCC.

Table 3. Summary of important regulatory transcriptional factor family genes identified from transcriptome data.

BRAD ID	E-Value	Identity	TAIR Description				Log2 Fold Change
			A. thaliana Id	TF Family	Gene Annotation	Gene Description	RPCC vs. GCC
					AP2/ERF Transcription Factors		
Bra011782	0	100	AT4G37750	AP2	*ANT, CKC, CKC1, DRG*	Integrase-type DNA-binding superfamily protein	2.21
Bra017656	0	99.88	AT4G34410	ERF	*RRTF1*	Redox responsive transcription factor 1	6.90
Bra011529	0	99.75	AT4G34410	ERF	*RRTF1*	Redox responsive transcription factor 1	6.72
Bra019087	0	100	AT2G20350	ERF	–	Integrase-type DNA-binding superfamily protein	6.50
Bra034624	0	99.8	AT4G34410	ERF	*RRTF1*	Redox responsive transcription factor 1	6.16
Bra028290	0	100	AT5G51990	ERF	*CBF4, DREB1D*	C-repeat-binding factor 4	5.94
Bra019777	0	98.04	AT1G12610	ERF	*DDF1*	Integrase-type DNA-binding superfamily protein	5.06
Bra037630	0	96.93	AT2G44840	ERF	*ATERF13, EREBP, ERF13*	Ethylene-responsive element binding factor 13	5.00
Bra015882	0	94.68	AT1G74930	ERF	*ORA47*	Integrase-type DNA-binding superfamily protein	4.96
Bra027612	0	100	AT1G63030	ERF	*DDF2*	Integrase-type DNA-binding superfamily protein	4.88
Bra016763	0	100	AT1G12610	ERF	*DDF1*	Integrase-type DNA-binding superfamily protein	4.84
Bra024539	0	99.48	AT1G22810	ERF	–	Integrase-type DNA-binding superfamily protein	4.79
Bra026963	0	98.87	AT1G12610	ERF	*DDF1*	Integrase-type DNA-binding superfamily protein	4.72
Bra031069	0	100	AT1G19210	ERF	–	Integrase-type DNA-binding superfamily protein	4.64
Bra028759	0	94.66	AT5G05410	ERF	*DREB2,DREB2A*	DRE-binding protein 2A	4.58
Bra015660	0	100	AT1G77640	ERF	–	Integrase-type DNA-binding superfamily protein	4.04
Bra028291	0	100	AT5G52020	ERF	–	Integrase-type DNA-binding superfamily protein	3.98
Bra016400	0	100	AT1G21910	ERF	–	Integrase-type DNA-binding superfamily protein	3.95
Bra032901	0	100	AT1G28370	ERF	*ATERF11,ERF11*	ERF domain protein 11	3.66
Bra014925	0	100	AT3G23230	ERF	–	Integrase-type DNA-binding superfamily protein	3.64
Bra027614	0	85	AT1G12630	ERF	–	Integrase-type DNA-binding superfamily protein	3.62
Bra002377	0	100	AT5G21960	ERF	–	Integrase-type DNA-binding superfamily protein	3.62
Bra010881	0	99.46	AT1G28360	ERF	*ATERF12,ERF12*	ERF domain protein 12	3.57
Bra016136	0	100	AT1G71450	ERF	–	Integrase-type DNA-binding superfamily protein	3.36

Table 3. *Cont.*

BRAD ID	E-Value	Identity	TAIR Description				Log2 Fold Change
			A. thaliana Id	TF Family	Gene Annotation	Gene Description	RPCC vs. GCC
Bra003780	0	100	AT1G74930	ERF	*ORA47*	Integrase-type DNA-binding superfamily protein	3.29
Bra036022	0	99.44	AT1G21910	ERF	–	Integrase-type DNA-binding superfamily protein	3.28
Bra016518	3.00E−162	95.32	AT1G19210	ERF	–	Integrase-type DNA-binding superfamily protein	3.23
Bra029147	0	100	AT5G52020	ERF	–	Integrase-type DNA-binding superfamily protein	3.21
Bra024953	0	100	AT5G47230	ERF	*ATERF-5, ATERF5, ERF5*	Ethylene responsive element binding factor 5	3.14
Bra040309	0	99.15	AT1G44830	ERF	–	Integrase-type DNA-binding superfamily protein	3.14
Bra032665	0	100	AT2G44840	ERF	*ATERF13, EREBP, ERF13*	Ethylene-responsive element binding factor 13	3.00
Bra030957	0	99.3	AT1G53170	ERF	*ATERF-8, ATERF8, ERF8*	Ethylene response factor 8	2.99
Bra010880	0	98.27	AT1G28370	ERF	*ATERF11, ERF11*	ERF domain protein 11	2.91
Bra011383	0	98.7	AT4G32800	ERF	–	Integrase-type DNA-binding superfamily protein	2.91
Bra015478	1.00E−10	74.84	AT2G44840	ERF	*ATERF13, EREBP, ERF13*	Ethylene-responsive element binding factor 13	2.86
Bra021048	0	97.12	AT4G17500	ERF	*ATERF-1, ERF-1*	Ethylene responsive element binding factor 1	2.63
Bra024954	0	99.54	AT5G47220	ERF	*ATERF-2, ATERF2, ERF2*	Ethylene responsive element binding factor 2	2.61
Bra008952	0	98.76	AT5G11590	ERF	*TINY2*	Integrase-type DNA-binding superfamily protein	2.53
Bra035732	0	97.33	AT5G51190	ERF	–	Integrase-type DNA-binding superfamily protein	2.51
Bra040158	0	100	AT4G17490	ERF	*ATERF6, ERF-6-6, ERF6*	Ethylene responsive element binding factor 6	2.48
Bra017493	0	98.76	AT5G47230	ERF	*ATERF-5, ATERF5, ERF5*	Ethylene responsive element binding factor 5	2.33
Bra040159	0	100	AT4G17500	ERF	*ATERF-1,ERF-1*	Ethylene responsive element binding factor 1	2.22
Bra034535	0	100	AT4G32800	ERF	–	Integrase-type DNA-binding superfamily protein	2.11
Bra027736	0	97.6	AT1G64380	ERF	–	Integrase-type DNA-binding superfamily protein	2.00
bHLH transcription factors							
Bra033690	0	100	AT5G43650	bHLH	*BHLH92*	Basic helix-loop-helix (bHLH) DNA-binding superfamily protein	4.69
Bra027501	0	100	AT5G43650	bHLH	*BHLH92*	Basic helix-loop-helix (bHLH) DNA-binding superfamily protein	2.68
Bra035639	0	97.42	AT5G56960	bHLH	–	Basic helix-loop-helix (bHLH) DNA-binding family protein	2.21
Bra036640	0	100	AT1G62975	bHLH	–	Basic helix-loop-helix (bHLH) DNA-binding superfamily protein	2.20
bZIP transcription factors							
Bra010035	0	94.65	AT5G49450	bZIP	*AtbZIP1,bZIP1*	Basic leucine-zipper 1	3.10

Table 3. *Cont.*

BRAD ID	E-Value	Identity	TAIR Description				Log2 Fold Change
			A. thaliana Id	TF Family	Gene Annotation	Gene Description	RPCC vs. GCC
C2H2 transcription factors							
Bra006692	1.00E−160	88.04	AT5G59820	C2H2	*RHL41, ZAT12*	C2H2-type zinc finger family protein	4.46
Bra002528	0	98.96	AT5G59820	C2H2	*RHL41, ZAT12*	C2H2-type zinc finger family protein	3.43
Bra022436	0	97.64	AT3G19580	C2H2	*AZF2, ZF2*	Zinc-finger protein 2	2.99
Bra010922	0	99.57	AT1G27730	C2H2	*STZ, ZAT10*	Salt tolerance zinc finger	2.96
Bra001752	0	99.87	AT3G19580	C2H2	*AZF2, ZF2*	Zinc-finger protein 2	2.96
Bra032845	0	100	AT1G27730	C2H2	*STZ, ZAT10*	Salt tolerance zinc finger	2.81
Bra038219	0	98.11	AT3G19580	C2H2	*AZF2, ZF2*	Zinc-finger protein 2	2.50
C3H transcription factors							
Bra000170	1.00E−156	100	AT2G40140	C3H	*ATSZF2, CZF1, SZF2, ZFAR1*	Zinc finger (CCCH-type) family protein	2.77
Bra007205	0	98.79	AT3G55980	C3H	*ATSZF1, SZF1*	Salt-inducible zinc finger 1	2.41
Bra004982	4.00E−135	81.1	AT2G40140	C3H	*ATSZF2, CZF1, SZF2, ZFAR1*	Zinc finger (CCCH-type) family protein	2.16
Dof transcription factors							
Bra014297	0	98.01	AT1G51700	Dof	*ADOF1, DOF1*	DOF zinc finger protein 1	2.18
GRAS transcription factors							
Bra021063	0	99.75	AT4G17230	GRAS	*SCL13*	SCARECROW-like 13	2.06
Bra033813	0	91.33	AT3G46600	GRAS	–	GRAS family transcription factor	2.02
HD-ZIP transcription factors							
Bra005259	0	100	AT2G36610	HD-ZIP	*ATHB22, HB22*	Homeobox protein 22	3.85
Bra016300	0	100	AT1G26960	HD-ZIP	*AtHB23, HB23*	Homeobox protein 23	2.02
LBD transcription factors							
Bra021433	5.00E-124	83.53	AT3G02550	LBD	*LBD41*	LOB domain-containing protein 41	2.90
MADS transcription factors							
Bra017376	0	99.84	AT2G03710	MIKC_ MADS	*AGL3,SEP4*	K-box region and MADS-box transcription factor family protein	2.33
Bra024521	0	99.59	AT1G22590	M-type_ MADS	*AGL87*	AGAMOUS-like 87	2.10
Bra005166	0	100	AT2G28700	M-type_ MADS	*AGL46*	AGAMOUS-like 46	2.05
MYB transcription factors							
Bra004162	2.00E−111	84.01	AT1G66390	MYB	*ATMYB90, MYB90, PAP2*	MYB domain protein 90	11.38

Table 3. *Cont.*

BRAD ID	E-Value	Identity	TAIR Description				Log2 Fold Change
			A. thaliana Id	TF Family	Gene Annotation	Gene Description	RPCC vs. GCC
Bra039763	0	92.86	AT1G56650	MYB	*ATMYB75, MYB75, PAP1, SIAA1*	Production of anthocyanin pigment 1	6.87
Bra029990	0	100	AT3G50060	MYB	*MYB77*	MYB domain protein 77	3.89
Bra012910	0	100	AT3G50060	MYB	*MYB77*	MYB domain protein 77	2.82
Bra013000	9.00E−118	100	AT5G60890	MYB	*ATMYB34, ATR1, MYB34*	MYB domain protein 34	2.51
Bra016164	0	100	AT1G71030	MYB_ related	*ATMYBL2, MYBL2*	MYB-like 2	5.55
Bra007957	0	97.7	AT1G71030	MYB_ related	*ATMYBL2, MYBL2*	MYB-like 2	4.06
Bra022637	5.00E−79	89.91	AT5G53200	MYB_ related	*TRY*	Homeodomain-like superfamily protein	2.34
				NAC transcription factors			
Bra008553	0	94.59	AT4G01550	NAC	*anac069, NAC069*	NAC domain containing protein 69	4.31
Bra020188	0	99.8	AT5G22380	NAC	*anac090, NAC090*	NAC domain containing protein 90	3.14
Bra006624	0	99.58	AT5G22380	NAC	*anac090, NAC090*	NAC domain containing protein 90	2.86
Bra027238	0	100	AT3G15500	NAC	*ANAC055,ATNAC3,NAC055, NAC3*	NAC domain containing protein 3	2.62
Bra037283	0	99.52	AT2G17040	NAC	*anac036, NAC036*	NAC domain containing protein 36	2.17
Bra013034	0	98.94	AT2G17040	NAC	*anac036, NAC036*	NAC domain containing protein 36	2.08
				WRKY transcription factors			
Bra023112	0	99.53	AT2G37260	WRKY	*ATWRKY44, DSL1,TTG2, WRKY44*	WRKY family transcription factor family protein	5.13
Bra014426	0	99.88	AT2G46400	WRKY	*ATWRKY46, WRKY46*	WRKY DNA-binding protein 46	3.87
Bra003588	0	99.5	AT1G80840	WRKY	*ATWRKY40, WRKY40*	WRKY DNA-binding protein 40	3.81
Bra035148	0	100	AT1G80840	WRKY	*ATWRKY40, WRKY40*	WRKY DNA-binding protein 40	3.69
Bra005210	0	100	AT2G37260	WRKY	*ATWRKY44, DSL1,TTG2, WRKY44*	WRKY family transcription factor family protein	3.54
Bra035147	0	99.52	AT1G80850	WRKY	–	DNA glycosylase superfamily protein	3.53
Bra033158	0	98.49	AT4G11070	WRKY	*AtWRKY41, WRKY41*	WRKY family transcription factor	2.95
Bra020196	0	90.34	AT5G22570	WRKY	*ATWRKY38, WRKY38*	WRKY DNA-binding protein 38	2.87
Bra010032	0	99.58	AT5G49520	WRKY	*ATWRKY48, WRKY48*	WRKY DNA-binding protein 48	2.86
Bra019265	0	100	AT4G23810	WRKY	*ATWRKY53, WRKY53*	WRKY family transcription factor	2.65
Bra013731	0	99.57	AT4G23800	WRKY	–	HMG (high mobility group) box protein	2.41
Bra023998	0	100	AT4G31550	WRKY	*WRKY11*	WRKY DNA-binding protein 11	2.27
Bra031900	0	98.34	AT5G64810	WRKY	*WRKY51*	WRKY DNA-binding protein 51	2.13

2.5. Functional Characteristics of DEGs and KEGG Pathway Enrichment Analysis

We performed GO analyses to annotate the DEGs. The top 20 enriched terms in each GO category of DEGs were selected based on significance ($p < 0.05$) and are summarized in Figure 3. In the BP category, the majority of DEGs were involved in "response to chemical stimulus" followed by "response to organic substance", "response to endogenous stimulus", and "cellular response to chemical stimulus" processes. Similarly, in the MF category, the majority of genes were involved in "DNA binding", followed by "transcription factor activity", "sequence specific DNA binding" and "calcium ion binding". In the CC category, many of the DEGs were related to "extracellular region", "cell wall", "external encapsulating structure" and "plant type cell wall" (Figure 3). We further categorized the up-regulated and down-regulated genes. Among the up-regulated DEGs, a total of 385 GO terms were identified, comprising 328 in the BP category, 35 in the MF category, and 22 in the CC category. Among the down-regulated DEGs, a total of 210 GO terms were identified, comprising 133 terms in the BP category, 50 in the MF category, and 22 in the CC category (Table S6.) In the BP category, a few genes were involved in "biosynthetic and metabolic processes of anthocyanins, ethylene signaling, flavonoids and phenylpropanoids", "catabolic and metabolic processes of L-phenylalanine", "cinnamic acid biosynthesis", "response to absence of light", and "response to temperature stimulus". In the MF category, some genes were involved in "phenylalanine ammonia-lyase activity" and "O-methyltransferase activity" (Table S6.). Some of the down-regulated genes were involved in "biosynthesis process", "negative regulation of cellular metabolic process", and "response to UV-C" (Table S6).

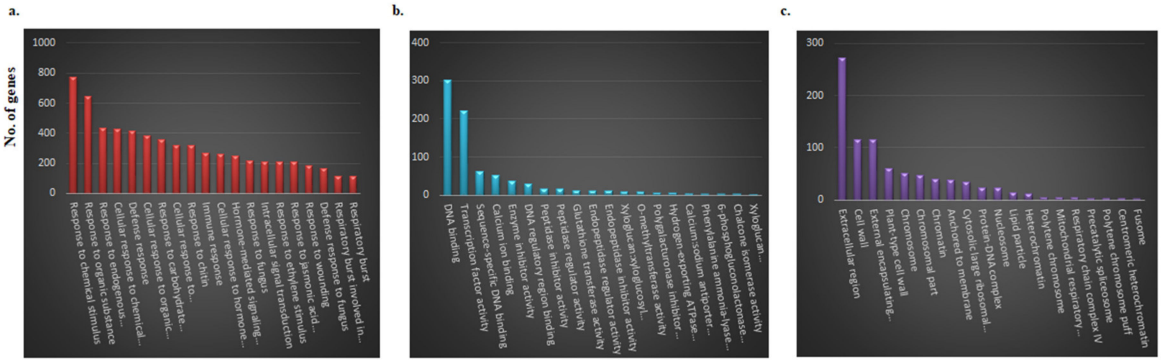


Figure 3. Gene ontology (GO) analysis of differentially expressed genes (DEGs) between red and green Chinese cabbage. DEGs were grouped into three categories: (**a**) Biological process; (**b**) molecular function; and (**c**) cellular component. X-axis shows gene annotation term; y-axis shows number of genes.

Next, we conducted a KEGG enrichment analysis to identify pathways significantly enriched with DEGs. The pathways enriched with up-regulated DEGs were "biosynthesis of secondary metabolites", "metabolic pathways", "biosynthesis of flavonoids and phenylpropanoids", "metabolism of fructose, mannose, starch, and sucrose", and several others (Figure 4 and Table S7). The down-regulated DEGs were involved in 46 types of functions related to organ development, and other functions related to plant growth and development (Table S7).

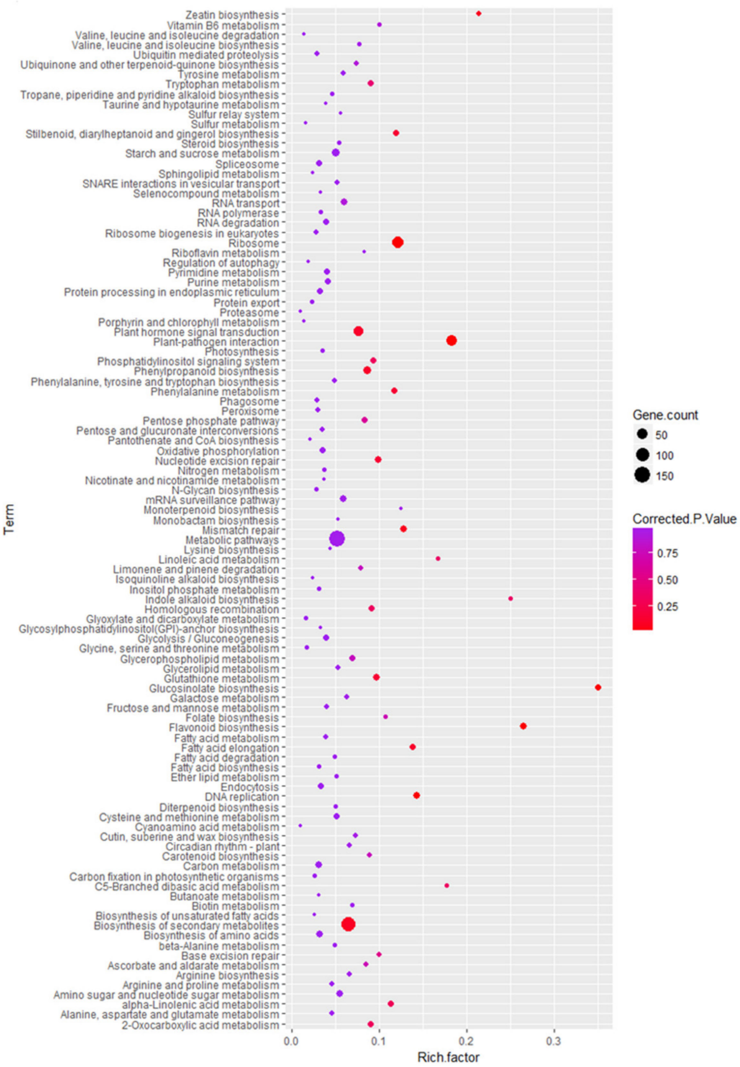


Figure 4. KEGG enrichment analysis of differentially expressed anthocyanin biosynthetic genes. *X*-axis shows KEGG terms and y-axis shows enrichment factor. Gene count and corrected *p*-values are shown on right.

2.6. Genes Related to Anthocyanin Biosynthesis Identified from Transcriptome Data

From the transcriptome data, we identified 255 ABGs (Table S8) comprising 58 phenylpropanoid biosynthetic genes (PBGs), 56 early biosynthetic genes (EBGs), 67 late biosynthetic genes (LBGs), 19 anthocyanin transporter genes (ATGs), 29 other anthocyanin biosynthesis regulatory genes (OABRGs) and 26 regulatory TF genes (Table S8). The main PBGs were those encoding cinnamyl alcohol dehydrogenase (CAD), caffeoyl CoA O-methyltransferase (CCoAMT), cinnamoyl-CoA reductase (CCR), cinnamate 4-hydroxylase (C4H), 4-coumarate: coenzyme A ligase (4CL), O-methyltransferase (OMT), and phenylalanine ammonia lyase (PAL). These genes are involved in different stages of the phenylpropanoid biosynthesis pathway and showed large differences in transcript levels (−0.7 to 3.2-fold change) between RPCC and GCC (Table S8). Similarly, EBGs encoding chalcone isomerase (CHI), chalcone synthase (CHS), flavanone-3-hydroxylase (F3H), flavonoid 3′-hydroxylase (F3′H), and flavonol synthase (FLS) showed large differences in their transcript levels between RPCC and GCC (−1.1 to 4.9-fold change) (Table S8). The LBGs showing large differences in transcript levels (−1.9 to 9.3-fold change) between RPCC and GCC encoded beta glucosidase (BGLU), dihydroflavonol 4-reductase (DFR), leucoanthocyanidin dioxygenase (LDOX), 5-*O*-glucoside-6-*O*-malonyltransferase (5MAT), UDP-glucose: flavonoid 3-o-glucosyltransferase (UF3GT), and UDP-glucosyltransferases

(UGT). Genes for anthocyanin transporters and regulatory TFs included those encoding glutathione S-transferase 26/TRANSPARENT TESTA 19 (GST26/TT19), multidrug and toxic compound extrusion (MATE), basic helix-loop-helix 32 (BHLH32), ENHANCER OF GLABRA 3 (EGL3), GLABRA 3 (GL3), myeloblastosis protein 75 (MYB75), myeloblastosis protein 90 (MYB90), TRANSPARENT TESTA 8 (TT8) and TRANSPARENT TESTA GLABRA 1 (TTG1), and TRANSPARENT TESTA GLABRA (TTG2). These genes showed differences in expression between RPCC and GCC ranging from a −0.5 to 11.3-fold change. Interestingly, we identified a few OABRGs with high log2fold expression values (≥ 2) (Table S8). These results indicate that ABGs have vital roles in anthocyanin biosynthesis, transportation, and accumulation in the leaf tissues of RPCC at the seedling stage (3 weeks).

2.7. Expression Analysis of Anthocyanin Biosynthetic Genes by qRT-PCR

In general, gene expression studies can demonstrate the biological activity of genes in plants. We confirmed the reproducibility and accuracy of DEGs identified in our transcriptome data through qRT-PCR analyses. Although transcriptome sequencing was performed using samples from 3-week-old plants (seedlings), we checked the expression profiles of genes in samples from 6-week-old (rosette stage) and 9-week-old (heading stage) plants of GCC and RPCC. This analysis of gene expression at three developmental stages sheds light on the expression profile of ABGs and their effect on anthocyanin biosynthesis and accumulation throughout plant development. In general, the anthocyanin biosynthesis pathway can be classified into three phases; 1. The phenylpropanoid pathway; 2. Early steps of the flavonoid pathway; and 3. The anthocyanin pathway [31]. For the validation of gene expression, we selected 60 genes identified from the transcriptome data and from a previous study [26] that are involved in various stages of anthocyanin biosynthesis (Table 4; Table S9).

In the phenylpropanoid pathway, three PAL genes (*BrPAL1*, *BrPAL2* and *BrPAL4*), *BrC4H*, and the 4CL homolog *Br4CL2* were detected in GCC and RPCC at three developmental stages. The transcriptome data from 3-week-old plants (heat map, Figure 5a) and the qRT-PCR analyses showed that the transcript levels of these genes were much higher in RPCC than in GCC. Gene expression was also compared among stages (seedling, rosette, and heading stages) (Figure 5a). Similar to PBGs, EBGs such as *BrCHS*, *BrCHI* and *BrCHI1, BrF3H,* and *BrF3′H-1* showed similar expression patterns in both the transcriptome analysis (heat map, Figure 5b) and qRT-PCR analyses (Figure 5b). Unlike PBGs and EBGs, the LBG *BrDFR* was expressed only in the RPCC at all stages, while the LBG *BrLDOX* was expressed at lower levels at the early stage. The *BrLDOX* transcript levels gradually increased from the rosette to the heading stage in RPCC (Figure 5c). The transcript levels of MYBs including *BrMYB90, BrMYB75, and BrMYBL2-1* varied among different stages. *BrMYB90* and *BrMYB75* showed maximum transcript levels in RPCC at the seedling stage, while *BrMYBL2-1* had higher transcript levels at the rosette and heading stages than at the seedling stage in RPCC (Figure 5d).

The downstream LBGs are involved in the acylation, glycosylation, and methylation of anthocyanins and include genes encoding acyltransferase (AT), glycosyltransferase (GT) and O-methyltransferase (OMT) [32,33]. The downstream LBGs selected from the transcriptome data for qRT validation included *BrUF3GT*, *BrUGT75C1-1, BrUGT73B2, Br5MAT, BrAT-1, and BrAT-2. BrUF3GT, BrUGT75C1-1*, and *BrUGT73B2* showed increased expression while *Br5MAT* and *BrAT-1* showed decreased expression from the seedling to heading stages in RPCC. The highest transcript level of *BrAT-2* was at the rosette stage in RPCC (Figure 5e). Interestingly, all the LBGs including regulatory MYB (RM) genes showed low or no expression in GCC compared with RPCC in the transcriptome data (heat map, Figure 5e) and in the qRT-PCR analyses (Figure 5e).

Table 4. Details of 60 genes selected for validation by qRT_PCR analysis.

Given I.D	**BRAD Id**			**Gene Position V 1.5**				***A. thaliana* Id**	**Gene Annotation**
	B. rapa Id	**Identity**	**E-Value**	**Chromosome**	**Start**	**End**	**Strand**		
Phenylpropanoid pathway genes									
BrPAL1	Bra017210	98.94	0	A04	16132008	16134574	-	AT2G37040	PAL1
BrPAL2	Bra003126	98.68	0	A07	14754085	14756786	-	AT3G53260	PAL2
BrPAL4	Bra029831	100	0	A05	22819640	22826303	-	AT3G10340	PAL4
BrC4H	Bra021637	97.23	0	A04	13688684	13690602	-	AT2G30490	C4H
Br4CL2	Bra031266	100	0	A05	17255035	17257878	+	AT3G21240	4CL2,AT4CL2
Early anthocyanin biosynthesis genes									
BrCHS	Bra006224	99.5	0	A03	2596137	2597594	+	AT5G13930	CHS,TT4
BrCHI	Bra007142	99.47	0	A09	29055564	29057157	-	AT3G55120	CHI,TT5
BrCHI1	Bra009101	100	0	A10	15229294	15230280	-	AT5G05270	CHI1
BrF3H	Bra036828	99.91	0	A09	27095567	27097080	+	AT3G51240	F3H,TT6
BrF3′H-1	Bra009312	100	0	A10	14356094	14358845	-	AT5G07990	F3′H,TT7
BrF3′H-2	Bra020459	90.02	0	A02	5846392	5848174	-	AT5G57220	F3′H,TT7
BrF3′H-3	Bra019366	99.68	0	A03	24701345	24702922	+	AT4G22690	F3′H,TT7
BrF3′H-4	Bra030246	100	0	A04	10022729	10024935	+	AT2G22330	F3′H,TT7
BrF3′H-5	Bra011280	94.06	0	A01	2958371	2960458	-	AT1G13080	F3′H,TT7
Late anthocyanin biosynthesis genes									
BrDFR	Bra027457	100	0	A09	10926334	10927890	-	AT5G42800	DFR,M318,TT3
BrLDOX	Bra013652	98.48	0	A01	6885692	6887113	-	AT4G22880	LDOX
BrUF3GT	Bra003021	99.81	0	A10	6063162	6064651	-	AT5G54060	UF3GT
BrUGT75C1-1	Bra038445	100	0	A08	8755970	8757334	-	AT4G14090	UGT75C1
BrUGT75C1-2	Bra039545	100	0	A01	11553894	11555345	+	AT4G15490	UGT75C1
BrUGT73B2	Bra034610	100	0	A08	11760868	11762581	+	AT4G34138	UGT73B2

Table 4. *Cont.*

Given I.D	BRAD Id			Gene Position V 1.5				A. thaliana Id	Gene Annotation
	B. rapa Id	**Identity**	**E-Value**	**Chromosome**	**Start**	**End**	**Strand**		
BrUGT78D2	Bra023594	-	-	A02	3087246	3088800	-	AT5G17050	UGT78D2
Br5MAT	Bra036208	98.82	0	A09	1925788	1927140	-	AT3G29590	5MAT
BrAT-1	Bra030550	98.52	0	A08	20600205	20602987	+	AT1G03940	Acyl-transferase family protein-1
BrAT-2	Bra034255	100	0	A04	11806054	11806788	-	AT3G29680	Acyl-transferase family protein
Anthocyanin transporter genes									
BrMATE2-1	Bra031776	88.51	0	A09	36544120	36546929	+	AT1G11670	TT12
BrMATE-2	Bra027073	99.8	0	A09	8642073	8645111	+	AT1G61890	TT12
BrTT19-1	Bra008570	100	0	A10	11677671	11678470	-	AT5G17220	GST26,TT19
BrTT19-2	Bra023602	99.69	0	A02	3117740	3118547	+	AT5G17220	TT19
Other anthocyanin biosynthesis genes									
BrOMT1-2	Bra011292	99.13	0	A01	2896202	2897686	-	AT1G77520	OMT1
BrCCR2	Bra008438	100	0	A02	14649009	14650738	-	AT1G80820	CCR2
BrRNS1-1	Bra026570	99.86	0	A02	20391492	20392500	+	AT2G02990	RNS1
BrCCoAMT	Bra033968	100	0	A02	9548647	9549923	+	AT1G67980	CCoAMT
BrFLS3	Bra029212	99.68	0	A02	25974752	25976735	-	AT5G63590	ATFLS3,FLS3
BrLAC17	Bra006683	90.74	0	A03	4611941	4613968	-	AT5G60020	LAC17
BrOxygenase protein	Bra012691	100	0	A03	22738914	22741827	-	AT4G16770	oxygenase superfamily protein
BrSCPL10-2	Bra025601	98.52	0	A04	7905792	7908695	-	AT2G22980	SCPL10
BrBGLU10	Bra037647	100	0	A04	18448229	18455545	+	AT3G60120	BGLU10
BrFLS1	Bra022378	100	0	A05	18803675	18805312	+	AT3G19010	FLS1
BrIRX12	Bra005140	99.73	0	A05	3679277	3682262	-	AT2G38080	IRX12
BrBGLU46-1	Bra018969	99.24	0	A06	976174	979392	-	AT1G52400	BGLU46
BrOST2	Bra024452	96.24	9.00E−83	A06	16477147	16483736	-	AT2G18960	AHA1,HA1,OST2,PMA

Table 4. *Cont.*

Given I.D	BRAD Id			Gene Position V 1.5				*A. thaliana* Id	Gene Annotation
	B. rapa Id	**Identity**	**E-Value**	**Chromosome**	**Start**	**End**	**Strand**		
BrSCPL10-1	Bra012153	88.53	0	A07	11924527	11938309	+	AT2G23000	SCPL10
BrOMT1-1	Bra012269	98.11	0	A07	11175574	11176865	+	AT1G21100	OMT1
BrOMT	Bra003707	100	3.00E−43	A07	17900548	17902240	+	AT1G76790	O-methyltransferase family protein
BrCAD1-2	Bra010819	99.59	0	A08	15846126	15848906	+	AT1G29690	CAD1
BrCAD1-3	Bra010879	99.78	0	A08	16130363	16132914	-	AT1G28380	CAD1
BrCCoAOMT1	Bra034600	100	0	A08	11803933	11805048	-	AT4G34050	CCoAOMT1
BrCAD1-1	Bra026804	100	0	A09	35535138	35538610	-	AT1G14780	CAD1
BrRNS1-2	Bra026846	100	0	A09	35723513	35724548	+	AT1G14220	RNS1
BrCCR5	Bra008743	100	0	A10	12472468	12475087	+	AT5G14700	CCR5
BrBGLU46-2	Bra002978	98.56	0	A10	6383388	6385871	-	AT5G54570	BGLU46
BrOMT1-3	Bra003009	100	0	A10	6154442	6160001	+	AT5G54160	ATOMT1,OMT1
Regulatory transcription factors									
BrMYBR1	Bra012149	100	0	A07	11952841	11953731	-	AT5G67300	MYBR1
BrMYBL2-1	Bra016164	100	0	A07	22386380	22387236	-	AT1G71030	MYB-like 2
BrMYB15	Bra001907	99.77	0	A03	19315200	19316380	+	AT3G23250	MYB15
BrMYB51	Bra025666	97.36	0	A06	6841688	6842966	+	AT1G18570	MYB51
BrMYB51	Bra016553	100	0	A08	18248352	18249890	-	AT1G18570	MYB51
BrMYB75	Bra039763	92.86	0	A02	8839008	8840737	+	AT1G56650	PAP1,MYB75
BrMYB77	Bra012910	100	0	A03	21598456	21599337	-	AT3G50060	MYB77
BrMYB90	Bra004162	84.01	2.00E−111	A07	20426416	20431671	+	AT1G66390	MYB90

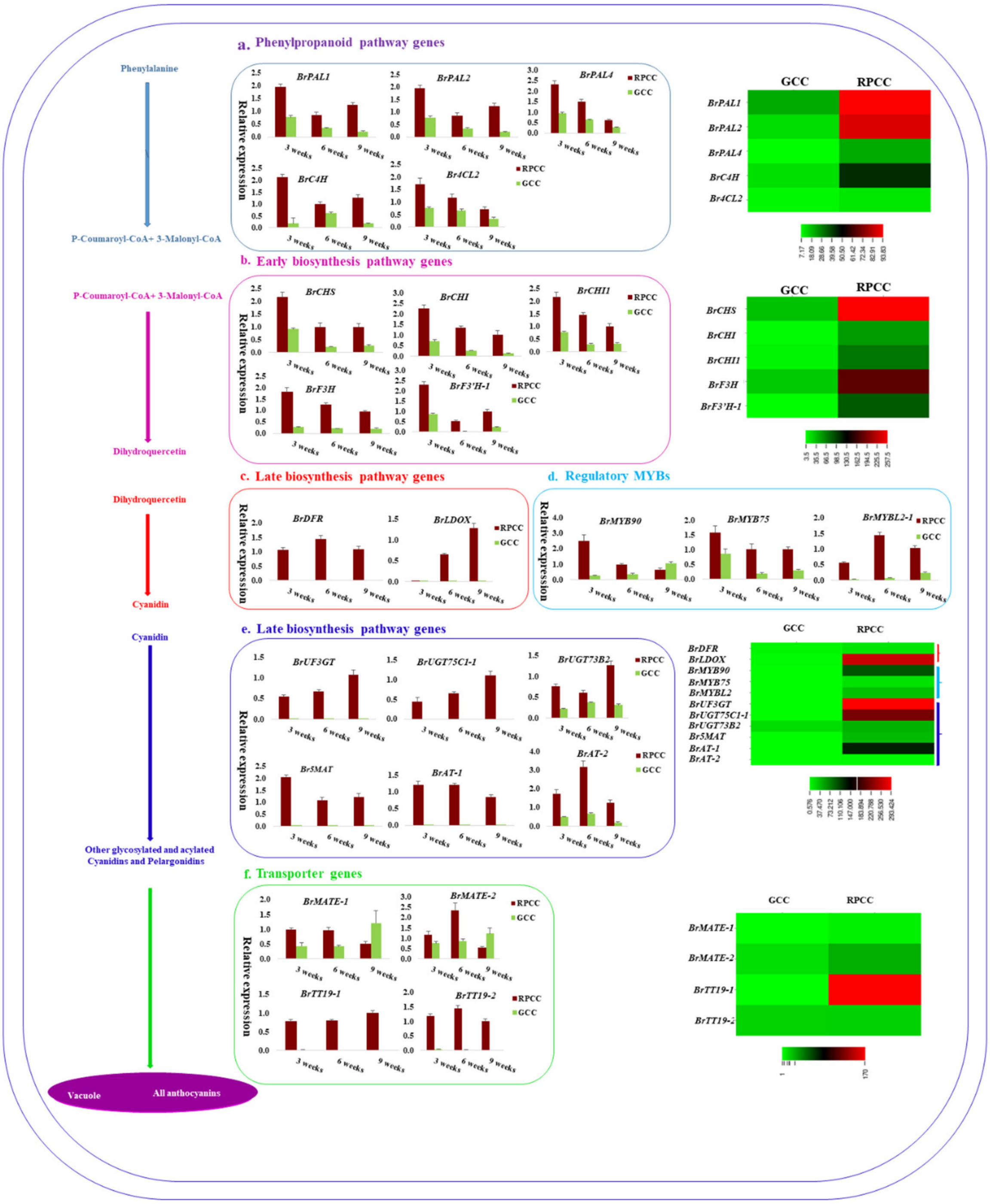


Figure 5. Validation of anthocyanin biosynthetic genes (ABGs) detected in transcriptome data by qRT-PCR analyses of reddish purple (RPCC) and green (GCC) leaf tissue samples. (**a**) Phenylpropanoid pathway genes; (**b**) early biosynthesis pathway genes; (**c**) and (**e**) late biosynthesis pathway genes; (**d**) regulatory MYB genes; and (**f**) transporter genes. Gene expression levels were normalized against that of *Actin*. Error bars are based on mean of three technical replicates. Schematic representation of anthocyanin biosynthetic pathway is shown in left corner. Heatmaps in middle and right corner indicate transcript abundance of ABGs.

Among many anthocyanin transporter genes, four *B. rapa* transporter genes including *BrMATE-1*, *BrMATE-2*, and *BrTT19-1* showed differential expression between GCC and RPCC (heat map, Figure 5f). At the seedling and rosette stages, both *BrMATE-1* and *BrMATE-2* were expressed at higher levels in

RPCC than in GCC. *BrTT19-1* and *BrTT19-2* transcript levels were high at all stages in RPCC, but at negligible or undetectable levels in GCC at all stages (Figure 5f). The remaining ABGs, showed diverse expression patterns in the qRT-PCR analyses (Figure S3). Most of the analyzed genes had higher transcript levels in RPCC than in GCC (Figure S3).

We also conducted qRT-PCR analyses for some MYB TF genes showing differences in expression levels between RPCC and GCC in transcriptome data (Table S4). Among the five selected MYB genes, BrMYB15 and BrMYB51-2 showed >1-fold expression in RPCC than in GCC at all stages, and BrMYB77 transcript levels increased from the rosette to the heading stage in RPCC but not in GCC (Figure 6).

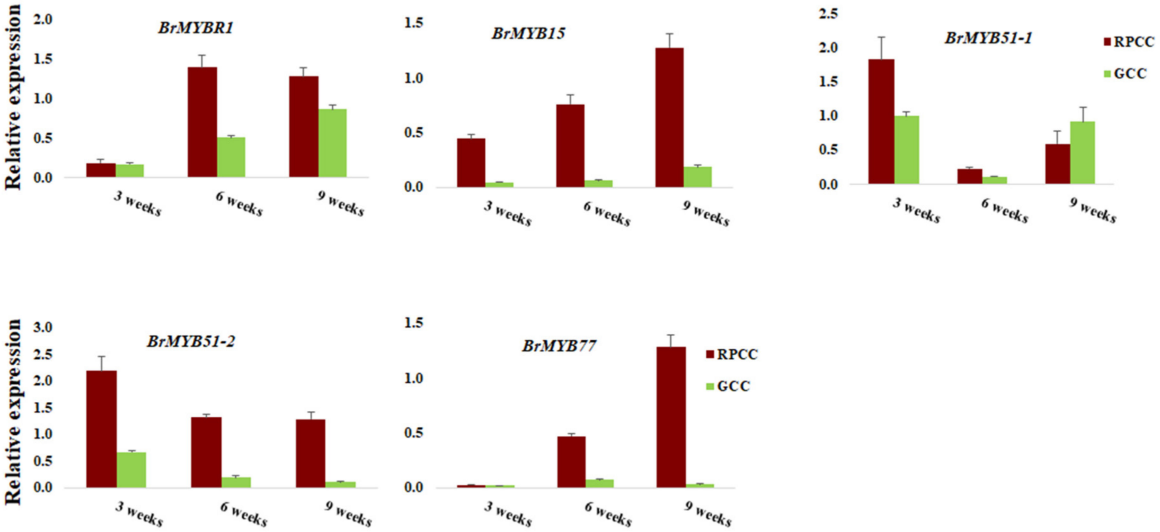


Figure 6. qRT-PCR validation of MYBs with high transcript abundance in transcriptome data. qRT-PCR expression values were normalized against that of *Actin*. Error bars are based on mean of three technical replicates.

A correlation analysis revealed a strong correlation between the transcriptome data and qRT-PCR data (R = 0.81) (Figure 7). Overall, similar trends in gene expression were detected from transcriptome data and qRT-PCR analyses. The expression patterns of PBGs, EBGs, LBGs, TGs, and RMs implied that LBGs, TGs and RMs play crucial roles in anthocyanin biosynthesis during different developmental stages of RPCC.

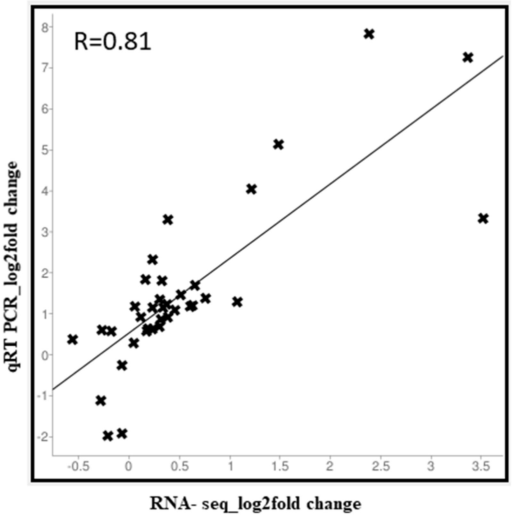


Figure 7. Correlation analysis between RNA-seq and qRT-PCR methods. Log2fold values of RNA-seq data (x-axis) are plotted against log2fold values of qRT-PCR (*y*-axis) data.

2.8. Promoter Analysis of Anthocyanin Biosynthetic Genes

Cis-regulatory elements (CREs) are binding sites for TFs in the promoters of target genes. To identify CREs, we analyzed the promoter regions of 22 important ABGs. The 2-kb region upstream of the transcription start site (TSS) was extracted and analyzed by the New PLACE program to find CRE motifs (Figure 8a). Among the predicted CREs, most were binding elements for MYB, bHLH, WRKY, bZIP and Ap2/ERF TFs (Figure 8b and Table S10). Those binding to MYB TFs were the most abundant, followed by those binding to bHLH, WRKY, bZIP and AP2/ERF TFs (Figure 8b). MYB CREs have been found in the promoters of genes related to secondary metabolism, flavonoid biosynthesis, anthocyanin biosynthesis, and plant defense (Table S10). The bHLH and bZIP CREs are known to be involved in the light response, tissue specific activation of phenylpropanoid biosynthetic genes, sugar repression, seed development, and the biosynthesis of phenylpropanoids, lignin, and flavonoids. AP2/ERF CREs are involved in functions related to the ethylene response, the jasmonate response, and secondary metabolism (Table S10). Therefore, the results of our study and other studies [34,35] indicate that MYB, and bHLH CREs regulate the expression of genes at all stages in the anthocyanin biosynthesis pathway.

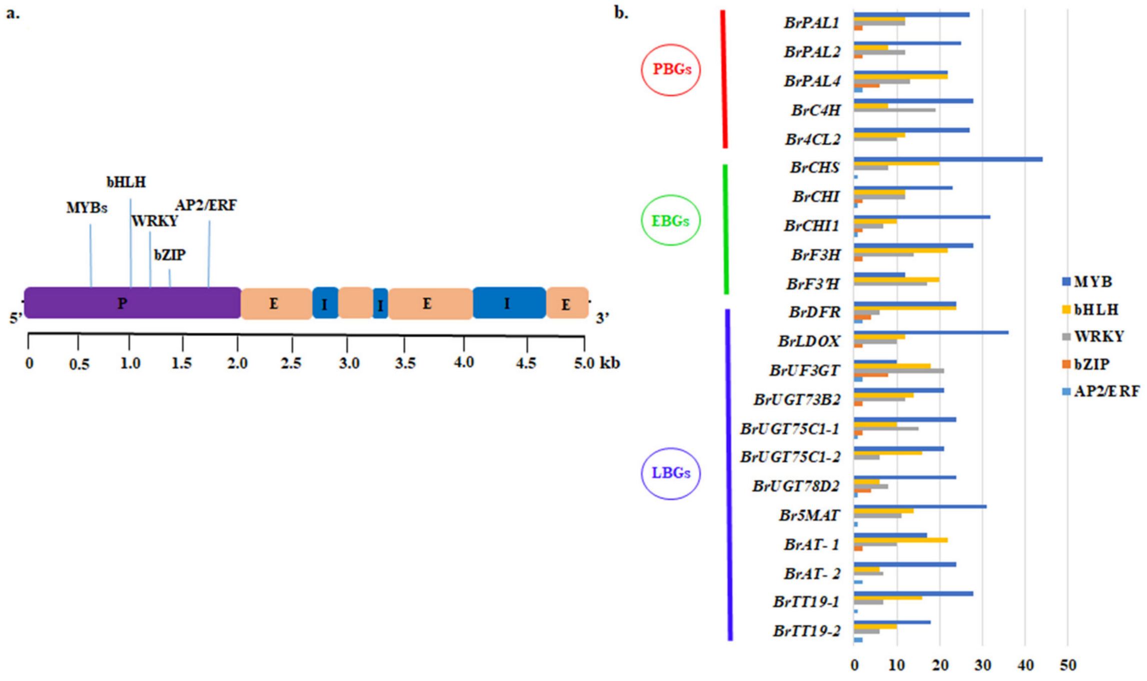


Figure 8. Cis-regulatory elements predicted in upstream promoter regions of anthocyanin biosynthetic genes (ABGs). (**a**) Example of plant gene organization and important cis-elements in promoter. (**b**) Number of each type of cis-element identified in ABGs. P, promoter; E, exon; I, intron.

2.9. Regulatory Network Analysis of Anthocyanin Biosynthesis Genes

To identify the interactions among anthocyanin biosynthetic genes and the transcription factor genes including MYB, bHLH, WRKY, bZIP, and AP2/ERF (with log2fold change > 2) (Tables 3 and 4), a putative interactive network was constructed (Figure 9). Among them, 37 *B. rapa* genes (yellow circles) showed 147 interactions, which could be classified into two types: activation (↓/↑) and repression (⊥) (Figure 9). The gene network results showed that MYB75 interacts with a gene encoding an acyl transferase family protein, as well as MYB90, 5MAT, TT5, AGT, TT4, UGT78D2, TT19, DFR, and UF3GT, and activates them to promote anthocyanin biosynthesis. Two LBGs (DFR and LDOX) are positively regulated by TFs such as PIF3, MYB32, HY5, and TT2 (Figure 9) and DFR is also positively regulated by the TT8 TF. Besides gene–gene interactions, this network analysis revealed many other interactions among TFs and structural genes involved in various functions. Among the repressors, the MYBL2 TF interacts with MYB75, DFR, TT2, TT8, GL2, and EGL3; MYB75 represses SCPL10; both EGL3 and GL3 repress LDOX; and bHLH32 represses DFR (Figure 9). These results indicate that interactions between TFs and their target genes play a vital role in the regulation of anthocyanin biosynthesis and other metabolic functions related to the growth and development of RPCC.

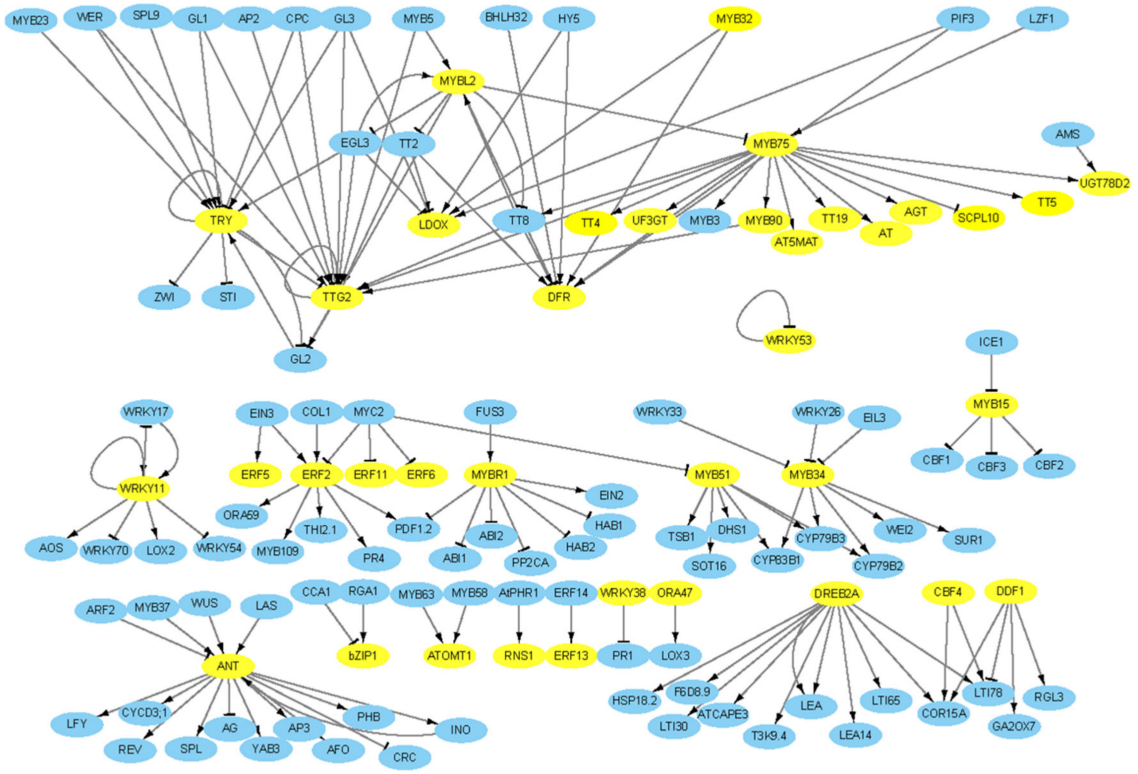


Figure 9. Gene regulatory network of anthocyanin biosynthetic genes and important transcription factors. DEGs detected from our transcriptome data are shown in yellow: other interactive genes involved in various functions including anthocyanin biosynthesis are shown in blue. Up tack (⊥) indicates repressors and arrow (↓/↑) indicates activators.

3. Discussion

The red color in Chinese cabbage has been introduced through different techniques of introgression breeding. Among the introduced varieties, Xie et al. [3] introgressed the red color phenotype by crossing a Chinese cabbage variety with red color *Brassica juncea* through the embryo rescue technique. While in our study, RPCC and red Chinese cabbage (RCC) reported by Lee et al. [4], have been developed through interspecific-crossing between Chinese cabbage and red cabbage.

3.1. Novel Anthocyanin Pigments Responsible for the Color of RPCC

Red Chinese cabbage is an economically important variety that is rich in various secondary metabolites including anthocyanins [4]. This variety accumulates red pigments at an early stage of plant growth (seedling stage), so it is an ideal system to study the genes and regulatory TFs that are involved in regulating color (anthocyanin) accumulation at the early stages of plant development (Figure 10). Fruits and vegetables contain five main types of anthocyanins, with different frequencies of occurrence: cyanidin (50%), delphinidin (12%), pelargonidin (12%), peonidin (12%), malvidin (7%), and petunidin (7%) [36]. We detected 11 anthocyanin variants in the RPCC samples, approximately 85% of which were cyanidin isoforms. Thus, our results and those of other studies show that cyanidin derivatives are the most abundant type of anthocyanins in red Chinese cabbage [4]. In addition, the anthocyanin pigments accumulated in RPCC and in previous studies [3,4] are entirely different, indicating that the color accumulation and regulation are due to different sets of anthocyanin pigments. In accordance with the results of the HPLC analysis, we propose a model of the anthocyanin biosynthesis pathway that generates cyanidin 3-(feruloyl) diglucoside-5-(malonoyl) glucoside, and pelargonidin 3-(caffeoyl) diglucoside-5-(malonoyl) glucoside in RPCC (Figure 11). Most of the anthocyanin components identified in this study have been detected in radish or other *Brassica* crops but not in red Chinese cabbage [20–23,37–43]. Hence, 11 of the 13 pigments identified in our study were detected for the first time in RPCC (Table S1). Joo et al [28] have proved that the anthocyanin rich extract has the ability to lower the risk of vascular inflammatory diseases.

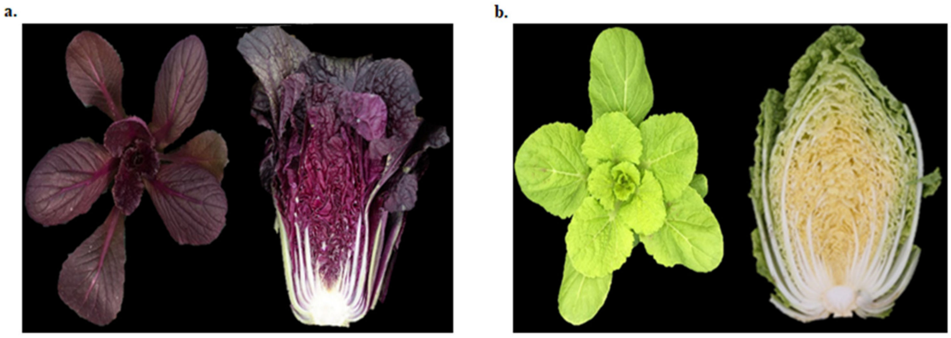
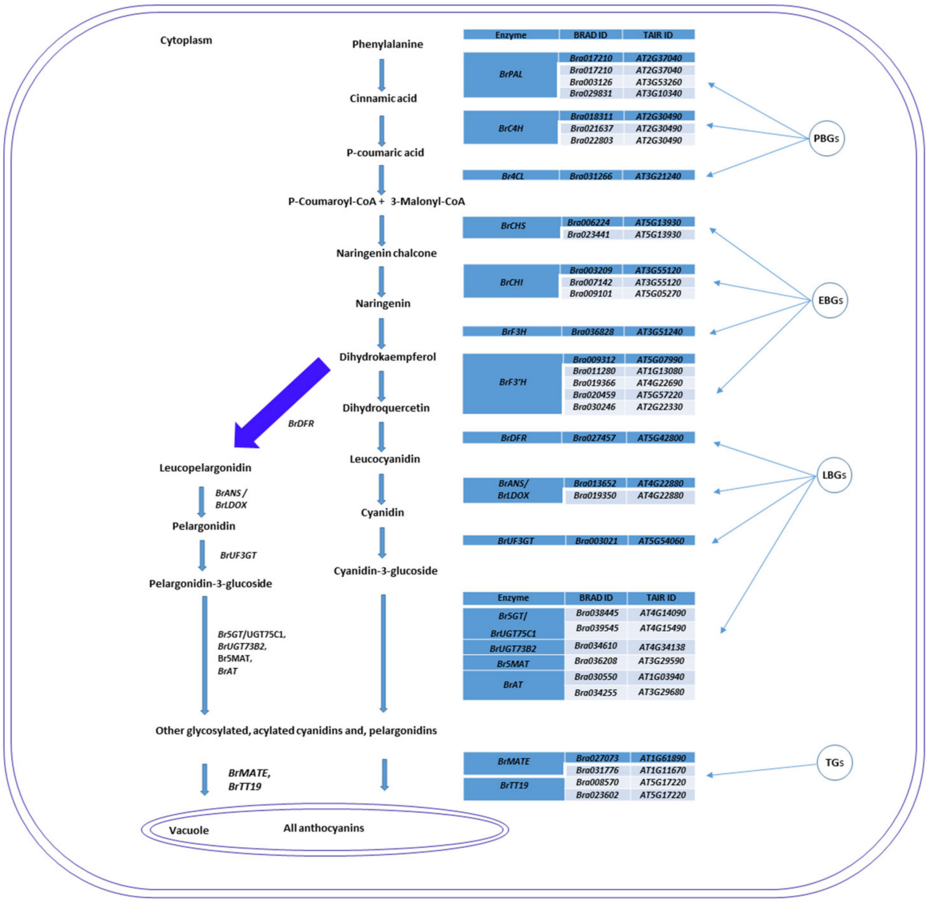


Figure 10. Chinese cabbage at young and mature stages. (**a**). Green (GCC) and (**b**). reddish purple (RPCC).



Enzyme	BRAD ID	TAIR ID
BrPAL	Bra017210	AT2G37040
	Bra017210	AT2G37040
	Bra003126	AT3G53260
	Bra029831	AT3G10340
BrC4H	Bra018311	AT2G30490
	Bra021637	AT2G30490
	Bra022803	AT2G30490
Br4CL	Bra031266	AT3G21240
BrCHS	Bra006224	AT5G13930
	Bra023441	AT5G13930
BrCHI	Bra003209	AT3G55120
	Bra007142	AT3G55120
	Bra009101	AT5G05270
BrF3H	Bra036828	AT3G51240
BrF3'H	Bra009312	AT5G07990
	Bra011280	AT1G13080
	Bra019366	AT4G22690
	Bra020459	AT5G57220
	Bra030246	AT2G22330
BrDFR	Bra027457	AT5G42800
BrANS/ BrLDOX	Bra013652	AT4G22880
	Bra019350	AT4G22880
BrUF3GT	Bra003021	AT5G54060

Enzyme	BRAD ID	TAIR ID
Br5GT/ BrUGT75C1	Bra038445	AT4G14090
	Bra039545	AT4G15490
BrUGT73B2	Bra034610	AT4G34138
Br5MAT	Bra036208	AT3G29590
BrAT	Bra030550	AT1G03940
	Bra034255	AT3G29680
BrMATE	Bra027073	AT1G61890
	Bra031776	AT1G11670
BrTT19	Bra008570	AT5G17220
	Bra023602	AT5G17220

Figure 11. Schematic representation of anthocyanin biosynthetic pathway based on anthocyanin pigments and important anthocyanin biosynthetic genes (ABGs) identified from transcriptome data.

3.2. Anthocyanin Biosynthesis Genes Are Differentially Regulated in RPCC

Transcriptome sequencing is an advanced NGS technique that can be used to predict novel genes, gene function, and genome evolution. Comparative transcriptome sequencing between two different phenotypes, GCC and RPCC, revealed differences in the expression levels of genes involved in anthocyanin biosynthesis and the regulation of this pathway [26,44]. The transcriptome sequencing of RPCC and GCC at the seedling stage revealed 3345 DEGs, which included unique genes with unknown functions and TF genes. About 255 DEGs were involved in various functions related to phenylpropanoids, lignins, flavonols, and anthocyanins. Further qRT-PCR analyses showed that PBGs and EBGs were expressed at levels 0.5- to 1.0-fold higher at the seedling stage than at the other two stages (rosette and heading) in RPCC. Many of these genes encoded proteins involved in the early phases of anthocyanin biosynthesis [7,45,46]. Our results indicate that *BrPAL, BrPAL2, BrPAL4, BrC4H, Br4CL2, BrCHS, BrCHI, BrCHI1, BrF3H*, and *BrF3′H-1* may be involved in the early phase of anthocyanin biosynthesis (i.e., the conversion of phenylalanine to dihydroquercetin) in RPCC. As shown in Figure 5e, the important LBGs *BrDFR* and *BrLDOX*, whose encoded products catalyze the conversion of dihydroquercetin to cyanidin at the late stage of anthocyanin biosynthesis, were expressed in RPCC at all stages, but not in GCC at any stage. Previous studies have shown that DFR plays crucial roles in anthocyanin accumulation in many plant species under different abiotic stress conditions [47,48]. Similarly, in *Arabidopsis,* sucrose and jasmonic acid have been shown to induce LBGs such as *DFR, LDOX*, and *UF3GT*, leading to anthocyanin accumulation [45,49]. Interestingly, we detected high transcript levels of the LBGs *BrDFR, BrLDOX*, and *BrUF3GT* in RPCC but not in GCC, indicating that anthocyanin biosynthesis occurs in RPCC under normal growth conditions without induction by external factors such as sugars or hormones.

Some downstream LBGs are involved in p-coumaroylation (*At3AT1*: At1g03940), glucosylation (*UGT75C1*:At4g14090) and malonylation (*At5MAT*: At3g29590) [33,50–52]; the orthologs of these genes (*BrAT-1, BrUGT75C1-1*, and *Br5MAT*) were expressed only in RPCC. Accordingly, the HPLC analyses detected p-coumaroyl (*At3AT1*: At1g03940) diglucoside (*UGT75C1*: At4g14090), indicating that anthocyanins have been modified as reported previously [33]. Our qRT-PCR analyses showed that the transcript levels of *BrUGT73B2* and *BrAT-2*, whose encoded proteins catalyze p-coumaroylation and glucosylation, respectively, were much higher in RPCC than in GCC at the rosette and heading stages. TT19 encodes a transporter involved in the movement and accumulation of anthocyanins in the Brassicaceae [53,54]. In this study, two TT19 paralogs, *BrTT19-1* and *BrTT19-2,* which have a common ortholog in *A. thaliana* (*AtTT19*: AT5G17220), showed very high transcript levels in RPCC, indicating that transport of anthocyanins from the cytosol to the vacuole is an important process in RPCC.

3.3. Differential Expression of MYBs Regulates Reddish Purple Color Accumulation

In general, ABGs are controlled by various TFs, including the MYB, bHLH, and WD40 TFs that make up the most well-known complex in plants, the MBW complex [55,56]. A study using the transcriptome approach in red Chinese cabbage identified that the anthocyanin pathway regulating genes are TT8 (*Bra037887*) and PAP1 (*c3563g1i2*) [3]. Most TFs that are known to regulate anthocyanins are MYB TFs. A study on grapes reported that the MYBA transcription factor regulates the anthocyanin biosynthesis pathway through controlling the expression of *UFGT* [57]. Because TFs, especially MYB, are known to be important for controlling expression of ABGs, we searched our transcription data for TF sequences. As a result, we identified 25 TF genes with log2fold change > 3: 12 TF genes with log2fold change > 4, five TF genes with log2fold change > 5 and > 6, and one TF gene with log2fold change > 10 (Table 3). They included two important MYB genes: *MYB90* (log2fold change, 11.3) and *MYB75* (log2fold change, 6.8), which are homologs of *PRODUCTION OF ANTHOCYANIN PIGMENT 2 (PAP2)* and *PRODUCTION OF ANTHOCYANIN PIGMENT 1 (PAP1)*. These genes showed similar transcript levels in qRT-PCR and transcriptome analyses, and their high expression levels in RPCC suggested that they might be involved in the positive regulation of LBGs during anthocyanin biosynthesis [13,58]. We identified duplicate copies of *MYBL2, Bra016164 (BrMYBL2-1)* and *Bra007957 (BrMYBL2-2)*, in *B.*

rapa, which are orthologs of *A. thaliana AtMYBL2* (*AT1G71030*). The transcript levels of both *BrMYBL2-1* and *BrMYBL2-2* were very high, as revealed by transcriptome analysis (log2fold change of 5.5 and 4.06, respectively) and qRT-PCR analysis. A previous study detected a similar expression pattern of *MYBL2* in *B. rapa,* indicative of its role as a positive regulator [59]. However, other studies on the Brassicaceae have demonstrated that MYBL2 can function as a negative regulator of anthocyanin biosynthesis [60,61]. Functional characterization of *BrMYBL2-1* and *Bra007957* (*BrMYBL2-2*) will clarify the molecular mechanisms of these genes in RPCC. Our results also showed that MYB binding motifs are highly conserved not only in LBGs but in most of the analyzed ABGs.

4. Materials and Methods

4.1. Plant Material and Sample Collection

The RPCC was developed (through introgression hybridization) and registered by the Kwonnong Seed Company (Cheongju, S. Korea) as described by Lee et al. [4]. The lines used in this study, green Chinese cabbage (GCC) and RPCC, belong to *B. rapa* L. ssp. *pekinensis* and were selected on the basis of their distinct color phenotypes (Figure 1). Seeds of GCC and RPCC were germinated in a growth chamber under a 16-h light/8-h dark photoperiod at 24 °C. Leaf samples (innermost and outermost leaves) were collected from three biological replicates at three growth stages: the seedling, rosette, and heading stages (at 3-, 6-, and 9-weeks-old, respectively). The leaf samples were stored at −70 °C until further analysis.

4.2. Anthocyanin Extraction and HPLC Analysis

The total anthocyanin content of freeze-dried outer and inner leaf tissues of 9-week-old GCC and RPCC was determined by HPLC and LC-MS/MS, as described previously [37]. Each 100-mg lyophilized leaf sample was mixed with 2 mL water:formic acid (95:5 *v*/*v*) followed by 5 min vortexing and 20 min sonication. The sample was centrifuged at 9200× *g* for 15 min at 4 °C and the supernatant was filtered through a 0.45-µm PTFE hydrophilic syringe filter. From this filtrate, 10 µL was used for estimating anthocyanin content. The sample was injected into an Agilent 1200 series HPLC connected to a 4000 Q-Trap LC-ECI-MS/MS system. A Synergy 4µL Polar-RP 80A column (250 × 4.6 mm i.d., particle size 4 µm; Phenomenex, Torrance, CA, USA) with a Security Guard Cartridge (AQ C18, 4 × 3.0 mm KJO-4282; Phenomenex, Torrance, CA, USA) were used. Anthocyanin pigments were detected at 520 nm. The oven temperature was set to 40 °C. The composition of the mobile phase was as follows: solvent A: water:formic acid (95:5 *v*/*v*), and solvent B (acetonitrile:formic acid, 95:5 *v*/*v*). The gradient conditions were as follows: 0–8 min, 8–13 min, 13% solvent B; 13–20 min, 20–23 min, 17% solvent B; 23–30 min, 30–40 min, 20% solvent B; 40–40.1 min, 5% solvent B; and 40.1–50 min, 5% solvent B. The anthocyanin concentration in each sample was measured by comparison of the area of each peak with that of the external standard (cyanidin-3-*O*-glucoside) on the HPLC chromatogram. Mean ± SD values were calculated from the three replicates of each sample.

4.3. Sequence Pre-Processing and Assembly

For transcriptome analysis, total RNA was extracted from leaf tissues of 3-week-old plants using an RNeasy Mini kit (Qiagen, Valencia, CA, USA) and sequenced with the Illumina Hi-seq2000 platform by SEEDERS Inc. (Korea). The sequence data was submitted to the NCBI database and are available under the accession number PRJNA612946. The raw reads were trimmed using the Dynamic-Trim and Length-Sort programs of the Solexa QA [62] package. Based on the Dynamic-Trim (phred score ≥ 20) and Length-Sort (short read length ≥ 25bp) parameters, clean reads were obtained. The clean reads were assembled according to the protocols of Velvet (version 1.2.08) and Oases (version 0.2.08) software [63]. The optimal k-mer was selected based on the max length, average length, and N50 according to the total length of the assembled sequence.

4.4. Mapping and Annotation of Transcripts

To identify gene function, the transcripts were used as queries in BlastX searches against the amino acid sequences in the BRAD [64] and KEGG databases with the following parameters: filter criterion: e-value ≤ 1e−10, best hits. Mapping was performed using Bowtie2 (v2.1.0) software with the following limitation: mismatch ≤ 2 bp, computed by the penalty method) [65]. Transcript levels were normalized using the R package of DESeq [66], and this software was also used to calculate the gene expression values for each sample with data deviation.

4.5. Identification of Transcription Factors

To identify TFs, sequences of all *B. rapa* 4127 TFs were downloaded from the Plant Transcription Factor Database [67] (http://planttfdb.cbi.pku.edu.cn/). The total assembled transcripts were compared and analyzed using BlastX software with the parameters e-value ≤ 1e−50 and identity ≥ 50, and annotated.

4.6. Prediction of Differentially Expressed Genes

Differentially expressed genes (DEGs) were defined as those with at least a log2fold difference in transcript levels between the RPCC and GCC samples. Up-regulated genes were those with log2 fold-change greater than 1, and down-regulated genes were those with log2 fold-change less than −1 [22].

4.7. Expression Analyses

Leaf samples collected from 3-, 6-, and 9-week-old plants were collected and immediately frozen in liquid nitrogen. Total RNA was extracted using an RNeasy Mini kit (Qiagen). cDNA was synthesized using a Reverse Transcription System kit (Promega, Madison, WI, USA). The resulting cDNA was used as a template for qRT-PCR analyses, which were conducted with the CFX96 Real-Time system (Bio-Rad, Hercules, CA, USA). qRT-PCR analyses were conducted for 60 anthocyanin biosynthetic genes (ABGs) with three biological replicates and three technical replicates using the following conditions: 95 °C for 3 min; 39 cycles of 95 °C for 15 s and 58 °C for 20 s. The relative expression levels were determined by normalizing the data with the comparative Ct method $2-[\Delta\Delta Ct]$ [68] using the *Actin* gene as a reference.

4.8. Gene Ontology Annotation and KEGG Enrichment Analyses

Gene ontology (GO) [69] alignments were performed using total transcripts and 'GO DB' with the threshold of 'counts ≥ 1' and 'GO depth' set to 3. Genes were classified into three functional categories, BP (Biological Process), CC (Cellular Component), and MF (Molecular Function). Then, gene annotation (filter criterion: e-value ≤ 1e−10, best hits) was performed through comparisons with amino acid sequences in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database using BLASTX.

To predict the functional characteristics of up- and down-regulated genes, we first computed *p*-values based on Fisher's exact test; these values were taken as indicators of significance of the gene in the respective function. Further KEGG enrichment analysis was performed using the KOBAS online tool (http://kobas.cbi.pku.edu.cn/kobas3) [70]. These analyses provided pathway annotations for the transcripts.

4.9. Identification of Cis-Regulatory Elements

The 2-kb region upstream from the transcription start site of selected genes was screened to identify cis-regulatory motifs using the New PLACE web tool [71].

4.10. Network Analysis

We carried out network analysis to construct an active gene-to-gene regulatory network of genes positively correlated with the anthocyanin biosynthetic pathway. The STRING 10.0 database (http://string-db.org/) was used to obtain an interaction network of these genes in *B. rapa* [72], according to orthologous genes in *A. thaliana*. Every link has a score from 0 to 1, where 1 is considered as the highest confidence link for reconstruction [73].

5. Conclusions

In conclusion, 11 of the 13 pigments detected in RPCC are reported for the first time for this variety. Analyses of the transcriptome data from two varieties at the seedling stage revealed many unique transcripts including DEGs and TF genes that are involved in a multitude of functions in growth and development. Our results show that many DEGs between the red and green varieties are involved in the biosynthesis of secondary metabolites such as phenylpropanoids, lignins, flavonoids, and anthocyanins, and in the regulation of these biosynthetic pathways. Further qRT-PCR expression analyses confirmed that ABGs and many TFs play essential roles in anthocyanin biosynthesis. The gene-to-gene interaction network illustrates the possible regulatory mechanism of MYBs with ABGs during anthocyanin biosynthesis in RPCC. Overall, our study describes the pigments in RPCC, identifies the important anthocyanin biosynthetic genes and TF genes that control the anthocyanin biosynthesis pathway, and proposes a model for the possible interaction mechanism between ABGs and TFs.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/21/8/2901/s1. References [74–85] are cited in Supplementary Materials. Figure S1. Statistics of non-redundant transcripts predicted in transcriptome data. (a) Length distribution of transcripts. (b) Transcripts annotated in public databases. Figure S2 MA plot of DEGs between green (G) vs. reddish purple (RP) Chinese cabbage identified from transcriptome data. X-axis shows log2fold change and y-axis shows differentially expressed genes. Red and green dots indicate up-regulated and down-regulated genes, respectively. Figure S3. Additional ABGs showing differential expression in transcriptome data validated by qRT-PCR. Expression values were normalized against that of *Actin*. Error bars are based on mean of three technical replicates. Table S1. Total anthocyanin pigments identified in reddish purple Chinese cabbage leaf samples. Table S2. Total up-regulated DEG transcripts between RPCC and GCC with annotations from BRAD and TAIR databases. Table S3. Total down-regulated DEG transcripts between RPCC and GCC with annotations from BRAD and TAIR databases. Table S4. Transcription factor genes up-regulated between RPCC and GCC samples, as identified from transcriptome data. Table S5. Transcription factor genes down-regulated between RPCC and GCC, as identified from transcriptome data. Table S6. Gene ontology (GO) annotation of DEG transcripts predicted from transcriptome data. GO annotations are shown for up-regulated genes (left) and down-regulated genes (right). Table S7. KEGG pathway annotation of DEG transcripts predicted in transcriptome data. KEGG annotations are shown for up-regulated genes (left) and down-regulated genes (right). Table S8. Genes identified from transcriptome data that are involved in phenylpropanoid and anthocyanin pathways with expression values and GO and KEGG annotations. Table S9. List of primer sequences used for qRT-PCR of anthocyanin biosynthetic genes. Table S10. *Cis*-regulatory elements (CREs) identified in upstream promoter regions of anthocyanin biosynthetic genes.

Author Contributions: Conceptualization, J.J.R., S.R.C., S.S.C., S.Y.Y., M.-H.O., S.U.P., S.-J.K., J.L., O.H.K. and Y.P.L.; Data curation, J.J.R., S.H.O.; Formal analysis, J.J.R., S.S.C., S.Y.Y., S.H.O., H.K., C.Y.L., S.U.P., K.S.-J.; Funding acquisition, S.R.C. and YPL; Investigation, J.J.R., S.Y.Y., H.K., C.Y.L., J.L. and O.H.K.; Methodology, J.J.R., S.S.C., S.H.O., H.K., C.Y.L., J.L. and O.H.K.; Project administration, S.R.C. and Y.P.L.; Resources, S.R.C., J.L., O.H.K. and Y.P.L.; Software, J.J.R. and M.-S.K.; Supervision, S.R.C. and Y.P.L.; Validation, J.J.R. and S.S.; Writing–original draft, J.J.R.; Writing–review & editing, J.J.R., S.R.C., S.S.C., S.Y.Y., M.-S.K.,S.S., M.-H.O., S.U.P., S.-J. K., J.L., O.H.K. and Y.P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, and Forestry (IPET) through Golden Seed Project (213006-05-4-SBD30, 213006-05-4-SB110), funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA), Ministry of Oceans and Fisheries (MOF), Rural Development Administration (RDA) and Korea Forest Services (KFS).

Acknowledgments: We thank Jennifer Smith, from Liwen Bianji, Edanz Group China (www.liwen bianji.cn/ac), for editing the English text of a draft of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ABGs	Anthocyanin biosynthetic genes
ABP	Anthocyanin biosynthetic pathway
AP2/ERF	Apetala 2/ethylene response factor
AT	Acyltransferases

bHLH	Basic helix-loop-helix
DEGs	Differentially expressed genes
DFR	Dihydroflavonol-4-reductase
EBGs	Early biosynthetic genes
GCC	Green Chinese cabbage
HPLC	High performance liquid chromatography
LBGs	Late biosynthetic genes
LDOX	Leucoanthocyanidin dioxygenases
MATE	Multidrug and toxic compound extrusion
MYB	Myeloblastosis
5MAT	5-O-glucoside-6-O-malonyltransferase
PBGs	Phenylpropanoid biosynthetic genes
qRT-PCR	*Quantitative* real-time polymerase chain reaction
RPCC	Reddish purple Chinese cabbage
RPKM	Reads per kilo base per million mapped reads
TT19	Transparent Testa 19
UF3GT	UDP- glucose: flavonoid 3-*O*-glucosyltransferase
UGT75c1	UDP-glucosyltransferase 75c1

References

1. Katche, E.; Quezada-Martinez, D.; Katche, E.I.; Vasquez-Teuber, P.; Mason, A.S. Interspecific hybridization for Brassica crop improvement Crop Breeding. *Genet. Genom.* **2019**, *1*, e190007.
2. Wang, G.-X.; Lv, J.; Zhang, J.; Han, S.; Zong, M.; Guo, N.; Zeng, X.-Y.; Zhang, Y.-Y.; Wang, Y.-P.; Liu, F. Genetic and epigenetic alterations of *Brassica nigra* introgression lines from somatic hybridization: A resource for cauliflower improvement. *Front. Plant Sci.* **2016**, *7*, 1258. [CrossRef]
3. Xie, L.; Li, F.; Zhang, S.; Zhang, H.; Qian, W.; Li, P.; Zhang, S.; Sun, R. Mining for candidate genes in an introgression line by using RNA sequencing: The anthocyanin over accumulation phenotype in Brassica. *Front. Plant Sci.* **2016**, *7*, 1245. [CrossRef]
4. Lee, H.; Oh, I.-N.; Kim, J.; Jung, D.; Cuong, N.P.; Kim, Y.; Lee, J.C.; Kwon, O.; Park, S.U.; Lim, Y.P.; et al. Phenolic compound profiles and their seasonal variations in new red-phenotype head-forming Chinese cabbages. *Lwt-Food Sci. Tech.* **2018**, *90*, 433–439. [CrossRef]
5. Vogt, T. Phenylpropanoid biosynthesis. *Mol. Plant* **2010**, *3*, 2–20. [CrossRef] [PubMed]
6. Tanaka, Y.; Sasaki, N.; Ohmiya, A. Biosynthesis of plant pigments: Anthocyanins, betalains and carotenoids. *Plant J.* **2008**, *54*, 733–749. [CrossRef]
7. Liu, Y.; Tikunov, Y.; Schouten, R.E.; Marcelis, L.F.M.; Visser, R.G.F.; Bovy, A. anthocyanin biosynthesis and degradation mechanisms in solanaceous vegetables: A review. *Front. Chem.* **2018**, *6*, 52. [CrossRef] [PubMed]
8. Petroni, K.; Tonelli, C. Recent advances on the regulation of anthocyanin synthesis in reproductive organs. *Plant Sci.* **2011**, *181*, 219–229. [CrossRef] [PubMed]
9. Sablowski, R.W.M.; Moyano, E.; Cullanez-Macia, F.A.; Schuch, W.; Martin, C.; Bevan, M. A flower specific Myb protein activates transcription of phenylpropanoid biosynthetic genes. *EMBO J.* **1994**, *13*, 128–137. [CrossRef] [PubMed]
10. Lloyd, A.; Brockman, A.; Aguirre, L.; Campbell, A.; Bean, A.; Cantero, A.; Gonzalez, A. Advances in the MYB–bHLH–WD repeat (MBW) pigment regulatory model: Addition of a WRKY factor and co-option of an anthocyanin MYB for betalain regulation. *Plant Cell Physiol.* **2017**, *58*, 1431–1441. [CrossRef] [PubMed]
11. Stracke, R.; Werber, M.; Weisshaar, B. The R2R3-MYB gene family in *Arabidopsis thaliana*. *Curr. Opin. Plant Biol.* **2001**, *4*, 447–456. [CrossRef]
12. Dubos, C.; Stracke, R.; Grotewold, E.; Weisshaar, B.; Martin, C.; Lepiniec, L. MYB transcription factors in *Arabidopsis*. *Trends Plant Sci.* **2010**, *15*, 573–581. [CrossRef] [PubMed]
13. Gonzalez, A.; Zhao, M.; Leavitt, J.M.; Lloyd, A.M. Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in *Arabidopsis* seedlings. *Plant J.* **2008**, *53*, 814–827. [CrossRef] [PubMed]

14. Xie, X.; Xie, X.B.; Li, S.; Zhang, R.F.; Zhao, J.; Chen, Y.C.; Zhao, Q.; Yao, Y.X.; You, C.X.; Zhang, X.S.; et al. The bHLH transcription factor MdbHLH3 promotes anthocyanin accumulation and fruit colouration in response to low temperature in apples. *Plant Cell Env.* **2012**, *35*, 1884–1897. [CrossRef]
15. Verweij, W.; Spelt, C.E.; Bliek, M.; de Vries, M.; Wit, N.; Faraco, M. Functionally similar WRKY proteins regulate vacuolar acidification in *Petunia* and hair development in *Arabidopsis*. *Plant Cell* **2016**, *28*, 786–803. [CrossRef] [PubMed]
16. An, J.; Qu, F.; Yao, J.; Wang, X.N.; You, C.X.; Wang, X.F.; Hao, Y.J. The bZIP transcription factor MdHY5 regulates anthocyanin accumulation and nitrate assimilation in apple. *Hortic. Res.* **2017**, *4*, 1702. [CrossRef] [PubMed]
17. An, J.P.; Yao, J.F.; Xu, R.R.; You, C.X.; Wang, X.F.; Hao, Y.J. Apple bZIP transcription factor MdbZIP44 regulates abscisic acid-promoted anthocyanin accumulation. *Plant Cell Env.* **2018**, *41*, 2678–2692. [CrossRef]
18. Guo, N.; Wu, J.; Zheng, S.; Cheng, F.; Liu, B.; Liang, J.; Cui, Y.; Wang, X. Anthocyanin profile characterization and quantitative trait locus mapping in zicaitai (*Brassica rapa* L. ssp.chinensis var. purpurea). *Mol. Breed.* **2015**, *35*, 113. [CrossRef]
19. He, Q.; Zhang, Z.; Zhang, L. Anthocyanin accumulation, antioxidant ability and stability, and a transcriptional analysis of anthocyanin biosynthesis in purple heading Chinese cabbage (*Brassica rapa L. ssp. pekinensis*). *J. Agricul. Food Chem.* **2016**, *64*, 132–145. [CrossRef]
20. Oh, Y.S.; Lee, J.H.; Yoon, S.H.; Oh, C.H.; Choi, D.S.; Choe, E.; Jung, M.Y. Characterization and quantification of anthocyanins in grape juices obtained from the grapes cultivated in Korea by HPLC/DAD, HPLC/MS, and HPLC/MS/MS. *J. Food Sci.* **2008**, *73*, C378–C389. [CrossRef]
21. Moreno, D.A.; Perez-Balibrea, S.; Ferreres, F.; Gil-Izquierdo, A.; Garcia-Viguera, C. Acylated anthocyanins in Broccoli sprouts. *Food Chem.* **2010**, *123*, 358–363. [CrossRef]
22. Jeon, J.; Kim, J.K.; Kim, H.; Kim, Y.J.; Park, Y.J.; Kim, S.J.; Kim, C.; Park, S.U. Transcriptome analysis and metabolic profiling of green and red kale (*Brassica oleracea* var. *acephala) seedlings. Food Chem.* **2018**, *241*, 7–13. [CrossRef] [PubMed]
23. Jeon, J.; Lim, C.J.; Kim, J.K.; Park, S.U. Comparative metabolic profiling of green and purple pakchoi (*Brassica rapa* Subsp. *Chinensis*). *Molecules* **2018**, *23*, 1613. [CrossRef] [PubMed]
24. Tong, C.; Wang, X.; Yu, J.; Wu, J.; Li, W.; Huang, J.; Dong, C.; Hua, W.; Liu, S. Comprehensive analysis of RNA-seq data reveals the complexity of the transcriptome in *Brassica rapa*. *BMC Genom.* **2013**, *14*, 689. [CrossRef] [PubMed]
25. Wang, S.; Zhou, G.; Huang, X.; Hu, J.; Wang, B.; Lin, C.; Li, X.; Jia, Y.; Wang, A. Transcriptome analysis of non-heading Chinese cabbage under heat stress by RNA-seq and marker identification. *Euphytica* **2017**, *213*, 109. [CrossRef]
26. Zhang, L.; Xu, B.; Wu, T.; Yang, Y.; Fan, L.; Wen, M.; Sui, J. Transcriptomic profiling of two pak choi varieties with contrasting anthocyanin contents provides an insight into structural and regulatory genes in anthocyanin biosynthetic pathway. *BMC Genom.* **2017**, *18*, 288. [CrossRef]
27. Eom, S.H.; Baek, S.-A.; Kim, J.K.; Hyun, T.K. Transcriptome analysis in Chinese cabbage (Brassica *rapa* ssp pekinensis) provides the role of glucosinolate metabolism in response to drought stress. *Molecules* **2018**, *23*, 1186.
28. Joo, H.K.; Choi, S.; Lee, Y.R.; Lee, E.O.; Park, M.S.; Park, K.B.; Kim, C.S.; Lim, Y.P.; Park, J.T.; Jeon, B.H. Anthocyanin-rich extract from red Chinese cabbage alleviates vascular inflammation in endothelial cells and Apo E–/–mice. *Int. J. Mol. Sci.* **2018**, *19*, 816. [CrossRef]
29. Ren, J.; Liu, Z.; Niu, R.; Feng, H. Mapping of Re, a gene conferring the red leaf trait in ornamental kale (*Brassica oleracea* L. var. acephala). *Plant Breed.* **2015**, *134*, 494–500. [CrossRef]
30. Outchkourov, N.; Karlova, R.; Hoelscher, M.; Schrama, X.; Blilou, I.; Jongedijk, E.; Simon, C.D.; van Dijk, A.D.J.; Bosch, D.; Hall, R.D.; et al. Transcription factor-mediated control of anthocyanin biosynthesis in vegetative tissues. *Plant Physiol.* **2018**, *176*, 1862–1878. [CrossRef]
31. Bu, C.; Zhang, Q.; Zeng, J.; Cao, X.; Hao, Z.; Qiao, D.; Cao, Y.; Xu, H. Identification of a novel anthocyanin synthesis pathway in the fungus Aspergillus sydowii H-1. *BMC Genom.* **2020**, *21*, 29. [CrossRef] [PubMed]
32. Cheng, J.; Wei, G.; Zhou, H.; Gu, C.; Vimolmangkang, S.; Liao, L.; Han, Y. Unraveling the mechanism underlying the glycosylation and methylation of anthocyanins in peach. *Plant Physiol.* **2014**, *166*, 1044–1058. [CrossRef] [PubMed]

33. Sasaki, N.; Nishizaki, Y.; Ozeki, Y.; Miyahara, T. The role of acyl-glucose in anthocyanin modifications. *Molecules* **2014**, *19*, 18747–18766. [CrossRef] [PubMed]
34. Hartmann, U.; Sagasser, M.; Mehrtens, F.; Stracke, R.; Weisshaar, B. Differential combinatorial interactions of *cis*-acting elements recognized by R2R3-MYB, bZIP, and bHLH factors control light-responsive and tissue-specific activation of phenylpropanoid biosynthesis genes. *Plant Mol. Biol.* **2005**, *57*, 155–171. [CrossRef]
35. Yu, M.; Man, Y.; Wang, Y. Light- and temperature-induced expression of an R2R3-MYB gene regulates anthocyanin biosynthesis in red-fleshed kiwifruit. *Int. J. Mol. Sci.* **2019**, *20*, 5228. [CrossRef]
36. Castañeda-Ovando, A.; Pacheco-Hernández, M.L.; Páez-Hernández, E.; Rodríguez, J.A.; Galán-Vidal, C.A. Chemical studies of anthocyanins: A review. *Food Chem.* **2009**, *113*, 859–871. [CrossRef]
37. Park, C.H.; Yeo, H.J.; Kim, N.S.; Eun, P.Y.; Kim, S.-J.; Arasu, M.V.; Al-Dhabi, N.A.; Park, S.Y.; Kim, J.K.; Park, S.U. Metabolic profiling of pale green and purple kohlrabi (*Brassica oleracea* var. gongylodes). *Appl. Biol. Chem.* **2017**, *60*, 249–257. [CrossRef]
38. Wu, X.; Prior, R.L. Identification and characterization of anthocyanins by High-performance liquid chromatography–electrospray ionization–tandem mass spectrometry in common foods in the United States: Vegetables, nuts, and grains. *J. Agric. Food Chem.* **2005**, *53*, 3101–3113. [CrossRef]
39. Charron, C.S.; Clevidence, B.A.; Britz, S.J.; Novotny, J.A. Effect of dose size on bioavailability of acylated and non-acylated anthocyanins from red cabbage (*Brassica oleracea* L. var. Capitata). *J. Agric. Food Chem.* **2007**, *55*, 5354–5362. [CrossRef]
40. Park, N.I.; Xu, H.; Li, X.; Jang, I.H.; Park, S.; Ahn, G.H.; Lim, Y.P.; Kim, S.J.; Park, S.U. Anthocyanin accumulation and expression of anthocyanin biosynthetic genes in radish (*Raphanus sativus*). *J. Agric. Food Chem.* **2011**, *59*, 6034–6039. [CrossRef]
41. Sun, J.; Xiao, Z.; Lin, L.-Z.; Lester, G.E.; Wang, Q.; Harnly, J.M.; Chen, P. Profiling Polyphenols in Five Brassica Species Microgreens by UHPLC-PDA-ESI/HRMSn. *J. Agric. Food Chem.* **2013**, *61*, 10960–10970. [CrossRef] [PubMed]
42. Wu, J.; Liu, W.; Yuan, L.; Guan, W.Q.; Brennan, C.S.; Zhang, Y.Y.; Zhang, J.; Wang, Z.D. The influence of postharvest UV-C treatment on anthocyanin biosynthesis in fresh-cut red cabbage. *Sci. Rep.* **2017**, *7*, 5232. [CrossRef] [PubMed]
43. Goswami, G.; Nath, U.K.; Park, J.-I.; Hossain, M.R.; Biswas, M.K.; Kim, H.T.; Kim, H.R.; Nou, I.S. Transcriptional regulation of anthocyanin biosynthesis in a high-anthocyanin resynthesized Brassica napus cultivar. *J. Biol. Res.* **2018**, *25*, 19.
44. Yuan, Y.; Chiu, L.-W.; Li, L. Transcriptional regulation of anthocyanin biosynthesis in red cabbage. *Planta* **2009**, *230*, 1141–1153. [CrossRef] [PubMed]
45. Solfanelli, C.; Poggi, A.; Loreti, E.; Alpi, A.; Perata, P. Sucrose-specific induction of the anthocyanin biosynthetic path-way in *Arabidopsis*. *Plant Physiol.* **2006**, *140*, 637–646. [CrossRef]
46. Guo, N.; Han, S.; Zong, M.; Wang, G.; Zheng, S.; Liu, F. Identification and differential expression analysis of anthocyanin biosynthetic genes in leaf color variants of ornamental kale. *BMC Genom.* **2019**, *20*, 564. [CrossRef]
47. Wang, H.; Fan, W.; Li, H.; Yang, J.; Huang, J.; Zhang, P. Functional characterization of dihydroflavonol-4-reductase in anthocyanin biosynthesis of purple sweet potato underlies the direct evidence of anthocyanins function against abiotic stresses. *PLoS ONE* **2013**, *8*, e78484. [CrossRef]
48. Ahmed, N.U.; Park, J.I.; Jung, H.J.; Yang, T.J.; Hur, Y.; Nou, I.S. Characterization of dihydroflavonol 4-reductase (DFR) genes and their association with cold and freezing stress in *Brassica Rapa*. *Gene* **2014**, *550*, 46–55. [CrossRef]
49. Li, T.; Jia, K.P.; Lian, H.L.; Yang, X.; Li, L.; Yang, H.Q. Jasmonic acid enhancement of anthocyanin accumulation is de-pendent on phytochrome A signaling pathway under far-red light in *Arabidopsis*. *Biochem. Biophys. Res. Commun.* **2014**, *454*, 78–83. [CrossRef]
50. Tohge, T.; Nishiyama, Y.; Hirai, M.Y.; Yano, M.; Nakajima, J.; Awazuhara, M.; Inoue, E.; Takahashi, H.; Goodenowe, D.B.; Kitayama, M.; et al. Functional genomics by integrated analysis of metabolome and transcriptome of *Arabidopsis* plants over-expressing an MYB transcription factor. *Plant J.* **2005**, *42*, 218–235. [CrossRef]

51. Luo, J.; Nishiyama, Y.; Fuell, C.; Taguchi, G.; Elliott, K.; Hill, L.; Tanaka, Y.; Kitayama, M.; Yamazaki, M.; Bailey, P.; et al. Convergent evolution in the BAHD family of acyl transferases: Identification and characterization of anthocyanin acyl transferases from Arabidopsis thaliana. *Plant J.* **2007**, *50*, 678–695.
52. D'Auria, J.C.; Reichelt, M.; Luck, K.; Svatos, A.; Gershenzon, J. Identification and characterization of the BAHD acyltransferase malonyl CoA: Anthocyanidin 5-O-glucoside-6″-O-malonyltransferase (At5MAT) in *Arabidopsis*. *Thaliana*. *FEBS Lett.* **2007**, *581*, 872–878. [CrossRef] [PubMed]
53. Sun, Y.; Li, H.; Huang, J.-R. Arabidopsis TT19 functions as a carrier to transport anthocyanin from the cytosol to tonoplasts. *Mol. Plant* **2012**, *5*, 387–400. [CrossRef] [PubMed]
54. Wang, Y.; Tang, Y.; Zhang, M.; Cai, F.; Qin, J.; Wang, Q.; Liu, C.; Wang, G.; Xu, L.; Yang, L.; et al. Molecular cloning and functional characterization of a glutathione S-transferase involved in both anthocyanin and proanthocyanidin accumulation in *Camelina sativa* (Brassicaceae). *Genet. Mol. Res.* **2012**, *11*, 4711–4719. [CrossRef]
55. Ramsay, N.A.; Glover, B.J. MYB-bHLH-WD40 protein complex and the evolution of cellular diversity. *Trends Plant Sci.* **2005**, *10*, 63–70. [CrossRef]
56. Zhang, B.; Chopra, D.; Schrader, A.; Hülskamp, M. Evolutionary comparison of competitive protein-complex formation of MYB, bHLH, and WDR proteins in plants. *J. Exp. Bot.* **2019**, *70*, 3197–3209. [CrossRef] [PubMed]
57. Kobayashi, S.; Ishimaru, M.; Hiraoka, K.; Honda, C. Myb-related genes of the Kyoho grape (*Vitis labruscana)* regulate anthocyanin biosynthesis. *Planta* **2002**, *215*, 924–933.
58. Borevitz, J.O.; Xia, Y.J.; Blount, J.; Dixon, R.A.; Lamb, C. Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. *Plant Cell* **2000**, *12*, 2383–2393. [CrossRef]
59. Mushtaq, M.A.; Pan, Q.; Chen, D.; Zhang, Q.; Ge, X.; Li, Z. Comparative leaves transcriptome analysis emphasizing on accumulation of anthocyanins in Brassica: Molecular regulation and potential interaction with photosynthesis. *Front. Plant Sci.* **2016**, *7*, 311. [CrossRef]
60. Matsui, K.; Umemura, Y.; Ohme-Takagi, M. AtMYBL2, a protein with a single MYB domain, acts as a negative regulator of anthocyanin biosynthesis in *Arabidopsis*. *Plant J.* **2008**, *55*, 954–967. [CrossRef]
61. Zhang, X.; Zhang, K.; Wu, J.; Guo, N.; Liang, J.; Wang, X.; Cheng, F. QTL-Seq and sequence assembly rapidly mapped the gene BrMYBL21 for the purple trait in Brassica rapa. *Sci. Rep.* **2020**, *10*, 2328. [PubMed]
62. Cox, M.P.; Peterson, D.A.; Biggs, P.J. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinform.* **2010**, *11*, 485. [CrossRef] [PubMed]
63. Schulz, M.H.; Zerbino, D.R.; Vingron, M.; Birney, E. Oases: Robust denovo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **2012**, *28*, 1086–1092. [CrossRef] [PubMed]
64. Wang, X.; Wang, H.; Wang, J.; Sun, R.; Wu, J.; Liu, S.; Bai, Y.; Mun, J.H.; Bancroft, I.; Cheng, F. The genome of the mesopolyploid crop species *Brassica Rapa*. *Nat. Genet.* **2011**, *28*, 1035–1039. [CrossRef]
65. Langmead, B.; Salizberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
66. Anders, S.; Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **2010**, *11*, R106. [CrossRef]
67. Jin, J.P.; Tian, F.; Yang, D.C.; Meng, Y.Q.; Kong, L.; Luo, J.C.; Gao, G. Plant TFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* **2017**, *45*, D1040–D1045. [CrossRef]
68. Livak, K.J.; Schmittgen, T.D. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods* **2001**, *25*, 402–408. [CrossRef]
69. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
70. Xie, C.; Mao, X.; Huang, J.; Ding, Y.; Wu, J.; Dong, S.; Kong, L.; Gao, G.; Li, C.Y.; Wei, L. KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* **2011**, *39*, W316–W322. [CrossRef]
71. Higo, K.; Ugawa, Y.; Iwamoto, M.; Korenaga, T. Plant *cis*-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* **1999**, *27*, 297–300. [CrossRef] [PubMed]
72. Szklarczyk, D.; Morris, J.H.; Cook, H.; Kuhn, M.; Wyder, S.; Simonovic, M.; Santos, A.; Doncheva, N.T.; Roth, A.; Bork, P.; et al. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **2017**, *45*, D362–D368. [CrossRef]

73. Kim, M.-S.; Hong, S.; Devaraj, S.P.; Im, S.; Kim, J.-R.; Lim, Y.P. Identification and characterization of the leaf specific networks of inner and rosette leaves in *Brassica rapa*. *Biochem. Biophys. Res. Commun.* **2017**, *490*, 821–826. [CrossRef] [PubMed]
74. Wong, D.C.J.; Gutierrez, R.L.; Gambetta, G.A.; Castellarin, S.D. Genome-wide analysis of cis-regulatory element structure and discovery of motif-driven gene co-expression networks in grapevine. *DNA Res.* **2017**, *24*, 311–326. [CrossRef] [PubMed]
75. Chakravarthy, S.; Tuori, R.P.; Dascenzo, M.D.; Fobert, P.R.; Despres, C.; Martin, G.B. The tomato transcription factor Pti4 regulates defence-related gene expression via GCC box and non-GCC box cis-elements. *Plant Cell* **2003**, *15*, 3033–3050. [CrossRef]
76. Abe, H.; Urao, T.; Ito, T.; Seki, M.; Shinozaki, K.; Shinozaki, K.-Y. *Arabidopsis* AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. *Plant Cell* **2003**, *15*, 63–78. [CrossRef]
77. Abe, H.; Shinozaki, K.-Y.; Urao, T.; Iwasaki, T.; Hosokawa, D.; Shinozaki, K. Role of *Arabidopsis* MYC and MYB homologs in drought- and abscisic acid-regulated gene expression. *Plant Cell* **1997**, *9*, 1859–1868.
78. Planchais, S.; Perennes, C.; Glab, N.; Mironov, V.; Inze, D.; Bergounioux, C. Characterization of *cis*-acting element involved in cell cycle phase-independent activation of Arath; CycB1;1 transcription and identification of putative regulatory proteins. *Plant Mol. Biol.* **2002**, *50*, 111–127. [CrossRef]
79. Gubler, F.; Kalla, R.; Roberts, J.K.; Jacobsen, J.V. Gibberellin-regulated expression of a myb gene in barley aleurone cells: Evidence for Myb transactivation of a high-pl alpha-amylase gene promoter. *Plant Cell* **1995**, *7*, 1879–1891. [CrossRef]
80. Tamagnone, L.; Merida, A.; Parr, A.; Mackay, S.; Culianez-Macia, F.A.; Roberts, K.; Martin, C. (1998). The AmMYB308 and AmMYB330 transcription factors from *Antirrhinum* regulate phenylpropanoid and lignin biosynthesis in transgenic tobacco. *Plant Cell* **1998**, *10*, 135–154. [CrossRef]
81. Baranowskij, N.; Frohberg, C.; Prat, S.; Willmitzer, L. A novel DNA binding protein with homology to Myb oncoproteins containing only one repeat can function as a transcriptional activator. *Embo J.* **1994**, *13*, 5383–5392. [CrossRef] [PubMed]
82. Yamamoto, S.; Nakano, T.; Suzuki, K.; Shinshi, H. Elicitor-induced activation of transcription via W box-related *cis*-acting elements from a basic chitinase gene by WRKY transcription factors in tobacco. *Biochem. Biophys. Acta* **2004**, *1679*, 279–287. [CrossRef] [PubMed]
83. Toyofuku, K.; Umemura, T.; Yamaguchi, J. Promoter elements required for sugar-repression of the RAmy3D gene for alpha-amylase in rice. *FEBS Lett.* **1998**, *428*, 275–280. [CrossRef]
84. Izawa, T.; Foster, R.; Nakajima, M.; Shimamoto, K.; Chua, N.-H. The rice bZIP transcriptional activator RITA-1 is highly expressed during seed development. *Plant Cell* **1994**, *6*, 1277–1287.
85. Brown, R.L.; Kazan, K.; McGrath, K.C.; Maclean, D.J.; Manners, J.M. A role for the GCC-box in jasmonate-mediated activation of the PDF1.2 gene of Arabidopsis. *Plant Physiol.* **2003**, *132*, 1020–1032. [CrossRef]

International Journal of
Molecular Sciences

Article

Identification and Characterization of microRNAs in the Developing Seed of Linseed Flax (*Linum usitatissimum* L.)

Tianbao Zhang [1], Zhen Li [1], Xiaxia Song [1], Lida Han [1], Limin Wang [2], Jianping Zhang [2], Yan Long [1,]* and Xinwu Pei [1,]*

[1] Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China; zhangtianbao@caas.cn (T.Z.); 18871433766@163.com (Z.L.); songxxia@163.com (X.S.); hanlida@caas.cn (L.H.)

[2] Crop Institute, Gansu Academy of Agricultural Sciences, Lanzhou 730070, China; 13893414680f@163.com (L.W.); Z13038703697@163.com (J.Z.)

* Correspondence: longyan@caas.cn (Y.L.); peixinwu@caas.cn (X.P.)

Received: 24 February 2020; Accepted: 9 April 2020; Published: 14 April 2020

Abstract: Seed development plays an important role during the life cycle of plants. Linseed flax is an oil crop and the seed is a key organ for fatty acids synthesis and storage. So it is important to understand the molecular mechanism of fatty acid biosynthesis during seed development. In this study, four small RNA libraries from early seeds at 5, 10, 20 and 30 days after flowering (DAF) were constructed and used for high-throughput sequencing to identify microRNAs (miRNAs). A total of 235 miRNAs including 114 known conserved miRNAs and 121 novel miRNAs were identified. The expression patterns of these miRNAs in the four libraries were investigated by bioinformatics and quantitative real-time polymerase chain reaction (qPCR) analysis. It was found that several miRNAs, including *Lus-miRNA156a* was significantly correlated with seed development process. In order to confirm the actual biological function of *Lus-miRNA156a*, over-expression vector was constructed and transformed to *Arabidopsis*. The phenotypes of homozygous transgenic lines showed decreasing of oil content and most of the fatty acid content in seeds as well as late flowering time. The results provided a clue that *miRNA156a* participating the fatty acid biosynthesis pathway and the detailed molecular mechanism of how it regulates the pathway needs to be further investigated.

Keywords: microRNA; miRNA156; seed development; fatty acid synthesis; linseed flax

1. Introduction

MiRNAs are non-coding RNAs about 20–25 nucleotide (nt) in length and they are encoded by endogenous genes, from which a primary non-protein coding message is transcribed (pri-miRNA) [1]. The pri-miRNA sequence contains an imperfect hairpin stem-loop structure allowing the molecule to fold-back onto itself to form dsRNA. Previous studies have confirmed that miRNAs regulate the expression of target genes through target mRNA degradation or translation inhibition [1–3]. As an important mechanism of post-transcriptional regulation, miRNAs could regulate genes involved in many developmental processes, such as flowering time, leaf development, auxin signaling and organ polarity [4–8].

MiRNAs are also found involving in the regulation of seed development process, such as *miR156*, *miR397*, *miR396* and *miR408*. For example, *miR396* regulates seed size and yield via its' target gene *OsGRF4* in rice [9]. In lettuce, Huo et al. found that the suppression of *DELAY OF GERMINATION1* (*DOG1*) gene could enable seed germination at high temperature in associated with reduced *miR156* and increased *miR172* levels [10]. Seeds are important storage organs. For oil crops, several fatty acids like oleic acid, linoleic acid and linolenic acid are storage in the seeds. Until now, several miRNAs have been discovered via high-throughput sequencing during the seed development in different oil crops, such as soybean, rapeseed. In soybean, 55 annotated miRNAs and 26 new soybean miRNAs were detected in a seed small RNA library [11]. Using high-throughput technology, there are 85 known miRNAs from 30 miRNA families as well as 1610 novel miRNA at stages of different seed development in *B. napus* [12]. Among the abundant miRNAs, some specific RNAs have been found relating with fatty acid biosynthesis. For example, *bBna-miR156b*, *bna-miR156g*, *bna-miR159*, *bna-miR395b*, *bna-miR6029* and 19 novel miRNAs were found to be involved in fatty acid biosynthesis [12]. As a conservative miRNA existing in different crops, *miRNA156* was found to be involved in different aspects of agronomic traits, such as fruit development in tomato, tuberization in potato, nodulation in soybean [13]. For example, *miR156* could play important roles in the modulation of grape berry development and ripening [14,15]. *Vv-miR156* exhibited an overall increasing expression trend during berry development and ripening in grape [15]. Previous studies have confirmed that *miR156* regulating developing processed through the *SPL* gene family [16]. In *Arabdopsis thaliana*, *miR156* has ten targets, such as *SPL2*, *SPL3*, *SPL4*, *SPL5*, *SPL6*, *SPL9*, *SPL10*, *SPL11*, *SPL13* and *SPL15* genes [16]. *MiR156* targets *SPL10* and *SPL11* genes, which can cause abnormal cell division and control the development of seeds in *Arabidopsis thaliana* [17]. There were too many researches about *miR156* but researches involving in fatty acids metabolism of *miR156* are limited.

Flax is an annual herbaceous dicotyledonous plant, which is divided into oil flax, fiber flax and oil fiber dual-purpose flax [18]. Flax seed is not only a reproductive organ to maintain generation continuity but also an organ to store oil and its storage capacity directly affects oil content and grain yield. Linseed oil is rich in a variety of unsaturated fatty acids, such as oleic acid, linoleic acid, linolenic acid and so forth, especially the higher content of α-linolenic acid, the average content of 40% to 60%. To identify the key genes involved in fatty acids biosynthesis, a cDNA library made from flax bolls collected at 12 days after anthesis was constructed and screened for ketoacyl CoA synthase, fatty acid elongase, stearoyl-ACP desaturase and fatty acid desaturase [19].

In recent years, some progress has been made in the identification and functional analysis of miRNA in flax. Neutelings identified 20 conserved miRNAs belonging to 13 families [20]. In the following years, miRNAs that play a role in the absorption and reaction of nutrients, such as N and P, were identified in flax [21,22]. For example, Melnikova identified a total of 96 conserved miRNAs under normal and deficient phosphorus conditions and found 475 new potential miRNAs [21]. However, the role of miRNAs in flax seed development remains unclear. Therefore, the identification of miRNAs in linseed flax seed development and the clarification of their functions will help to understand the regulatory process of flax seed development.

2. Results

2.1. Small RNA Libraries Data Analysis

Four small RNA libraries, M5, M10, M20 and M30, from four development stages of flax seed were constructed. After high-throughput sequencing, 22,179,284, 34,145,577, 19,947,423 and 18,881,760 reads were successively obtained respectively for the four libraries.

Reads without 3′ adaptor sequence and insert fragment were removed. Sequences shorter than 18 or longer than 30 nucleotides were removed. 20,601,320, 13,039,680, 17,501,654 and 13,870,666 reads were successively obtained (Table 1). Using Bowtie software, clean reads were screened against Silva database, GtRNAdb database, Rfam database and Repbase database. NcRNAs including ribosomal RNA (rRNA), transfer RNA (tRNA), small nuclear RNA (snRNA), small nuclear RNA (snoRNA) and repeat sequences were filtered to obtain unannotated reads containing miRNA. The average ratio of the content of rRNA in four libraries was less than 20%, indicating that the quality of the four libraries constructed was reliable. The unannotated reads was 91.8%, 75.69%, 82.4% and 52.67%, respectively (Table 1). Then, the unannotated reads were compared to the flax genome and 11,568,108, 6,180,914, 9,606,035, 4,181,608 reads were matched to the flax genome, which accounted for 61.17%, 62.64%, 66.62% and 57.24% of the unannotated reads (Table 1).

Table 1. The number of sequencing reads and distribution of sRNAs obtained from 4 libraries.

Samples	M5	M10	M20	M30
Raw_reads	22,179,284	34,145,577	19,947,423	18,881,760
Containing N′ reads	4476	6996	3921	3772
Length<18	432,915	20,339,545	1,489,610	1,056,229
Length>30	1,140,573	759,356	952,238	3,951,093
Clean_reads	20,601,320(100%)	13,039,680(100%)	17,501,654(100%)	13,870,666(100%)
rRNA	1,537,810(7.46%)	2,099,680(16.10%)	2,701,051(15.43%)	5,230,679(37.71%)
scRNA	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)
snRNA	2(0.00%)	3(0.00%)	0(0.00%)	0(0.00%)
snoRNA	1364(0.01%)	5601(0.04%)	2587(0.01%)	10,255(0.07%)
tRNA	146,574(0.71%)	1,052,700(8.07%)	371,121(2.12%)	1,314,919(9.48%)
Repbase	3693(0.02%)	13,668(0.10%)	6974(0.04%)	9299(0.07%)
Unannotated	18,911,877(91.80%)	9,868,028(75.69%)	14,419,921(82.40%)	7,305,514(52.67%)
Mapped_Reads to genome	11,568,108	6,180,914	9,606,035	4,181,608

2.2. Identification of Known and Novel miRNAs

After screening the miRNAs, it was found that the length of the miRNAs varied from 18 nt-24 nt and the miRNA content of 21 nt was the highest, indicating that miRNAs with 21 nt played important roles in linseed flax seed development (Figure 1). MiRDeep2 software was used to identify known and new miRNAs. Through the alignment of reads to the flax genome, possible precursor sequences are obtained. Based on the distribution information of reads on the precursor sequences and the energy information of the precursor structures, the Bayesian model was used to score and finally realize the identification of miRNAs. Totally, there were 235 miRNAs predicted for all samples, including 114 known miRNAs and 121 newly predicted miRNAs (Table 2). For the 114 known unique miRNAs, there were 23 families, including the conserved miRNA156, *miRNA166*, *miRNA169* and et al. The family members ranged from one to 11. The *miR166* family was the largest family which having 11 members, followed by *miR156*, *miR167*, *miR169*, *miR171* and *miR172* with 9 members. The 121 newly predicted miRNAs distributed in 89 scaffolds and 5 contigs of the flax genome. The pri-sequences and mature sequences of all of the 235 miRNAs were listed in Supplementary Table S1.

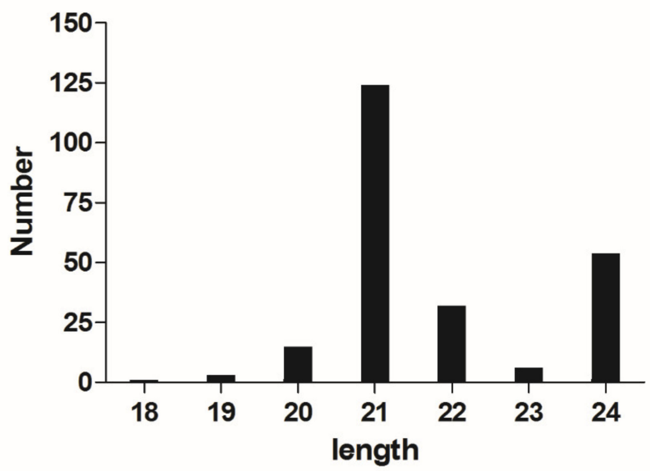


Figure 1. The length of miRNAs in the four libraries.

Table 2. Distribution of miRNAs in four libraries.

Samples	Known-miRNAs	Novel-miRNAs	Total
M5	108	117	225
M10	110	120	230
M20	104	120	224
M30	87	118	205
Total	114	121	235

2.3. Analysis of miRNA Expression in Four Developing Stages of Seeds

After getting the miRNA sequences, the expression values of all the miRNAs were calculated. Among all the 235 identified miRNAs, 199 miRNAs co-expressed in the four stages of seed development. Two miRNAs, *lus-miR169g* and *lus-miR169l*, specifically expressed in M5 library and four miRNAs (*lus-miR156d*, *lus-miR171a*, *lus-miR171f* and *lus-miR828a*) specifically expressed in the M10 library (Supplementary Table S2). There were no miRNAs specifically expressed in M20 and M30 libraries. There were 221 miRNAs co-expressed in M5 and M10 libraries, while 4 and 9 expressed separately in M5 and M10 libraries. 118 miRNAs co-expressed in M5 and M20 libraries while 7 and 6 expressed separately in M5 and M20 libraries. There were 24 and 4 expressed separately in M5 and M30 libraries, while 201 miRNAs co-expressed in M5 and M30 (Figure 2, Supplementary Table S2).

Then, the expression values of miRNAs in each sample were statistically analyzed and normalized by TPM algorithm. |log2(FC)| ≥ 1 and FDR ≤ 0.01 were used as screening criteria in the detection of differentially expressed miRNAs. After calculating the expression values of the miRNAs, 101, 158 and 154 miRNAs showed significantly different expression between 5 DAF and 10 DAF, 5 DAF and 20 DAF and 5 DAF and 30 DAF libraries, respectively. For 101 differentially expressed miRNAs were detected between M10 and M5 libraries, 48 were up-regulated and 53 were down-regulated. There were 158 differentially expressed miRNAs in M20 and M5 libraries, among which 91 were up-regulated and 67 were down-regulated. A total of 154 differentially expressed miRNAs were detected in M30 and M5 libraries, of which 68 were up-regulated and 86 down-regulated (Figure 2).

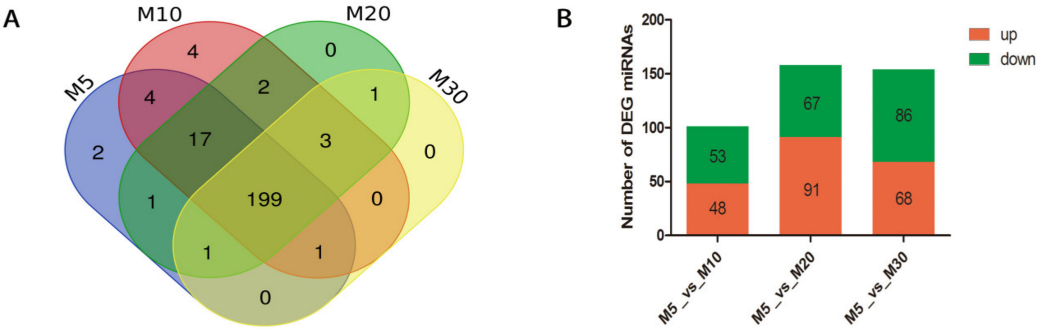


Figure 2. Comparison of the expression of miRNAs in the four libraries. (**A**). The numbers of miRNAs expressing in each of the library. (**B**). Differentially expressed miRNAs during seed development. M5 library was as a control.

2.4. Target Gene Prediction and GO Analysis for the Differentially Expressed Target Genes

In general, miRNA and target genes are generally negatively regulated. In order to find target genes of the identified miRNAs, TargetFinder software was used to predict. As a result, for the 235 identified miRNAs, only 109 miRNAs have been predicted 630 targets (Supplementary Table S3). For the 114 known conserved miRNAs, 85 miRNAs had predicted target genes, while for the novel miRNAs, only 24 miRNAs had predicted target genes. Many known miRNA families, such as *lus-miRNA156* (*a*–*i*), *lus-miRNA160* (*a*–*j*), had different potential functional target genes, which mean that these miRNAs are involved in regulating multiple genes' expression in linseed flax.

Comparing the M5 and M10 libraries, there were 43 significantly altered genes. There were 67 significantly altered genes between M5 and M20 libraries. Meanwhile, 59 significantly altered genes between M5 and M30 libraries. Then GO analysis including three major types was used to classify the gene function of the target and differentially expressed target genes. For the targets, there were 6 molecular functions, 16 biological processes and 9 cellular components were included. For differentially expressed target genes between M5 vs. M10 library, 4 molecular functions, 14 biological processes and 6 cellular components were included. For M5 vs. M20 library, 6 molecular functions, 16 biological processes and 8 cellular components were included. And for M5 vs. M30 library, the differentially expressed genes could be divided into 5 molecular functions, 15 biological processes and 8 cellular components (Figure 3). Several biological processes including metabolic process, biological regulation and cellular process were included.

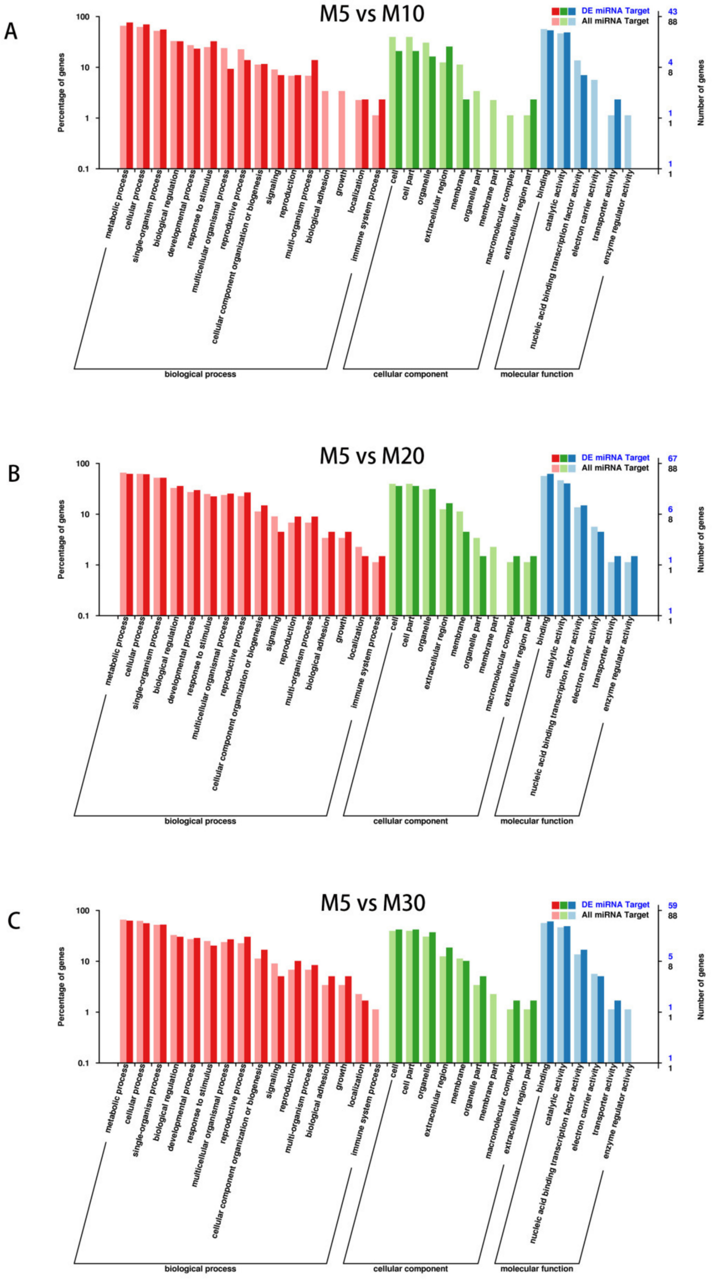


Figure 3. Gene ontology classifications of miRNA targets and differentially expressed targets in seed development. (**A**) M5 vs. M10, (**B**) M5 vs. M20, (**C**) M5 vs. M30.

2.5. MiRNA Expression Verification by Using qRT-PCR

To confirm the sequencing results and examine the expression patterns of the miRNAs at different stages of seed development (M5, M10, M20 and M30), four known conserved miRNAs including *Lus-miR156a*, *Lus-miR172e*, *Lus-miR159b*, *Lus-miR397a* and two novel identified miRNAs, *Lus-miR-10* and *Lus-miR-24* were used to perform qPCR (Figure 4). As the seed develops to maturity, *Lus-miR156a*, *Lus-miR159b* and *Lus-miR-10* was steadily increasing from M5 to M30 stage. For *Lus-miR397a* and *Lus-miR-24*, they expressed very low in the 5 DAF to 20 DAF, while the expression values increased much in the 30 DAF. And *Lus-miR172e* was down-regulated during the seed development process. The expression tendency was same after comparing with the high-throughput sequencing result (Figure 4). The results showed that the sequencing results of the four libraries in this study was reliable and the identified miRNAs could be further investigated for illustrating the relationships between them and seed development and even fatty acid synthesis.

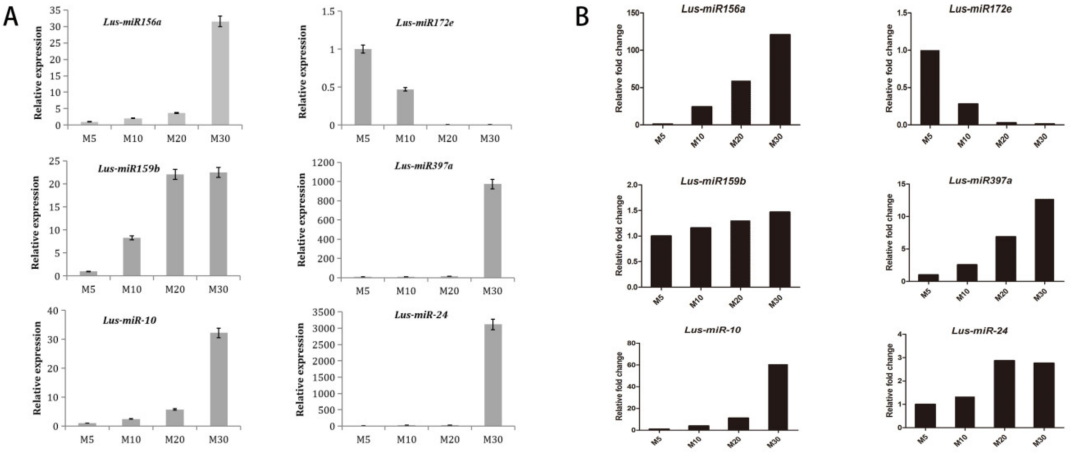
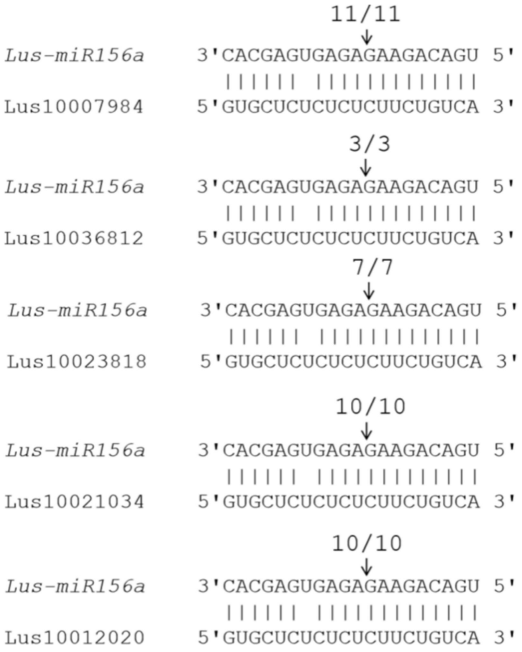


Figure 4. Expression patterns of *miR156a*, *miR172e*, *miR159b*, *miR397a*, *Lus-miR-10* and *Lus-miR-24* in seed developmental stages (M5, M10, M20 and M30) of qPCR and Next-generation sequencing (NGS) data. Error bars indicated standard deviation of three replicates. (**A**)The results of qPCR, (**B**) The results from NGS data.

2.6. Screening of Target Genes of Lus-miR156a in Flax Genome

Based on the miRNA identification and differentially expressed gene analysis, it was inferred that *Lus-miR156a* and its' related target genes could participate the seed development process. So then the cleavage sites of target genes were screened by using the 5′-RLM-RACE method, so as to determine the real target gene of miRNA156. The results were shown in Figure 5. It showed that totally five target genes of *lus-miR156a* having cleavage sites, which are cleaved between the 11th and 12th bases in the complementary region. After searching the potential gene functions of these five genes, it was found that they were homology with the *SPL* genes, especially had highly similarity with *SPL6* and *SPL9* genes in *Arabidopsis* genome.



```
                         11/11
                           ↓
Lus-miR156a     3'CACGAGUGAGAGAAGACAGU 5'
                  |||||| |||||||||||||
Lus10007984     5'GUGCUCUCUCUCUUCUGUCA 3'
                          3/3
                           ↓
Lus-miR156a     3'CACGAGUGAGAGAAGACAGU 5'
                  |||||| |||||||||||||
Lus10036812     5'GUGCUCUCUCUCUUCUGUCA 3'
                          7/7
                           ↓
Lus-miR156a     3'CACGAGUGAGAGAAGACAGU 5'
                  |||||| |||||||||||||
Lus10023818     5'GUGCUCUCUCUCUUCUGUCA 3'

                         10/10
                           ↓
Lus-miR156a     3'CACGAGUGAGAGAAGACAGU 5'
                  |||||| |||||||||||||
Lus10021034     5'GUGCUCUCUCUCUUCUGUCA 3'

                         10/10
                           ↓
Lus-miR156a     3'CACGAGUGAGAGAAGACAGU 5'
                  |||||| |||||||||||||
Lus10012020     5'GUGCUCUCUCUCUUCUGUCA 3'
```

Figure 5. Mapping target mRNA cleavage sites by 5′-RLM-RACE. The arrows indicate the cleavage sites and the numbers show the frequency of clones sequenced.

2.7. Overexpression of miR156a Affects the Flowering Time, Rosette Leaves and Fatty Acids

To analyze the function of *Lus-miR156a*, the overexpression plasmid vector was constructed and transformed it to *Arabidopsis*. Then the phenotypes of flowering time, rosette leaves and fatty acids of the homologous transgenic lines of T3 generation were investigated. For the rosette leaves of *Lus-miR156a* overexpression line were 3.67–4.00 leaves than *Arabidopsis* wild-type (Figure 6A). And for flowering time, it was found that the flowering time of *Lus-miR156a* overexpression lines delayed 3.67–4.00 day than *Arabidopsis* wild-type (Figure 6B). After getting the mature seeds, the fatty acid content and oil content were analyzed by GC-MS. The total oil content of the transgenic lines decreased 10% compared to that of the wild type (Table 3). Besides the total oil content, the fatty acid contents of the seed were investigated. It was found that the levels of C16:0, C18:0, C18:2$^{\Delta9,12}$, C18:3$^{\Delta9,12,15}$, C20:2$^{\Delta11,14}$ in *Lus-miR156a* overexpression lines were significantly lower than that of the wild type. It means that *Lus-miR156a* actually participate the fatty acid metabolism pathway.

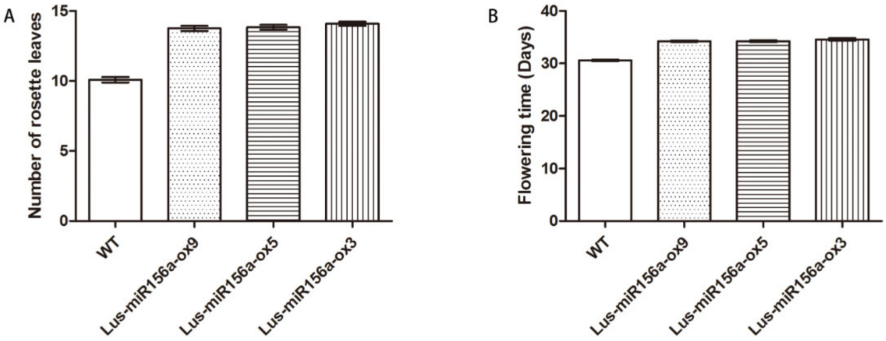


Figure 6. Number of rosette leaves and flowering time of the transgenic lines and wild type. (**A**) Number of rosette leaves, (**B**) Flowering time.

Table 3. Fatty acids profiles of *Arabidopsis* seeds over-expressing *Lus-miR156a.*

Fatty Acids	WT	OXmiR156a-3	OXmiR156a-5	OXmiR156a-9
C16:0	0.03 ± 0.002	0.02 ± 0.003	0.021 ± 0.004 *	0.02 ± 0.004 *
C16:1$^{\Delta9}$	0.002 ± 0.001	0.002 ± 0.001	0.003 ± 0.001	0.003 ± 0.001
C18:0	0.009 ± 0.001	0.005 ± 0 **	0.006 ± 0.001 **	0.006 ± 0.001 **
C18:1$^{\Delta9}$	0.047 ± 0.006	0.033 ± 0.009	0.04 ± 0.005	0.036 ± 0.002 *
C18:2 $^{\Delta9,12}$	0.106 ± 0.002	0.079 ± 0.013 **	0.085 ± 0.004 **	0.081 ± 0.008 **
C18:3 $^{\Delta9,12,15}$	0.06 ± 0.005	0.042 ± 0.01 *	0.044 ± 0.005 **	0.04 ± 0.007 **
C20:0	0.004 ± 0.001	0.003±0.001	0.003 ± 0.001	0.003 ± 0.001
C20:1 $^{\Delta11}$	0.04 ± 0.01	0.028±0.005*	0.034 ± 0.006 *	0.035 ± 0.011
C20:2 $^{\Delta11,14}$	0.004 ± 0.001	0.003 ± 0 **	0.003 ± 0 **	0.003 ± 0 **
C22:1 $^{\Delta13}$	0.005 ± 0.002	0.004 ± 0	0.006 ± 0.002	0.005 ± 0.002
Sum	0.323 ± 0.011	0.229 ± 0.026 **	0.25 ± 0.006 **	0.238 ± 0.024 **

* $p < 0.05$ and ** $p < 0.01$ indicated significant differences with WT.

2.8. Seed Oil Synthesis Genes Were Regulated by Lu-miR156a in Arabidopsis Transgenic Lines

In the *Lu-miR156a* over-expression transgenic lines, the target gene, including *SPL6* and *SPL9*, were significantly decreased than WT, while the other *SPL* genes, such as *SPL3*, *SPL10, SPL11* were not significantly changed compared to WT (Figure 7). This result showed that *SPL6* and *SPL9* genes were target genes of miR156a in seed development process. In order to evaluate how the seed oil synthesis genes regulated by *Lu-miR156a*, some oil synthesis related genes were used. The results showed that *FAD2*, *FAD3* and *FAE1* were significantly decreased in transgenic lines (Figure 7). It mean that over-expression of *Lu-miRNA156a* influenced the gene expression of seed oil synthesis genes.

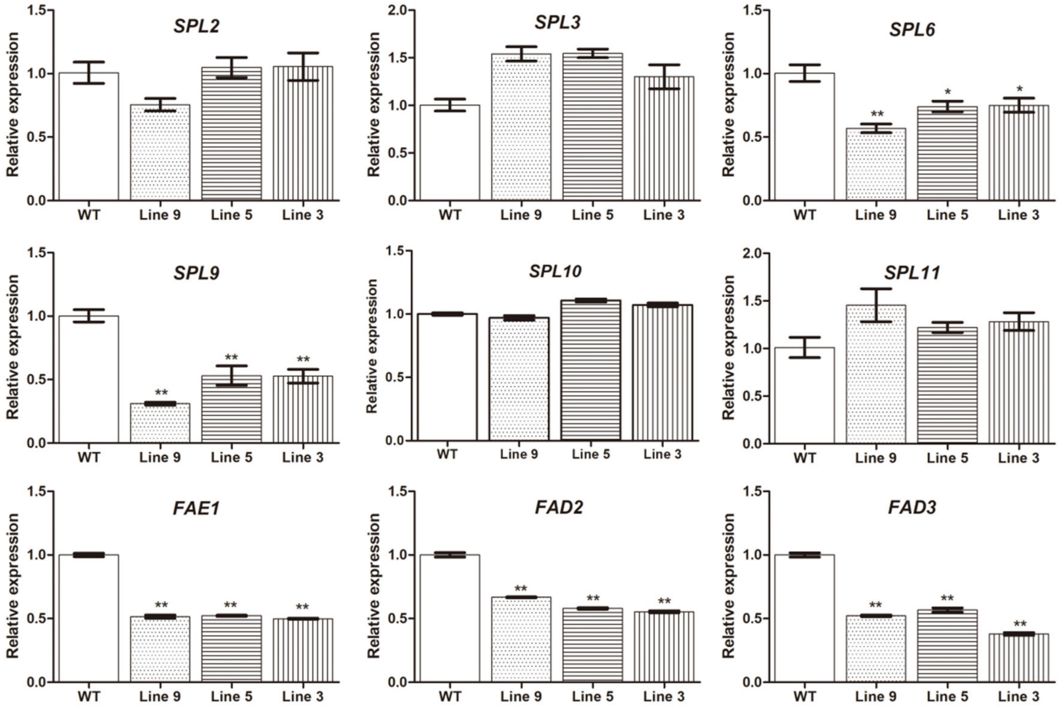


Figure 7. Expression levels of SPL genes and seed oil synthesis genes in the over-expression *Arabidopsis* transgenic lines and WT. Line 9: Lus-MIR156a-OX9, Line 5: Lus-MIR156a-OX5, Line 3: Lus-MIR156a-OX3. * $p < 0.05$ and ** $p < 0.01$ indicated significant differences with WT.

3. Discussion

As an important variety of oil crops, there are some studies involving in miRNAs for linseed flax, while until now there are few studies focusing on seed development or fatty acid synthesis. In this study, 235 miRNAs were predicted based on four libraries from four seed developing stages, which including 114 known miRNAs belonged to 23 families and 121 novel miRNAs. Then, a conservative miRNA, *lus-miRNA156a* was further investigated for its' target genes and potential biological functions. The phenotypes of transgenic lines showed *lus-miRNA156a* could regulate flowering time. The most interesting result of this study was that over expression of *Lus-miRNA156a* could change the fatty acid content and total amount of the oil content in seeds. This result provided the new evidence that miRNA156a could attend regulating the fatty acid synthesis pathway in linseed flax.

In this study, four small RNA libraries, M5, M10, M20 and M30, were constructed for sequencing. Totally 235 miRNAs were identified correlating with the seed developing process. Compared the miRNA numbers to the previous published researches, it was found that the miRNA numbers would increase when the libraries increased. In soybean, a small RNA library, 15 DAF, was used for miRNA identification and totally 207 miRNAs were identified [11]. There were also 85 unique miRNAs identified when using three small RNA libraries in *B.napus* [12]. This mean that more samples will help to identify more miRNAs which regulate specific developing process in plants, including seed development. Among the 235 miRNAs, the 114 unique miRNAs belonged to 23 families. With the 23 families, the *miR166* family was the largest family which having 11 members, followed by *miR156*, *miR167*, *miR169*, *miR171* and *miR172* with 9 members. The tendencies of the results were similar with previous studies and with some differences. Wang et al. found that the *miR169* family was the largest family with 10 members in the three small miRNA libraries of *B.napus* seeds [12]. While also in *B.napus* seeds, Korbes et al. found *miR156/157* was the largest family (24 members), followed by the *miR165/166* (21 members) and *miR169* (15 members) families. There were 2–6 members were found existing in the remaining miRNA families identified [23]. The different expression of these small RNAs in different oil corps suggest that they function in common and unique regulation pathways.

Since the miRNA and target duplexes are near-perfectly matched in plants, it is possible to find targets by computational approach. In the current study, for the 235 identified miRNAs, 109 miRNAs have 630 target genes. The 85 conserved miRNAs had target genes, while only 24 novel miRNAs had target genes, which meant that conserved miRNAs had more target genes than that of the novel discovered miRNAs. The reason is that conserved miRNAs play key roles in universal mechanisms of regulation in different plant varieties. While the linseed flax specific miRNAs may only function in regulation of gene expression during flax seed developing stages. Meanwhile, most of the target genes of conserved miRNAs are transcription factors and the linseed flax specific miRNAs may regulate various types of genes. It means that there is a new feature of miRNA regulation pathway in linseed flax.

In general, the miRNAs act to down-regulate their target genes by directing cleavage of the highly complementary target transcripts. In *Arabidopsis*, of the 16 Squamosa-promoter Binding Protein (SBP)-like genes, ten have *miR156* complementary sites [24,25]. 19 rice *SPL* genes and 12 rice *miRNA156* precursors were identified in the rice genome. Sequence and experimental analysis suggested that 11 *OsSPL* genes were putative targets of *OsmiR156* [26]. Ten *SlySBP* genes carry putative miR156-binding sites in tomato [27]. Using the 5′-RLM-RACE method, we found the five target genes of *miR156a* have cleavage sites, whose cleaved site is between the 11th and 12th bases in the complementary region. BLAST analysis showed that these five target genes were closet similarity with *SPL6* and *SPL9* of *Arabdopsis*. *SPL9* have a major function in both the vegetative-to-reproductive transition and the juvenile-to-adult vegetative transition [16]. *SPL6* does not contribute to in shoot morphogenesis but may be important for certain physiological processes [16]. The gene expressions of *SPL6* and *SPL9* were significantly down-regulated in *Lu-miR156a* over-expression plants, while the other four *SPL* genes were not regulated. It mean that *SPL6* and *SPL9* were target genes of *miR156a* in seed developing process. The data suggested that different crops contain miRNA with different target genes in different organs.

MiRNAs play important roles during plant growth, developmental transitions and determination of cell identity [28,29]. Seed is a crucial organ of plant. There are many researches on the development of seed development. While for miRNAs, most of the researches are based on high-throughput sequencing. For example, in oil seed crops, such as soybean, rapeseed, many miRNAs have been discovered for correlating with the seed development. While until now, there were no direct clues which could confirm the regulating pathway between miRNAs and seed development or fatty acid synthesis. The current study provided a new clue for *miRNA156* regulating the fatty acid metabolism. Previous studies have shown that *miR156* regulates many different aspects of developing. For flowering time, overexpression of *miR156* in both *Arabidopsis* and maize prolongs the expression of juvenile vegetative traits and delays flowering [30,31]. Also in our study, *miR156a* can increase the number of rosettes leaves, leading to delay the flowering time. Besides the flowering time and rosette leaves phenotypes, the fatty acids profiles of overexpression *miR156a* lines were changed, the total oil content of the transgenic lines decreased comparing with that in wild type, also the fatty acid content in the seeds. It indicates that *Lus-miR156a* plays a certain role in fatty acid metabolism. For the specific fatty acid, the linoleic acid, the linolenic acid contents were decreased. The three genes, *FAD2*, *FAD3* and *FAE1*, which mainly regulated the synthesis of these two fatty acids, were down-regulated in the transgenic plants compared to WT. It mean that miRNA156a actually regulated the gene expression of seed oil synthesis pathway. While the actual regulating mechanism of how miRNA156a work on seed oil synthesis and metabolism through regulating its' target genes need to be further investigated.

4. Materials and Methods

4.1. Plant Materials and Growth Condition

One linseed flax cultivar, Macbeth, was used as material for miRNA identification. Its oil content is about 46.7% and linolenic content is about 44.5% [32]. The cultivar was planted in the field trial station of Chinese Academy of Agricultural Sciences in Langfang city, Hebei province, China. The blooming flower was tagged and then the siliques were collected respectively in 5, 10, 20 and 30 days after flowering. All the samples were put in liquid nitrogen immediately for later RNA extraction.

4.2. RNA Extraction and Small RNA Library Construction

Total RNA was extracted using the EASY spin Plant microRNA Kit (Aidlab, Beijing, China). NEB Next Ultra small RNA Sample Library Prep Kit for Illumina (NEB, Ipswich, MA, USA) was used for small RNA Library construction. Sequencing was performed using Illumina HiSeq2500 high-throughput sequencing platform (Biomarker Technology, Beijing, China).

4.3. Analysis of Sequencing Results

The sequencing raw reads with the content of unknown base N greater than or equal to 10% were removed. Reads without 3 'adaptor sequence and inserted fragment were removed. Sequences shorter than 18 nucleotides or longer than 30 nucleotides were removed and clean reads were obtained. Using Bowtie software, clean reads respectively with Silva database(http://www.arb-silva.de/), GtRNAdb database(http://lowelab.ucsc.edu/GtRNAdb/), Rfam database(http://rfam.xfam.org/) and Repbase database(http://www.girinst.org/repbase/) sequence alignment. To obtain unannotated reads containing miRNA, ribosomal RNA (rRNA), transfer RNA (tRNA), small nuclear RNA (snRNA), small nuclear RNA (snoRNA) and repeat sequences were filtered. Clean reads were compared with the flax genome (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Lusitatissimum) using the Bowtie software and the location information on the reference genome was obtained, that is, mapped reads. Identification of known and new miRNAs were performed with miRDeep2 software using the criteria for plant miRNAs annotation [33]. According to the known miRNA and newly predicted miRNA and flax gene sequence information, TargetFinder (https://www.acunetix.com/blog/docs/target-finder/) was used for target gene prediction.

4.4. Expression of miRNAs

High-throughput sequencing data was normalized the expression levels with TPM algorithm [34]. |log2(FC)|≥1 and FDR≤0.01 were used as screening criteria in the detection of differentially expressed miRNA. qPCR was then used to verify miRNAs expression. Ten ng RNA was reversed using TaqManH miRNAs Reverse Transcription Kit (Applied Biosystems, USA). The obtained cDNA was amplified using a miRNA Universal SYBR qPCR Master Mix (Vazyme, Nanjing, China) Kit in ABI7500 real-time System (Applied Biosystems, Foster, CA, USA) with actin as internal reference. The $2^{-\Delta\Delta Ct}$ method expression levels of miRNA in different materials were calculated using the comparative Ct method [35,36]. Four known miRNAs, such as *miR156a*, *miR159b*, *miR172e* and *miR397a* and two new miRNAs, such as *Lus-miR-10* and *Lus-miR-24*, were performed qPCR to analyze their expression level.

4.5. GO Analysis for Target Genes and Differentially Expressed Genes

In order to understand the targets of miRNAs and classifications as well as the metabolic regulatory networks associated with linseed flax miRNAs and their targets, all of the target genes and differentially expressed targets were mapped to Gene Ontology (GO) terms(http://www.geneontology.org). The number of the genes of each term was calculated. GO terms with a p-value less than the threshold of 0.05 were considered to be significantly enriched. GO annotation results were plotted using WEGO (http://wego.genomics.org.cn/).

4.6. Verification of Cleavage Sites of miRNA Target Genes

In order to verify the target gene cleavage sites of miRNA, the 5′-RLM-RACE technique was applied [37]. First of all, the 500 µg total RNA by Oligolex ®mRNA mini Kit (Qiagen, Hilden, Germany) enrichment of mRNA, further use of First Choice ®RLM-RACE Kit (Ambion, Foster, CA, USA) for the enrichment of mRNA and 5′ joints, reverse transcription cDNA, the use of nested PCR for two rounds of amplification, recycling series connected to 5minTM TA/Blunt-Zero Cloning Kit (Vazyme, Nanjing, China) and sequencing. Ten to fifteen clones were used for sequencing to confirm the cleavage sites.

4.7. Arabidopsis Plant Transformation

In order to check whether *LumiR156a* is correlating with the agronomic traits, *miR156* over-expression vectors (Pro35S: miR156a) were constructed. To construct Pro35S: miR156a, a 0.5 KB fragment encompassing the pri-miR156a sequence was amplified (MIR156a-F: ACGGGGGACTGAA TTCTGTGTAAGGACAAGAGAGGTAGC; MIR156a-R: CCGCCTCGAGCCCGGGAGTAAGGACA CCTGGAGGCT) introducing *EcoR* I and *Xma* I restriction sites and subcloned into pBinGlyRed vector [38]. Plants were transformed with Agrobacterium EHA105 containing the Pro35S: *miR156*a construct. The seeds of Arabidopsis *Col*-0 ectotype were directly sown into the soil and grew in the culture room under conditions (16 h light / 8 h dark, 23 °C). Floral dip method was used to transform the wild type plants [39]. Red seeds were selected as positive under green light.

4.8. Phenotypes of Transgenic Arabidopsis Plant

Homozygous transgenic *Arabidopsis* lines in T3 generation contained *LumiR156a* were obtained and their characters were identified. Phenotypes of the flowering time, the number of rosette leaves and the oil content in seeds were measured. The phenotypes were collected from 12 plants in each line. Fatty acid methyl esters were extracted into hexane and analyzed by GC-MS [40]. Fatty acid methyl esters were formed by transesterification of 5–10 mg seeds by heating with MeOH-H2SO4(19:1) at 70 °C for 30 min. Fatty acid compositions were calculated against the internal control.

4.9. MiRNA and Gene Expression Analysis in Transgenic Arabidopsis Plants

As *SPL6* and *SPL9* genes were screened as the target genes of *Lu-miRNA156a* in linseed flax, the homologous genes of them in *Arabidopsis* genome were selected as target genes to evaluate their expression values in *Arabidopsis* transgenic lines by using qPCR method. Besides *SPL6* and *SPL9*, the gene expression values during seed development of some other SPL genes, including *SPL2*, *SPL3*, *SPL10* and *SPL11* were also evaluated. The siliques with 20DAF were used as materials. Meanwhile, the important seed oil synthesis genes, *FAD2*, *FAD3* and *FAE1* were chosen for gene expression analysis. The primers were listed in Supplementary Table S4.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/21/8/2708/s1.

Author Contributions: Conceptualization, Y.L. and X.P.; methodology, L.H.; software, T.Z.; investigation, Z.L. and X.S.; resources, L.W. and J.Z.; data curation, T.Z. and Y.L.; writing—original draft preparation, T.Z.; writing—review and editing, Y.L. and X.P.; supervision: X.P.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No.31570330).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

miRNA	microRNA
DAF	days after flowering
qPCR	quantitative real-time polymerase chain reaction
nt	nucleotide
FC	fold change
FDR	false discovery rate
TPM	transcripts per million
NGS	next-generation sequencing
GO	Gene Ontology
SPL	SQUAMOSA PROMOTER BINDING PROTEIN-LIKE
GC-MS	Gas Chromatography-Mass Spectrometer

References

1. Xie, M.; Zhang, S.; Yu, B. microRNA biogenesis, degradation and activity in plants. *Cell. Mol. Life Sci.* **2015**, *72*, 87–99. [CrossRef]
2. Ameres, S.L.; Zamore, P.D. Diversifying microRNA sequence and function. *Nat. Rev. Mol. Cell Biol.* **2013**, *14*, 475–488. [CrossRef]
3. Huntzinger, E.; Izaurralde, E. Gene silencing by microRNAs: Contributions of translational repression and mRNA decay. *Nat. Rev. Genet.* **2011**, *12*, 99–110. [CrossRef] [PubMed]
4. Baulcombe, D. RNA silencing in plants. *Nature* **2004**, *431*, 356–363. [CrossRef] [PubMed]
5. Chapman, E.J.; Carrington, J.C. Specialization and evolution of endogenous small RNA pathways. *Nat. Rev. Genet.* **2007**, *8*, 884–896. [CrossRef] [PubMed]
6. Huang, D.; Koh, C.; Feurtado, J.A.; Tsang, E.W.; Cutler, A.J. MicroRNAs and their putative targets in Brassica napus seed maturation. *BMC Genom.* **2013**, *14*, 140. [CrossRef]
7. Zhang, J.; Zhang, S.; Han, S.; Wu, T.; Li, X.; Li, W.; Qi, L. Genome-wide identification of microRNAs in larch and stage-specific modulation of 11 conserved microRNAs and their targets during somatic embryogenesis. *Planta* **2012**, *236*, 647–657. [CrossRef]
8. Chen, Z.; Li, F.; Yang, S.; Dong, Y.; Yuan, Q.; Wang, F.; Li, W.; Jiang, Y.; Jia, S.; Pei, X. Identification and functional analysis of flowering related microRNAs in common wild rice (Oryza rufipogon Griff.). *PLoS ONE* **2013**, *8*, e82844. [CrossRef]
9. Duan, P.; Ni, S.; Wang, J.; Zhang, B.; Xu, R.; Wang, Y.; Chen, H.; Zhu, X.; Li, Y. Regulation of OsGRF4 by OsmiR396 controls grain size and yield in rice. *Nat. Plants* **2015**, *2*, 15203. [CrossRef]

10. Huo, H.; Wei, S.; Bradford, K.J. DELAY OF GERMINATION1 (DOG1) regulates both seed dormancy and flowering time through microRNA pathways. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E2199–E2206. [CrossRef]
11. Song, Q.X.; Liu, Y.F.; Hu, X.Y.; Zhang, W.K.; Ma, B.; Chen, S.Y.; Zhang, J.S. Identification of miRNAs and their target genes in developing soybean seeds by deep sequencing. *BMC Plant. Biol.* **2011**, *11*, 5. [CrossRef] [PubMed]
12. Wang, J.; Jian, H.; Wang, T.; Wei, L.; Li, J.; Li, C.; Liu, L. Identification of microRNAs Actively Involved in Fatty Acid Biosynthesis in Developing Brassica napus Seeds Using High-Throughput Sequencing. *Front. Plant Sci.* **2016**, *7*, 1570. [CrossRef] [PubMed]
13. Wang, H.; Wang, H. The *miR156*/SPL Module, a Regulatory Hub and Versatile Toolbox, Gears up Crops for Enhanced Agronomic Traits. *Mol. Plant* **2015**, *8*, 677–688. [CrossRef] [PubMed]
14. Wang, B.; Wang, J.; Wang, C.; Shen, W.; Jia, H.; Zhu, X.; Li, X. Study on Expression Modes and Cleavage Role of *miR156*b/c/d and its Target Gene Vv-SPL9 During the Whole Growth Stage of Grapevine. *J. Hered.* **2016**, *107*, 626–634. [CrossRef]
15. Cui, M.; Wang, C.; Zhang, W.; Pervaiz, T.; Haider, M.S.; Tang, W.; Fang, J. Characterization of Vv-*miR156*: Vv-SPL pairs involved in the modulation of grape berry development and ripening. *Mol. Genet. Genom.* **2018**, *293*, 1333–1354. [CrossRef]
16. Xu, M.; Hu, T.; Zhao, J.; Park, M.Y.; Earley, K.W.; Wu, G.; Yang, L.; Poethig, R.S. Developmental Functions of *miR156*-Regulated SQUAMOSA PROMOTER BINDING PROTEIN-LIKE (SPL) Genes in Arabidopsis thaliana. *PLoS Genet.* **2016**, *12*, e1006263. [CrossRef]
17. Nodine, M.D.; Bartel, D.P. MicroRNAs prevent precocious gene expression and enable pattern formation during plant embryogenesis. *Genes. Dev.* **2010**, *24*, 2678–2692. [CrossRef]
18. Vrinten, P.; Hu, Z.; Munchinsky, M.A.; Rowland, G.; Qiu, X. Two FAD3 desaturase genes control the level of linolenic acid in flax seed. *Plant Physiol.* **2005**, *139*, 79–87. [CrossRef]
19. Fofana, B.; Duguid, S.; Cloutier, S. Cloning of fatty acid biosynthetic genes β-ketoacyl CoA synthase, fatty acid elongase, stearoyl-ACP desaturase, and fatty acid desaturase and analysis of expression in the early developmental stages of flax (Linum usitatissimum L.) seeds. *Plant Sci.* **2004**, *166*, 1487–1496. [CrossRef]
20. Neutelings, G.; Fenart, S.; Lucau-Danila, A.; Hawkins, S. Identification and characterization of miRNAs and their potential targets in flax. *J. Plant Physiol.* **2012**, *169*, 1754–1766. [CrossRef]
21. Melnikova, N.V.; Dmitriev, A.A.; Belenikin, M.S.; Koroban, N.V.; Speranskaya, A.S.; Krinitsina, A.A.; Krasnov, G.S.; Lakunina, V.A.; Snezhkina, A.V.; Sadritdinova, A.F.; et al. Identification, Expression Analysis, and Target Prediction of Flax Genotroph MicroRNAs Under Normal and Nutrient Stress Conditions. *Front Plant. Sci.* **2016**, *7*, 399. [CrossRef] [PubMed]
22. Melnikova, N.V.; Dmitriev, A.A.; Belenikin, M.S.; Speranskaya, A.S.; Krinitsina, A.A.; Rachinskaia, O.A.; Lakunina, V.A.; Krasnov, G.S.; Snezhkina, A.V.; Sadritdinova, A.F.; et al. Excess fertilizer responsive miRNAs revealed in Linum usitatissimum L. *Biochimie* **2015**, *109*, 36–41. [CrossRef] [PubMed]
23. Korbes, A.P.; Machado, R.D.; Guzman, F.; Almerao, M.P.; de Oliveira, L.F.; Loss-Morais, G.; Turchetto-Zolet, A.C.; Cagliari, A.; dos Santos Maraschin, F.; Margis-Pinheiro, M.; et al. Identifying conserved and novel microRNAs in developing seeds of Brassica napus using deep sequencing. *PLoS ONE* **2012**, *7*, e50663. [CrossRef] [PubMed]
24. Riechmann, J.L.; Heard, J.; Martin, G.; Reuber, L.; Jiang, C.; Keddie, J.; Adam, L.; Pineda, O.; Ratcliffe, O.J.; Samaha, R.R.; et al. Arabidopsis transcription factors: Genome-wide comparative analysis among eukaryotes. *Science* **2000**, *290*, 2105–2110. [CrossRef] [PubMed]
25. Rhoades, M.W.; Reinhart, B.J.; Lim, L.P.; Burge, C.B.; Bartel, B.; Bartel, D.P. Prediction of plant microRNA targets. *Cell* **2002**, *110*, 513–520. [CrossRef]
26. Xie, K.; Wu, C.; Xiong, L. Genomic organization, differential expression, and interaction of SQUAMOSA promoter-binding-like transcription factors and microRNA156 in rice. *Plant Physiol.* **2006**, *142*, 280–293. [CrossRef]
27. Salinas, M.; Xing, S.; Hohmann, S.; Berndtgen, R.; Huijser, P. Genomic organization, phylogenetic comparison and differential expression of the SBP-box family of transcription factors in tomato. *Planta* **2012**, *235*, 1171–1184. [CrossRef]
28. Jones-Rhoades, M.W.; Bartel, D.P.; Bartel, B. MicroRNAS and their regulatory roles in plants. *Annu. Rev. Plant Biol.* **2006**, *57*, 19–53. [CrossRef]
29. Voinnet, O. Origin, biogenesis, and activity of plant microRNAs. *Cell* **2009**, *136*, 669–687. [CrossRef]

30. Chuck, G.; Cigan, A.M.; Saeteurn, K.; Hake, S. The heterochronic maize mutant Corngrass1 results from overexpression of a tandem microRNA. *Nat. Genet.* **2007**, *39*, 544–549. [CrossRef]
31. Wu, G.; Poethig, R.S. Temporal regulation of shoot development in Arabidopsis thaliana by *miR156* and its target SPL3. *Development* **2006**, *133*, 3539–3547. [CrossRef] [PubMed]
32. Duguid, S.D.; Kenaschuk, E.O.; Rashid, K.Y. Macbeth flax. *Can. J. Plant Sci.* **2003**, *83*, 803–805. [CrossRef]
33. Mackowiak, S.D. Identification of novel and known miRNAs in deep-sequencing data with miRDeep2. *Curr. Protoc. Bioinform.* **2011**, *36*. [CrossRef] [PubMed]
34. Zhang, M.; Li, L.; Liu, Y.; Gao, X. Effects of a Sudden Drop in Salinity on Immune Response Mechanisms of Anadara kagoshimensis. *Int. J. Mol. Sci.* **2019**, *20*, 4365. [CrossRef]
35. Chen, C.; Ridzon, D.A.; Broomer, A.J.; Zhou, Z.; Lee, D.H.; Nguyen, J.T.; Barbisin, M.; Xu, N.L.; Mahuvakar, V.R.; Andersen, M.R.; et al. Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Res.* **2005**, *33*, e179. [CrossRef]
36. Schmittgen, T.D.; Livak, K.J. Analyzing real-time PCR data by the comparative C(T) method. *Nat. Protoc.* **2008**, *3*, 1101–1108. [CrossRef]
37. Allen, E.; Xie, Z.; Gustafson, A.M.; Carrington, J.C. microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell* **2005**, *121*, 207–221. [CrossRef]
38. Hasan Nudin, N.F.; van Kronenburg, B.; Tinnenbroek, I.; Krens, F. The importance of salicylic acid and an improved plant condition in determining success in agrobacterium-mediated transformation. *Acta Hortic.* **2015**, *1087*, 65–69. [CrossRef]
39. Clough, S.J.; Bent, A.F. Floral dip: A simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant J.* **1998**, *16*, 735–743. [CrossRef]
40. Poirier, Y.; Ventre, G.; Caldelari, D. Increased flow of fatty acids toward beta-oxidation in developing seeds of Arabidopsis deficient in diacylglycerol acyltransferase activity or synthesizing medium-chain-length fatty acids. *Plant Physiol.* **1999**, *121*, 1359–1366. [CrossRef]

International Journal of
Molecular Sciences

MDPI

Article

A Tale of Two Families: Whole Genome and Segmental Duplications Underlie Glutamine Synthetase and Phosphoenolpyruvate Carboxylase Diversity in Narrow-Leafed Lupin (*Lupinus angustifolius* L.)

Katarzyna B. Czyż [1,*], Michał Książkiewicz [2], Grzegorz Koczyk [1], Anna Szczepaniak [2], Jan Podkowiński [3] and Barbara Naganowska [2]

[1] Department of Biometry and Bioinformatics, Institute of Plant Genetics, Polish Academy of Sciences, 60-479 Poznan, Poland; gkoc@igr.poznan.pl
[2] Department of Genomics, Institute of Plant Genetics, Polish Academy of Sciences, 60-479 Poznan, Poland; mksi@igr.poznan.pl (M.K.); bnag@igr.poznan.pl (B.N.)
[3] Department of Genomics, Institute of Bioorganic Chemistry, Polish Academy of Sciences, 61-704 Poznan, Poland
[*] Correspondence: kwyr@igr.poznan.pl

Received: 17 February 2020; Accepted: 6 April 2020; Published: 8 April 2020

Abstract: Narrow-leafed lupin (*Lupinus angustifolius* L.) has recently been supplied with advanced genomic resources and, as such, has become a well-known model for molecular evolutionary studies within the legume family—a group of plants able to fix nitrogen from the atmosphere. The phylogenetic position of lupins in Papilionoideae and their evolutionary distance to other higher plants facilitates the use of this model species to improve our knowledge on genes involved in nitrogen assimilation and primary metabolism, providing novel contributions to our understanding of the evolutionary history of legumes. In this study, we present a complex characterization of two narrow-leafed lupin gene families—glutamine synthetase (*GS*) and phosphoenolpyruvate carboxylase (*PEPC*). We combine a comparative analysis of gene structures and a synteny-based approach with phylogenetic reconstruction and reconciliation of the gene family and species history in order to examine events underlying the extant diversity of both families. Employing the available evidence, we show the impact of duplications on the initial complement of the analyzed gene families within the genistoid clade and posit that the function of duplicates has been largely retained. In terms of a broader perspective, our results concerning *GS* and *PEPC* gene families corroborate earlier findings pointing to key whole genome duplication/triplication event(s) affecting the genistoid lineage.

Keywords: Fabaceae; *Lupinus*; glutamine synthetase (*GS*); phosphoenolpyruvate carboxylase (*PEPC*); phylogeny; evolution; gene families; duplication/triplication; structural genomics; genome organization; genome evolution

1. Introduction

The last decade has seen gradual progress in evolutionary studies on plants, mainly due to simultaneous, rapid advancement in theory, computing, and molecular technology. Legumes, which are the third largest plant family, have attracted the focus of active and collaborative international groups of researchers in the area of systematics and evolution [1–3]. Fabaceae, consisting of three major clades—Papilionoideae, Caesalpinioideae, and Mimosoideae—includes important grain, pasture, and agroforestry species that are characterized by an unusual flower structure, podded fruit, and the ability of most species to form nodules with rhizobia [4,5]. Recently, high-quality genome sequences of ten Fabaceae species have been published: *Arachis duranensis*, *Arachis ipaensis* [6], *Cajanus cajan* [7], *Cicer arietinum* [8], *Glycine max* [9], *Lotus japonicus* [10], *Lupinus angustifolius* [11], *Medicago truncatula* [12], *Phaseolus vulgaris* [13], and *Vigna radiata* [14].

Among legume species, due to their outstanding agronomic potential and complex evolutionary history, involving whole-genome duplication [15] and subsequent chromosome rearrangements, *L. angustifolius* has become an object of extensive molecular studies in terms of genomics, proteomics, and metabolomics. Altogether, several thousand molecular markers have been developed, including restriction fragment length polymorphisms (RFLPs), intron targeted amplified polymorphisms (ITAPs), amplified fragment length polymorphisms (AFLPs), molecular fragment length polymorphisms (MFLPs), single sequence repeats (SSRs), expressed sequence tags (ESTs), restriction site associated DNA markers (RADs), and EST-SSRs [16–19]. Reference genetic linkage maps carrying these markers have been built [17,20–22]. As a consequence, sequence-defined markers have been associated with major agronomic traits for this species, including soft seededness, anthracnose and *Phomopsis* stem blight resistance, pod shattering, vernalization requirement, and alkaloid content [16,18,23–27]. Two *L. angustifolius* nuclear genome bacterial artificial chromosome (BAC) libraries have been constructed and almost 15,000 BAC-end sequences have been obtained and annotated [28,29]. Selected BAC clones have been used as anchors for the integration of linkage groups in particular chromosomes by the molecular cytogenetic approach [30,31] and have served as material in evolutionary studies of the *Lupinus* genus [32,33]. Strong microsynteny in gene-rich regions between narrow-leafed lupin and other model legumes has also been observed [17,19,20,34,35]. Moreover, new evidence of widespread triplication within the *L. angustifolius* genome, possibly arising from a polyploidization event, has been found [11]. However, other duplication mechanisms, such as segmental duplications or chromosome additions, from related species cannot be ruled out [36].

Whole genome duplication/triplication and chromosomal rearrangements result in the multiplication of gene content within a particular genome. Gene pairs formed by duplication/triplication usually have a relatively short life span as, due to the relaxed selection constraints, some copies may be lost, others will be pseudogenized, and only a limited number will survive [31,37]. Various factors can alter the size of gene families [38–43]. Moreover, the relaxation of selective pressure may have created new developmental opportunities, conferred a selective advantage, and served as an engine for evolutionary changes [44]. Utilizing the explicit reconciliation of gene and species history [45], it is possible to elucidate the optimal sequence of duplication/speciation/loss events under a maximum parsimony framework, as well as derive the topological dating of key events in relation to the reference species tree [46,47]. Taken together, this allows for, as an example, the selection of likely orthologs for investigation as suitable taxonomic markers or for translational studies aimed at understanding neo/subfunctionalization in divergent species.

Taking into consideration the phylogenetic distances and the main characteristics of all legume plants, the most valuable sequences for genetic and evolutionary studies of Fabaceae belong to small gene families which originated early in the tree of life and participate in key enzymatic processes, such as genes encoding glutamine synthetase (GS). *GS* genes are considered to be among the oldest existing and functioning genes in the history of gene evolution [48]. GS is the key enzyme involved in the nitrogen metabolism of higher plants, catalyzing primary ammonium assimilation to form glutamine (GS1—cytosolic GS isoenzyme), as well as the reassimilation of ammonium released by a number of biochemical processes (such as photorespiration, protein catabolism, and deamination of amino acids), and is also related to storage protein accumulation in seeds (GS2—plastid GS isoenzyme) [49]. The central role of GS in nitrogen metabolism in all higher plants is unquestionable. The other gene important in legume evolutionary studies due to its functional correlation with *GS* genes may be phosphoenolpyruvate carboxylase (*PEPC*). PEPC plays a crucial role in the regulation of respiratory carbon flux in vascular plant tissues and green algae that actively assimilate nitrogen. The organic acids supplied by PEPC have several roles within nitrogen metabolism [50]. PEPC proteins are also encoded by a small multigene family with an insufficiently elucidated evolutionary history. However, it is assumed that gene duplication from pre-existing genes, followed by a few amino acid changes and the acquisition of a new gene transcription control, have led to the appearance of new isoforms such as C4 PEPC [51].

Here, we provide characterization of the *L. angustifolius* glutamine synthetase (*GS1* and *GS2*) and phosphoenolpyruvate carboxylase (*PEPC*) gene families, including gene structure determination; genetic localization within narrow-leafed lupin linkage groups (NLLs) and estimations of the *GS1*, *GS2*, and *PEPC* copy number in the narrow-leafed lupin genome. As sequences of narrow-leafed lupin [11,52] were only available in draft form prior to the start of this study, we decided to combine the screening of the BAC library with available data from genome sequencing. We also address several fundamental questions regarding the evolution of *GS* and *PEPC* gene families in legume plants and 40 other dicots and monocots. We support our evolutionary conclusions with a cross-genera microsynteny analysis of selected genome regions carrying particular *GS* and *PEPC* gene variants in the genomes of narrow-leafed lupin and several legume and non-legume species. Moreover, Fabaceae *GS* and *PEPC* representatives were sampled for selection pressure parameters by both pairwise and branch-site assays.

2. Results and Discussion

2.1. Narrow-Leafed Lupin GS and PEPC are Encoded by Multigene Families

To tag/select cytosolic *GS* and *PEPC* genes, two sequence-specific probes targeting *GS* and *PEPC* genes, respectively, were amplified and used for narrow-leafed lupin genome BAC library screening. As a result, two BAC clone sub-libraries were created, with BACs representing *L. angustifolius* genome regions carrying *GS* and *PEPC* genes. The presence of analyzed genes within selected BACs was positively verified by PCR amplification and Sanger sequencing with gene-specific primers. The similarity level between particular *GS* and *PEPC* homologs identified in the selected clones was determined. Fragments of analyzed genes (300–400 bp) with a similarity level above 97% were classified as one gene variant and assigned to one contig. Two such contigs and two singletons were constructed for the *GS* sub-library and two contigs with one singleton were constructed for *PEPC*. The composition of the *GS* sub-library is as follows: contig1, clones 015C08 and 087N22; contig2, clones 038E09, 047P22, 088E07, 094A04, and 131H20; and singletons, 036L23 and 059J08. The *PEPC* sub-library contains contig1, clones 067C07 and 083F23; contig2, clones 064J15 and 077K22; and a singleton, 131K15. Taking into consideration these results, the accuracy of BAC library screening with the use of the Southern blot method was calculated to be 50% for both sub-libraries and was considered as being relatively low. It was expected that post-hybridization signals would represent the coverage of the *L. angustifolius* genome in the BAC library [28,53]. The observed phenomenon may reflect the general characteristics of the lupin BAC library and incorporated cloning system used, with the noted instability depending on the carried sequence [28,54,55].

Gene copy number estimation with ddPCR revealed that BAC sub-libraries were lacking some gene duplicates. When the study was initially conceived and the experimental part was conducted, the lupin draft genome had not been officially released. Moreover, the availability of both the scaffold-level [52] genome draft and the latter LupinExpress pseudochromosome-level [11] assemblies has, in some of our other studies, failed to entirely resolve certain areas of the genome, including, for example, the placement of RAP2-7 transcription factor, crucial to alkaloid biosynthesis, reported by Kroc et al. (2019). Therefore, our recent BAC-based study aimed at molecular control of the vernalization response *Ku* locus in the narrow-leafed lupin highlighted a candidate gene (a homolog of FLOWERING LOCUS T) and provided the sequence of the domesticated allele carrying a functional mutation (large indel in the promoter) before the release of the lupin pseudochromosome sequence [25,30]. This finding was later confirmed by genome assembly-based studies. Furthermore, BAC clones may be used as chromosome-specific cytogenetic landmarks for chromosome-scale analysis, as well as for inter-species tracking of conserved chromosome regions and profiling of their structural variation. Both approaches have been used in lupin molecular cytogenetic studies [30–33]. Indeed, BAC clones from this study (047P22, 036L23, 059J08, 067C07, and 131K15) were recently exploited in parallel research addressing lupin karyotype evolution, providing single-locus anchors for the visualization of chromosomal rearrangements across the panel of ten European and African lupin species. Therefore, even after updating the bioinformatic results to include the newly available genomic data, we decided to retain BAC-derived sequences in the final analysis, both as a record of the train of thought and as valuable supporting evidence directly linking recently developed cytomolecular resources for comparative fluorescent in situ hybridization mapping.

To obtain data on *GS* and *PEPC* genes, sequences of interest were blasted against the narrow-leafed lupin annotated gene set cds v1.0. The search resulted in the identification of nine narrow-leafed lupin *GS* genes in total: seven *GS1* genes (named *GS1a1*, *GS1a2*, *GS1a3*, *GS1b1*, *GS1b2*, *GS1c1*, and *GS1c2*) and two *GS2* genes (named *GS2a1* and *GS2a2*). Nine *PEPC* homologs were identified: *PEPC1a*, *PEPC1b*, *PEPC1c*, *PEPC2a*, *PEPC2b*, *PEPC3a*, *PEPC3b*, *PEPC4*, and *PEPC5* (Table 1). The observed trend in the *L. angustifolius GS* and *PEPC* gene copy number is consistent with the data gathered for other legumes. The *P. vulgaris GS1* gene family contains three active *GS1* genes and one pseudogene [56]. In pea, three active *GS1* genes have been characterized: *GS1*, *GS3A*, and *GS3B* [57]. In *M. truncatula*, two active *GS1* genes (*MtGS1a* and *MtGS1b*), two *GS2* genes *(MtGS2a* and *MtGS2b)*, and one pseudogene (*MtGSc*) were revealed [58]. Two major classes of *GS1* genes have been investigated in *M. sativa* [59]. In the *G. max* genome, there are three *GS1* classes, each represented by at least two functional members [60]. Only one copy of the *GS1* gene was identified in the *A. ipaensi*s and *A. duranensis* species.

According to the proposed evolutionary history of narrow-leafed lupin, it was stated that this species has undergone duplication and/or triplication with several chromosome rearrangements [11,21,36]. Based on a cytogenetic analysis of several species from the *Lupinus* genus, it was also hypothesized that the lupin karyotype has evolved through polyploidy and subsequent aneuploidy [61]. Global analysis of the narrow-leafed lupin transcriptome and legume genome sequence comparative mapping enabled whole genome duplication (WGD) events to be dated. Hane et al. estimated the Papilionoideae radiation at 58 mya with genistoid lineage separation from the other Papilionoideae legumes at 54.6 mya, followed by whole-genome triplication in the genistoid lineage at 24.6 mya [11]. Additionally, the ancient polyploidy event has been confirmed based on an analysis of several genes, such as chalcone isomerases (*CHI*) [62], phosphatidylethanolamine binding proteins (*PEBP*) [30], isoflavone synthetases (*IFS*) [63], and cytosolic and plastid acetyl-coenzyme A carboxylases (*ACCase*) [64]. All listed genes are present in the narrow-leafed lupin genome in multiple variants and evolved by WGDs, evidenced by shared synteny and Bayesian phylogenetic inference. Our results concerning *GS* and *PEPC* gene families support the whole genome duplication/triplication(s) hypothesis.

Table 1. Characterization of *Lupinus angustifiolius* bacterial artificial chromosomes (BACs)/scaffolds carrying glutamine synthetase (*GS*) and phosphoenolpyruvate carboxylase (*PEPC*) sequences, including anchoring genes to the scaffolds and narrow-leafed lupin linkage groups (NLLs), cytogenetic marker representation, and repetitive content analysis within selected scaffolds. NLL—narrow-leafed lupin linkage group, RE—repetitive element, and CDS—coding sequence.

Gene Variant	Gene ID		BAC nb	Scaffold nb	NLL nb	Cyto marker	GC%	% RE	RE (bp)	RE type	CDS nb
	Lupin Express ID	GenBank ID									
GS1a1	**Lup21297**	**gene6261**	**047P22**	**4_25**	**4**	**047P22_5**	**33.1**	**8.58**	8584	Ty1/Copia	12
GS1a2	Lup001512	gene27466	087N22	106	16	087N22_2	32.94	15.63	15635	TY1/Copia; Gypsy/DIRS1; DNA transposons	10
GS1a3	Lup009916	gene24502	-	192	14	-	36.11	10.54	9282	Ty1/Copia; Gypsy/DIRS1; DNA transpozons	17
GS1b1	Lup029429	gene 19431	036L23	73	11	036L23_3	33	0	0	-	15
GS1b2	Lup032636	gene17555	059J08	94_15	9	059J08_3	32.43	0.17	174	Ty1/Copia	16
GS1c1	Lup002132	gene34907	-	11_68	UN	-	30.56	9.19	2621	Ty1/Copia; DNA transposons	3
GS1c2	Lup04581	gene4422	-	13	3	-	31.89	7.2	7202	Ty1/Copia; DNA transposons	15
GS2a1	Lup023221	gene31805	-	45_213	19	-	33.88	9.96	9963	Ty1/Copia; Gypsy/DIRS1; DNA transpozons	12
GS2a2	no	gene6462	-	186	4	-	32.66	7.73	7732	Ty1/Copia; Gypsy/DIRS1; DNA transpozons	12
PEPC1a	Lup022696	gene23490	064J15	437	13	-	34.25	8.85	8852	Ty1/Copia; Gypsy/DIRS1	13
PEPC1b	Lup029825	gene15450	-	74_10	8	-	32.28	1.88	1879	Ty1/Copia	13
PEPC1c	Lup015178	gene12998	-	274	7	-	32.36	1.17	1169	Ty1/Copia	14
PEPC2a	Lup002214	gene31196	067C07	110_41	19	067C07_2	32.41	3.63	3634	Ty1/Copia; Gypsy/DIRS1	14
PEPC2b	Lup26946	gene9184	131K15	59_19	5	131K15_5_3	33.21	6.49	6748	Ty1/Copia; Gypsy/DIRS1	14
PEPC3a	Lup031846	gene18605	-	9_1	10	-	33.97	5.17	1628	Ty1/Copia	3
PEPC3b	Lup016482	gene7147	-	296	4	-	32.76	15.64	15641	Ty1/Copia; DNA transposon	5
PEPC4	Lup002996	no	-	12_32	7	-	33.76	8.64	8644	Ty1/Copia; Gypsy/DIRS1	16
PEPC5	Lup031638		-	88_60	20	-	32.73	8.93	8933	Ty1/Copia; Gypsy/DIRS1; DNA transposons	10

2.2. GS and PEPC Gene Variants are Localized in Distinct Narrow-leafed Lupin Genome Regions

All identified representatives of *GS* and *PEPC* gene families, originating from BACs and in silico genome analyses, were mapped against narrow-leafed lupin genome assembly v1.0, revealing their localization within the analyzed genome. *GS1a1*, *GS1a2*, *GS1a3*, *GS1b1*, *GS1b2*, and *GS1c2* were assigned to narrow-leafed lupin pseudochromosomes (NLL-04, NLL-16, NLL-14, NLL-11, NLL-09, and NLL-03, respectively), whereas *GS1c1* was assigned to unlinked scaffold11_68. *GS2a1* and *GS2a2* were localized in NLL-19 and NLL-04, respectively. The physical distance between two NLL-04 *GS* genes—*GS1a1* and *GS2a2*—was calculated as approximately 3 Mbp. *PEPC* genes were allocated to nine different NLL pseudochromosomes, as follows: *PEPC1a* to NLL-13, *PEPC1b* to NLL-08, *PEPC1c* to NLL-07, *PEPC2a* to NLL-19, *PEPC2b* to NLL-05, *PEPC3a* to NLL-10, *PEPC3b* to NLL-04, *PEPC4* to NLL-7, and *PEPC5* to NLL-20. Employing the BAC-based results and including those obtained in our previous studies, we provide genomic localization for all identified *GS* and *PEPC* gene variants, as well as the cytogenetic position of four *GS1* and two *PEPC* gene copies in lupin chromosomes. The described gene variants correspond to chromosome-specific cytogenetic markers [31], as follows: *GS1a1*, 047P22_5; *GS1a2*, 087N22_2; *GS1b1*, 036L22_3; *GS1b2*, 059J08_3; *PEPC2a*, 067C07_2; and *PEPC2b*, 131K15_5_3 (Table 1).

In order to resolve the organization of multiple genome regions carrying distinct sequence variants of *GS* and *PEPC*, narrow-leafed lupin genome regions carrying these genes were extracted from the assembly and, together with seven sequenced BAC clone inserts (three with the *PEPC* genes 064J15, 067C07, and 131K15, and four with the *GS* sequences 036L23, 047P22, 059J08, and 087N22), were annotated with putative functions. BAC sequences were mapped onto narrow-leafed lupin scaffolds and selected regions were truncated into a uniform length of 100 Mbp. Four scaffolds remained with the original lengths: scaffold192, 88,054 bp; scaffold11_68, 28,507 bp; scaffold9_1, 31,494 bp; and scaffold59_19, 103,921 bp. Analysis revealed the average GC content of 32.95% and 33.23% for *GS* and *PEPC* regions, respectively. The observed occurrence of repetitive elements in genome fragments flanking *GS* and *PEPC* gene variants varied from 0% (*GS1b1*, scaffold73) to 15.63% (*GS1a2*, scaffold106), and from 1.17% (*PEPC1c*, scaffold274) to 15.64% (*PEPC3b*, scaffold296), with retrotransposons (Ty1/Copia and Gypsy/DIRS1) and DNA transposons (DNA/Mule-MuDR type) being the most abundant.

It has been confirmed that the narrow-leafed lupin genome is highly repetitive (57%) [11], with well-organized gene-rich regions. In addition to satellites sensu lato, long terminal repeat (LTR) retrotransposons and DNA transposons were revealed as the most common, with only a small proportion of non-coding RNA [11,19,31,65]. Due to the “copy and paste” mechanism underlying the amplification of LTR retrotransposons, they have been shown to make up the largest classes of transposable element (TE) content in the genomes of most flowering plants, greatly contributing to increases in size of their host genome [66]. As reported in studies concerning *Arabidopsis*, soybean, and flax genomes, *Copia* elements are largely located within and/or close to gene-coding regions, which suggests that these elements may have the dominant influence on the evolution of some gene families [67–69]. Gene prediction revealed features characteristic of gene-rich regions, with an average of 13 coding sequences per 100 Mbp for both *GS* and *PEPC* gene regions (Table 1, Supplementary file 1). The obtained data for the frequency of coding sequences within analyzed regions of the narrow-leafed lupin genome showed a lower coding sequence (CDS) abundance than in our previous studies [19,31]. This low number of genes in *GS1a2*, *GS2a1*, and *PEPC3b* neighborhoods is primarily due to the high content of repetitive elements in the surrounding regions.

2.3. GS and PEPC Gene Variants Present a Conserved Sequence Structure among All L. angustifolius Homologs and Other Legume Species

To investigate the structural changes of the *GS* and *PEPC* genes, sequence data from 46 species originating from 26 plant families were gathered (Supplementary file 2). In total, 244 sequences of *GS* homologs were subjected to exon/intron determination. The average CDS length for *GS1* (178 sequences analyzed) was established as 3259 bp, with 12 exons as the dominant structure, whereas for *GS2*

(46 sequences analyzed), the value was 3866 bp, with 13 exons. Legume GS homologs (36 sequences of *GS1* and *GS2*) presented a conserved gene structure consistent with the pattern described above. Indeed, only the structures of four *GS1* genes were different: Lj0g3v0335159 from *L. japonicus*—nine exons; TR_5g077950 from *M. truncatula*—nine exons; gene13764 (LOC107631250) from *A. ipaensis*—10 exons; and GLYMA02G41106 from *G. max*—10 exons. In the case of *GS2* homologs, only gene3699 (LOC107637831) from *A. ipaensis* with 14 exons and Tp57577_TGAC_v2_gene28916 from *Trifolium pratense* with 20 exons showed an atypical gene structure (Supplementary file 3).

To establish the structure of *PEPC* gene family representatives among higher plants, 223 sequences were analyzed. Based on the exon/intron organization, two groups were formed. The first group contained 167 sequences with an average length of 5645 bp (min. 3102 bp, max. 17,375 bp) structured into 10 exons. Nevertheless, some variation in exon composition was found, particularly in the sequences GSMUA_Achr9G06420_001 from *Musa acuminata* and MDP0000258440 from *Malus domestica*, consisting of 17 and 19 exons, respectively. The second group carried 57 sequences with an average length of 9268 bp (min. 4144 bp, max. 26,587 bp), mainly organized into 18–24 exons (mode value 20). Sixty-four sequences originating from the Fabaceae family presented very low variation in sequence organization. Only MTR_8g463920 and MTR_0002s0890 from *M. truncatula*, gene 1498 (LOC101500264) and gene 3089 (LOC101497901) from *C. arietinum*, and Tp57577_TGAC_v2_gene11496 from *T. pratense* showed differences in the gene structures (Supplementary file 3).

The structures of all identified *L. angustifolius GS* and *PEPC* genes were established. The *GS* sequence lengths varied from 3550 to 8730 for *GS1* homologs and from 4002 to 4890 bp for *GS2*. Coding sequence organization was highly conserved within *GS1* (12 exons) and *GS2* (13 exons) groups, despite the observed dissimilarities in lengths. CDS lengths were as follows: *GS1a*, 1071 bp (356 aa); *GS1b1*, 1071 bp (356 aa); *GS1b2*, 1062 bp (353 aa); *GS1c*, 1074 bp (357 aa); and *GS2a1* and *GS2a2*, 1299 bp (432 aa). A major structural difference in *GS* genes was observed for *GS1b2*, where exon number 12 was significantly shorter than in other homologs (144 vs. 153 bp, respectively). Moreover, 5′ and 3′ *GS* regulatory regions revealed high variation between all analyzed sequences, both in length and composition. *PEPC* genes were organized into 10 exons, and the coding sequence length varied from 2901 to 2907 bp (from 966 to 968 aa), excluding *PEPC5*, which had a 3135 bp (1044 aa) length structured into 20 exons and thus significantly deviated from the other *PEPC* sequence variants. The observed level of sequence similarities within the *PEPC* clade is considered as being high. However, major differences in the length and composition of 5′ and 3′UTR regions were noted (Supplementary file 4).

2.4. The Initial GS and PEPC Complement was Subsequently Duplicated in a Lineage-Specific Manner and Can be Traced to the Common Ancestor of Legumes

The reconstructed phylogeny of plant *GS* genes yielded several insights with regards to legume enzymes. Firstly, the initial representation of this family in Fabaceae is shown to have consisted of three ancestral clades (Figure 1, Figure 2, and Supplementary file 5) for a simplified phylogenetic tree of relationships. The first monophyletic clade (denoted as *GS2*—Table 2) encompasses the known types of *GS2* loci, which are annotated as chloroplastic proteins encoded in the nuclear genome. Duplicates are present in multiple, rather than singular, cases of divergent legumes and were previously found to be expressed in seeds, at least in the case of *M. truncatula* [70]. The other two clades (*GS1cs1* and *GS1cs2*) carry genes encoding cytosolic proteins corresponding to cytosolic isoforms preferentially expressed in different organs/at different developmental stages (i.e., *GS1cs2*—α, and *GS1cs1*—β and γ subunits described in early comparative analyses [71]). The placement of *Vitis vinifera* and multiple malvid sequences between the two clades points to the *GS1cs1*/*GS1cs2* ancestral split either coinciding or shortly following the γ triplication common to both rosids and asterids [72]. Additionally, the *GS1cs1* ancestral split, which resulted in the separation of β and γ subclades (constitutively expressed vs. nodule enhanced, respectively), is shown to have occurred early in the evolution of legumes (possibly prior to the separation of genistoid lineage, with *GS1cs1-β* encoding loci seemingly not having been retained in the NLL reference genome).

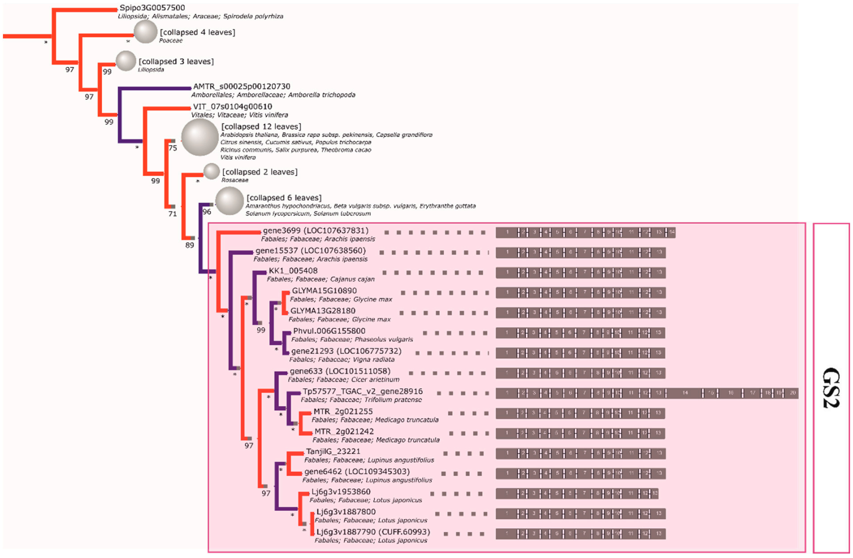

Spipo3G0057500
Liliopsida; Alismatales; Araceae; Spirodela polyrhiza
[collapsed 4 leaves]
Poaceae
[collapsed 3 leaves]
Liliopsida
AMTR_s00025p00120730
Amborellales; Amborellaceae; Amborella trichopoda
VIT_07s0104g00610
Vitales; Vitaceae; Vitis vinifera
[collapsed 12 leaves]
Arabidopsis thaliana, Brassica rapa subsp. pekinensis, Capsella grandiflora
Citrus sinensis, Cucumis sativus, Populus trichocarpa
Ricinus communis, Salix purpurea, Theobroma cacao
Vitis vinifera
[collapsed 2 leaves]
Rosaceae
[collapsed 6 leaves]
Amaranthus hypochondriacus, Beta vulgaris subsp. vulgaris, Erythranthe guttata
Solanum lycopersicum, Solanum tuberosum
gene3699 (LOC107637831)
Fabales; Fabaceae; Arachis ipaensis
gene15537 (LOC107638560)
Fabales; Fabaceae; Arachis ipaensis
KK1_005408
Fabales; Fabaceae; Cajanus cajan
GLYMA15G10890
Fabales; Fabaceae; Glycine max
GLYMA13G28180
Fabales; Fabaceae; Glycine max
Phvul.006G155800
Fabales; Fabaceae; Phaseolus vulgaris
gene21293 (LOC106775732)
Fabales; Fabaceae; Vigna radiata
gene633 (LOC101511058)
Fabales; Fabaceae; Cicer arietinum
Tp57577_TGAC_v2_gene28916
Fabales; Fabaceae; Trifolium pratense
MTR_2g021255
Fabales; Fabaceae; Medicago truncatula
MTR_2g021242
Fabales; Fabaceae; Medicago truncatula
TanjilG_23221
Fabales; Fabaceae; Lupinus angustifolius
gene6462 (LOC109345303)
Fabales; Fabaceae; Lupinus angustifolius
Lj6g3v1953860
Fabales; Fabaceae; Lotus japonicus
Lj6g3v1887800
Fabales; Fabaceae; Lotus japonicus
Lj6g3v1887790 (CUFF.60993)
Fabales; Fabaceae; Lotus japonicus
GS2

Figure 1. The reconstructed phylogeny of plant plastid GS isoenzyme (*GS2*) genes. A collapsed phylogeny tree was used in order to highlight Fabaceae family relationships.

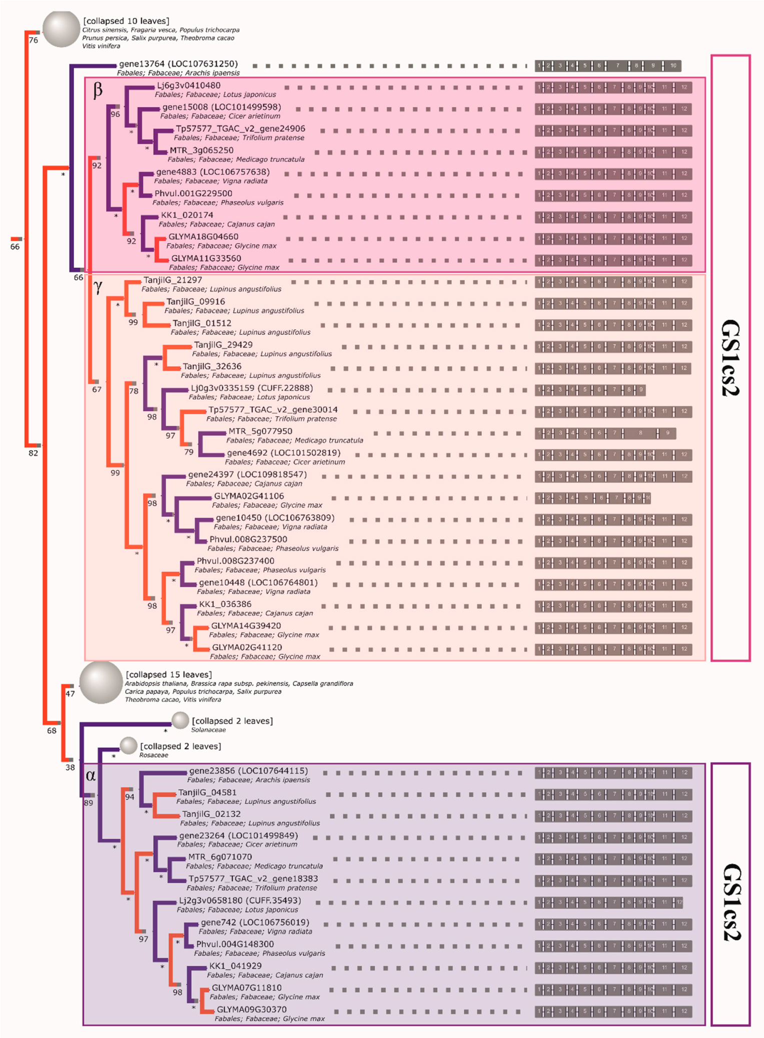

[collapsed 10 leaves]
Citrus sinensis, Fragaria vesca, Populus trichocarpa
Prunus persica, Salix purpurea, Theobroma cacao
Vitis vinifera
gene13764 (LOC107631250)
Fabales; Fabaceae; Arachis ipaensis
β
Lj6g3v0410480
Fabales; Fabaceae; Lotus japonicus
gene15008 (LOC101499598)
Fabales; Fabaceae; Cicer arietinum
Tp57577_TGAC_v2_gene24906
Fabales; Fabaceae; Trifolium pratense
MTR_3g065250
Fabales; Fabaceae; Medicago truncatula
gene4883 (LOC106757638)
Fabales; Fabaceae; Vigna radiata
Phvul.001G229500
Fabales; Fabaceae; Phaseolus vulgaris
KK1_020174
Fabales; Fabaceae; Cajanus cajan
GLYMA18G04660
Fabales; Fabaceae; Glycine max
GLYMA11G33560
Fabales; Fabaceae; Glycine max
γ
TanjilG_21297
Fabales; Fabaceae; Lupinus angustifolius
TanjilG_09916
Fabales; Fabaceae; Lupinus angustifolius
TanjilG_01512
Fabales; Fabaceae; Lupinus angustifolius
TanjilG_29429
Fabales; Fabaceae; Lupinus angustifolius
TanjilG_32636
Fabales; Fabaceae; Lupinus angustifolius
Lj0g3v0335159 (CUFF.22888)
Fabales; Fabaceae; Lotus japonicus
Tp57577_TGAC_v2_gene30014
Fabales; Fabaceae; Trifolium pratense
MTR_5g077950
Fabales; Fabaceae; Medicago truncatula
gene4692 (LOC101502819)
Fabales; Fabaceae; Cicer arietinum
gene24397 (LOC109818547)
Fabales; Fabaceae; Cajanus cajan
GLYMA02G41106
Fabales; Fabaceae; Glycine max
gene10450 (LOC106763809)
Fabales; Fabaceae; Vigna radiata
Phvul.008G237500
Fabales; Fabaceae; Phaseolus vulgaris
Phvul.008G237400
Fabales; Fabaceae; Phaseolus vulgaris
gene10448 (LOC106764801)
Fabales; Fabaceae; Vigna radiata
KK1_036386
Fabales; Fabaceae; Cajanus cajan
GLYMA14G39420
Fabales; Fabaceae; Glycine max
GLYMA02G41120
Fabales; Fabaceae; Glycine max
GS1cs2
[collapsed 15 leaves]
Arabidopsis thaliana, Brassica rapa subsp. pekinensis, Capsella grandiflora
Carica papaya, Populus trichocarpa, Salix purpurea
Theobroma cacao, Vitis vinifera
[collapsed 2 leaves]
Solanaceae
[collapsed 2 leaves]
Rosaceae
α
gene23856 (LOC107644115)
Fabales; Fabaceae; Arachis ipaensis
TanjilG_04581
Fabales; Fabaceae; Lupinus angustifolius
TanjilG_02132
Fabales; Fabaceae; Lupinus angustifolius
gene23264 (LOC101499849)
Fabales; Fabaceae; Cicer arietinum
MTR_6g071070
Fabales; Fabaceae; Medicago truncatula
Tp57577_TGAC_v2_gene18383
Fabales; Fabaceae; Trifolium pratense
Lj2g3v0658180 (CUFF.35493)
Fabales; Fabaceae; Lotus japonicus
gene742 (LOC106756019)
Fabales; Fabaceae; Vigna radiata
Phvul.004G148300
Fabales; Fabaceae; Phaseolus vulgaris
KK1_041929
Fabales; Fabaceae; Cajanus cajan
GLYMA07G11810
Fabales; Fabaceae; Glycine max
GLYMA09G30370
Fabales; Fabaceae; Glycine max
GS1cs2

Figure 2. The reconstructed phylogeny of plant cytosolic GS isoenzyme (*GS1*) genes. A collapsed phylogeny tree was used in order to highlight Fabaceae family relationships.

Both one *PEPC2* (PTPC, plant-type PEPC [50]) clade and two *PEPC1* (BTPC, bacterial-type PEPC) clades can be clearly characterized as monophyletic in legumes. Therefore, three ancestral genes inherited from an early rosid are indicated, each of which was duplicated prior to the divergence of genistoid/dalbergioid lines and can be traced to the common ancestor of legumes (*PEPC1a*, *PEPC1b*, *PEPC2*—see Table 3 for a full summary and Figure 3, Figure 4, and Supplementary file 6 for relevant fragments of phylogenetic reconstruction). The ancestral duplication giving rise to *PEPC1a* and *PEPC1b* legume plant-type *PEPC* subgroups likely dates back to core eudicots (coincident with γ triplication or closely following the event). An additional legume-specific duplication event is implied in *PEPC1b*, although incomplete lineage sorting artefacts cannot be ruled out. Indeed, as with available reconstructions of legume phylogeny based on housekeeping genes, the ordering of early diverging dalbergioid and genistoid lineages is seen to alternate between two possibilities.

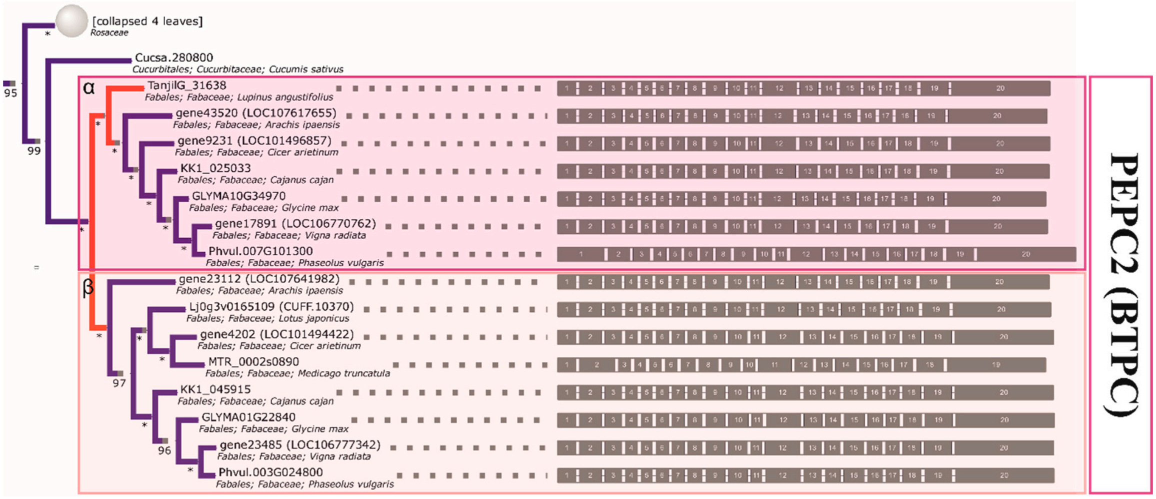

[collapsed 4 leaves]
Rosaceae
Cucsa.280800
Cucurbitales; Cucurbitaceae; Cucumis sativus
95
99
α
TanjilG_31638
Fabales; Fabaceae; Lupinus angustifolius
gene43520 (LOC107617655)
Fabales; Fabaceae; Arachis ipaensis
gene9231 (LOC101496857)
Fabales; Fabaceae; Cicer arietinum
KK1_025033
Fabales; Fabaceae; Cajanus cajan
GLYMA10G34970
Fabales; Fabaceae; Glycine max
gene17891 (LOC106770762)
Fabales; Fabaceae; Vigna radiata
Phvul.007G101300
Fabales; Fabaceae; Phaseolus vulgaris
β
gene23112 (LOC107641982)
Fabales; Fabaceae; Arachis ipaensis
Lj0g3v0165109 (CUFF.10370)
Fabales; Fabaceae; Lotus japonicus
gene4202 (LOC101494422)
Fabales; Fabaceae; Cicer arietinum
MTR_0002s0890
Fabales; Fabaceae; Medicago truncatula
97
KK1_045915
Fabales; Fabaceae; Cajanus cajan
GLYMA01G22840
Fabales; Fabaceae; Glycine max
96
gene23485 (LOC106777342)
Fabales; Fabaceae; Vigna radiata
Phvul.003G024800
Fabales; Fabaceae; Phaseolus vulgaris
PEPC2 (BTPC)

Figure 3. The reconstructed phylogeny of plant *PEPC* genes. A collapsed phylogeny tree was used in order to highlight Fabaceae family relationships.

The initial *GS1* complement was subsequently duplicated in a lineage-specific manner and available evidence (including intact intron-exon structure, which is prior published evidence in the case of alfalfa and common bean) indicates that the functionality of these duplicates has been largely retained in extant crop species. In regard to lupin, the narrow-leafed lupin enzymes are shown to be the result of such duplications and are thus paralogous to the closest counterparts from non-genistoid groups. As a closing side note, the overall resolution of events on the basis of the phylogeny (evolution of cytosolic GSI-encoding genes) suggests that monocot family members might be more ancient than dicot ones, stemming from the selective culling of duplicates predating the separation of both lineages (in line with the split between cytosolic and plastid eukaryotic GS, likely predating monocot/dicot divergence) [48]. However, it is worth noting that the resolution of these basal events lacks the support necessary to make strong inferences (less than 50% bootstrap probabilities for consensual clades).

Analogous to the *GS* case, most of the retained *PEPC* duplications are late and species-specific (as seen in the soybean, lotus, and lupin genomes). In this case, the reconciled *PEPC* phylogeny supports most lupin gene family members being late paralogs (*PEPC1a.2* and *PEPC1b.1*—single duplication, and *PEPC1b.2*—either two rounds of duplication and loss or triplication in the lineage). The inference of possible subsequent duplications/triplication (both here and in the GS1cs1 γ clade) corroborates the earlier findings, pointing to events affecting the genistoid lineage [36].

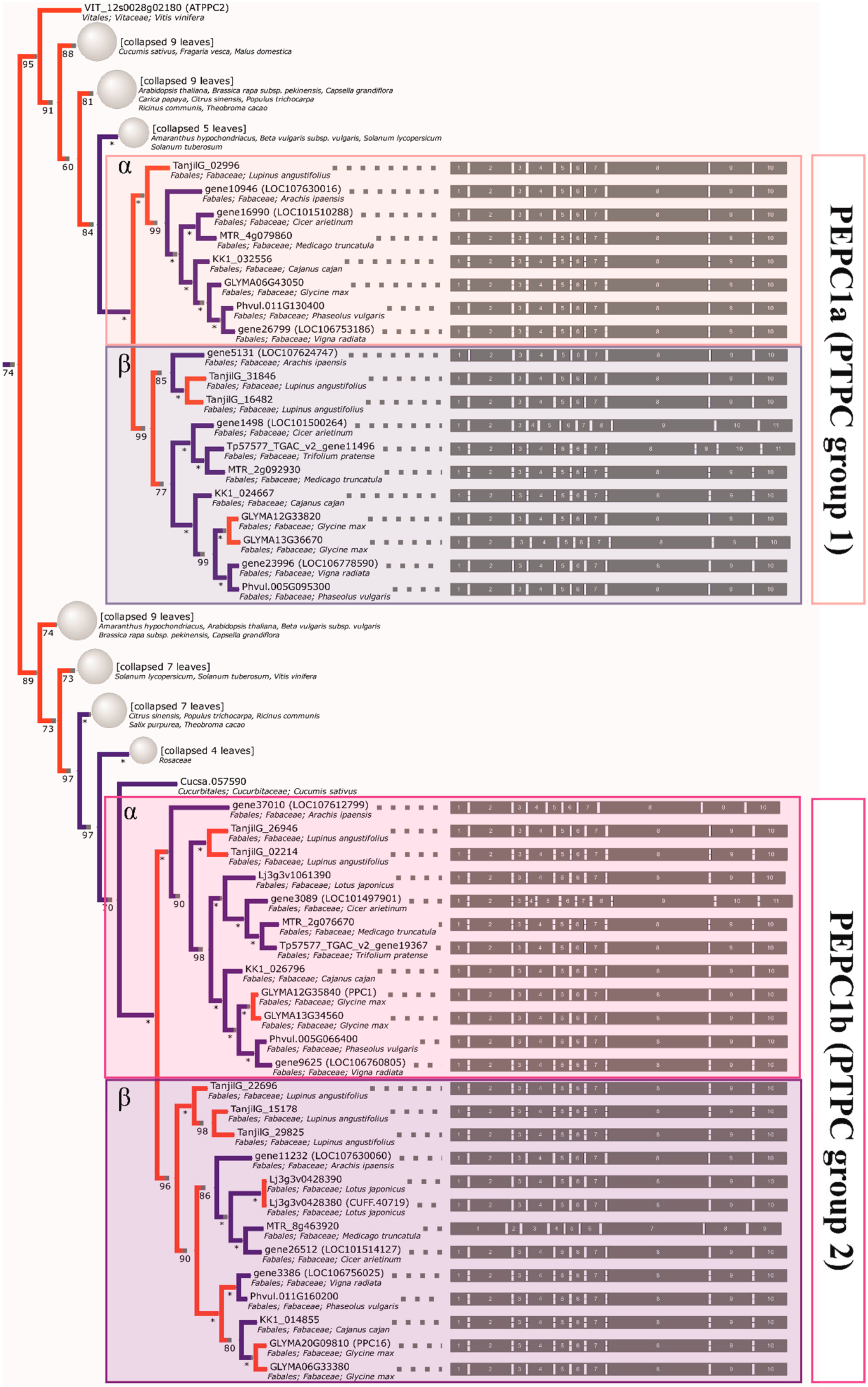

VIT_12s0028g02180 (ATPPC2)
Vitales; Vitaceae; Vitis vinifera
[collapsed 9 leaves]
Cucumis sativus, Fragaria vesca, Malus domestica
[collapsed 9 leaves]
Arabidopsis thaliana, Brassica rapa subsp. pekinensis, Capsella grandiflora
Carica papaya, Citrus sinensis, Populus trichocarpa
Ricinus communis, Theobroma cacao
[collapsed 5 leaves]
Amaranthus hypochondriacus, Beta vulgaris subsp. vulgaris, Solanum lycopersicum
Solanum tuberosum
α
TanjilG_02996
Fabales; Fabaceae; Lupinus angustifolius
gene10946 (LOC107630016)
Fabales; Fabaceae; Arachis ipaensis
gene16990 (LOC101510288)
Fabales; Fabaceae; Cicer arietinum
MTR_4g079860
Fabales; Fabaceae; Medicago truncatula
KK1_032556
Fabales; Fabaceae; Cajanus cajan
GLYMA06G43050
Fabales; Fabaceae; Glycine max
Phvul.011G130400
Fabales; Fabaceae; Phaseolus vulgaris
gene26799 (LOC106753186)
Fabales; Fabaceae; Vigna radiata
β
gene5131 (LOC107624747)
Fabales; Fabaceae; Arachis ipaensis
TanjilG_31846
Fabales; Fabaceae; Lupinus angustifolius
TanjilG_16482
Fabales; Fabaceae; Lupinus angustifolius
gene1498 (LOC101500264)
Fabales; Fabaceae; Cicer arietinum
Tp57577_TGAC_v2_gene11496
Fabales; Fabaceae; Trifolium pratense
MTR_2g092930
Fabales; Fabaceae; Medicago truncatula
KK1_024667
Fabales; Fabaceae; Cajanus cajan
GLYMA12G33820
Fabales; Fabaceae; Glycine max
GLYMA13G36670
Fabales; Fabaceae; Glycine max
gene23996 (LOC106778590)
Fabales; Fabaceae; Vigna radiata
Phvul.005G095300
Fabales; Fabaceae; Phaseolus vulgaris
PEPC1a (PTPC group 1)
[collapsed 9 leaves]
Amaranthus hypochondriacus, Arabidopsis thaliana, Beta vulgaris subsp. vulgaris
Brassica rapa subsp. pekinensis, Capsella grandiflora
[collapsed 7 leaves]
Solanum lycopersicum, Solanum tuberosum, Vitis vinifera
[collapsed 7 leaves]
Citrus sinensis, Populus trichocarpa, Ricinus communis
Salix purpurea, Theobroma cacao
[collapsed 4 leaves]
Rosaceae
Cucsa.057590
Cucurbitales; Cucurbitaceae; Cucumis sativus
α
gene37010 (LOC107612799)
Fabales; Fabaceae; Arachis ipaensis
TanjilG_26946
Fabales; Fabaceae; Lupinus angustifolius
TanjilG_02214
Fabales; Fabaceae; Lupinus angustifolius
Lj3g3v1061390
Fabales; Fabaceae; Lotus japonicus
gene3089 (LOC101497901)
Fabales; Fabaceae; Cicer arietinum
MTR_2g076670
Fabales; Fabaceae; Medicago truncatula
Tp57577_TGAC_v2_gene19367
Fabales; Fabaceae; Trifolium pratense
KK1_026796
Fabales; Fabaceae; Cajanus cajan
GLYMA12G35840 (PPC1)
Fabales; Fabaceae; Glycine max
GLYMA13G34560
Fabales; Fabaceae; Glycine max
Phvul.005G066400
Fabales; Fabaceae; Phaseolus vulgaris
gene9625 (LOC106760805)
Fabales; Fabaceae; Vigna radiata
β
TanjilG_22696
Fabales; Fabaceae; Lupinus angustifolius
TanjilG_15178
Fabales; Fabaceae; Lupinus angustifolius
TanjilG_29825
Fabales; Fabaceae; Lupinus angustifolius
gene11232 (LOC107630060)
Fabales; Fabaceae; Arachis ipaensis
Lj3g3v0428390
Fabales; Fabaceae; Lotus japonicus
Lj3g3v0428380 (CUFF.40719)
Fabales; Fabaceae; Lotus japonicus
MTR_8g463920
Fabales; Fabaceae; Medicago truncatula
gene26512 (LOC101514127)
Fabales; Fabaceae; Cicer arietinum
gene3386 (LOC106756025)
Fabales; Fabaceae; Vigna radiata
Phvul.011G160200
Fabales; Fabaceae; Phaseolus vulgaris
KK1_014855
Fabales; Fabaceae; Cajanus cajan
GLYMA20G09810 (PPC16)
Fabales; Fabaceae; Glycine max
GLYMA06G33380
Fabales; Fabaceae; Glycine max
PEPC1b (PTPC group 2)

Figure 4. The reconstructed phylogeny of plant *PEPC* genes. A collapsed phylogeny tree was used in order to highlight Fabaceae family relationships.

Table 2. Summary of major glutamine synthetase clades traced to the ancestral legume genome (monophyletic, support over 90%).

GS Subset	Legume Clade	Taxon	Locus Tag (NCBI: Gene Locus ID [1])
GS2	dalbergioids	*Arachis ipaensis*	gene15537 (LOC107638560), gene3699 (LOC107637831)
	genistoids	*Lupinus angustifolius*	TanjilG_23221, gene6462 (LOC109345303)
	IRLC	*Cicer arietinum*	gene633 (LOC101511058)
		Medicago truncatula	MTR_2g021242, MTR_2g021255
		Trifolium pratense	Tp57577_TGAC_v2_gene28916
	milletioids	*Cajanus cajan*	KK1_005408
		Glycine max	GLYMA13G28180, GLYMA15G10890
		Phaseolus vulgaris	Phvul.006G155800
		Vigna radiata	gene21293 (LOC106775732)
	robinioids	*Lotus japonicus*	Lj6g3v1887790 (CUFF.60993), Lj6g3v1887800, Lj6g3v1953860
GS1cs1	dalbergioids	*Arachis ipaensis*	gene13764 (LOC107631250)
	genistoids	*Lupinus angustifolius*	TanjilG_32636, TanjilG_29429, TanjilG_09916, TanjilG_01512, TanjilG_21297
	IRLC	*Cicer arietinum*	gene15008 (LOC101499598), gene4692 (LOC101502819)
		Medicago truncatula	MTR_3g065250, MTR_5g077950
		Trifolium pratense	Tp57577_TGAC_v2_gene24906, Tp57577_TGAC_v2_gene30014
	milletioids	*Cajanus cajan*	gene24397 (LOC109818547), KK1_036386, KK1_020174
		Glycine max	GLYMA02G41106, GLYMA02G41120, GLYMA11G33560, GLYMA14G39420, GLYMA18G04660
		Phaseolus vulgaris	Phvul.001G229500, Phvul.008G237400, Phvul.008G237500
		Vigna radiata	gene10448 (LOC106764801), gene10450 (LOC106763809), gene4883 (LOC106757638)
		Lotus japonicus	Lj0g3v0335159 (CUFF.22888), Lj6g3v0410480
GS1cs2	dalbergioids	*Arachis ipaensis*	gene23856 (LOC107644115)
	genistoids	*Lupinus angustifolius*	TanjilG_02132, TanjilG_04581
	IRLC	*Cicer arietinum*	gene23264 (LOC101499849)
		Medicago truncatula	MTR_6g071070
		Trifolium pratense	Tp57577_TGAC_v2_gene18383
	milletioids	*Cajanus cajan*	KK1_041929
		Glycine max	GLYMA07G11810, GLYMA09G30370
		Phaseolus vulgaris	Phvul.004G148300
		Vigna radiata	gene742 (LOC106756019)
	robinioids	*Lotus japonicus*	Lj2g3v0658180 (CUFF.35493)

[1] Where a locus tag is not available (gene designated as the NCBI reannotation only), the NCBI Gene database ID is given in the parentheses, prefixed with LOC.

Table 3. Summary of major phosphoenolpyruvate carboxylase clades traced to the ancestral legume genome (monophyletic, support over 90%).

PEPC Subset	Legume Clade	Taxon	Locus Tag (NCBI:Gene Locus ID [1])
PEPC1a	dalbergioids	*Arachis ipaensis*	gene10946 (LOC107630016), gene5131 (LOC107624747)
	genistoids	*Lupinus angustifolius*	TanjilG_02996, TanjilG_31846, TanjilG_16482
	IRLC	*Cicer arietinum*	gene1498 (LOC101500264), gene16990 (LOC101510288)
		Medicago truncatula	MTR_2g092930, MTR_4g079860
		Trifolium pratense	Tp57577_TGAC_v2_gene11496
	milletioids	*Cajanus cajan*	KK1_024667, KK1_032556
		Glycine max	GLYMA06G43050, GLYMA12G33820, GLYMA13G36670
		Phaseolus vulgaris	Phvul.005G095300, Phvul.011G130400
		Vigna radiata	gene23996 (LOC106778590), gene26799 (LOC106753186)
PEPC1b	dalbergioids	*Arachis ipaensis*	gene11232 (LOC107630060), gene37010 (LOC107612799)
	genistoids	*Lupinus angustifolius*	TanjilG_15178, TanjilG_29825, TanjilG_22696, TanjilG_02214, TanjilG_26946
	IRLC	*Cicer arietinum*	gene26512 (LOC101514127), gene3089 (LOC101497901)
		Medicago truncatula	MTR_2g076670, MTR_8g463920
		Trifolium pratense	Tp57577_TGAC_v2_gene19367
	milletioids	*Cajanus cajan*	KK1_014855, KK1_026796
		Glycine max	GLYMA06G33380, GLYMA12G35840 (PPC1), GLYMA13G34560, GLYMA20G09810 (PPC16)
		Phaseolus vulgaris	Phvul.005G066400, Phvul.011G160200
		Vigna radiata	gene3386 (LOC106756025), gene9625 (LOC106760805)
	robinioids	*Lotus japonicus*	Lj3g3v0428380 (CUFF.40719), Lj3g3v0428390, Lj3g3v1061390
PEPC2	dalbergioids	*Arachis ipaensis*	gene23112 (LOC107641982), gene43520 (LOC107617655)
	genistoids	*Lupinus angustifolius*	TanjilG_31638
	IRLC	*Cicer arietinum*	gene4202 (LOC101494422), gene9231 (LOC101496857)
		Medicago truncatula	MTR_0002s0890
	milletioids	*Cajanus cajan*	KK1_025033, KK1_045915
		Glycine max	GLYMA01G22840, GLYMA10G34970
		Phaseolus vulgaris	Phvul.003G024800, Phvul.007G101300
		Vigna radiata	gene17891 (LOC106770762), gene23485 (LOC106777342)
	robinioids	*Lotus japonicus*	Lj0g3v0165109 (CUFF.10370)

[1] Where a locus tag is not available (gene designated as the NCBI reannotation only), the NCBI Gene database ID is given in the parentheses, prefixed with LOC.

2.5. Compared to GS Genes, the History of Coding Sequences of PEPC Genes More Closely Recapitulates the History of Species

A maximum likelihood codon-based phylogenetic species tree of 46 reference plant genomes, based on 29 putative single-copy orthologs with the best coverage and uniqueness, was generated in order to track species evolution. The obtained species phylogeny (Figure 5) is highly supported, with only two major differences from the accepted consensus (e.g., The Angiosperm Phylogeny Group 2016). One of these is the alliance of lycopod *Selaginella* and moss *Physcomitrella*. The grouping of these lineages is likely an artefact of rapid diversification in early land plant lineages and could be observed in PEPC/GS phylogenies. Additionally, a significant observed difference is the grouping of *Citrus sinensis* (malvid, order *Sapindales*) with representatives of the rosid order *Malpighiales* (*Ricinus communis*, *Populus trichocarpa*, and *Salix purpurea*). Notably, the phylogeny of the latter order has still not been entirely resolved, with the whole COM (*Celastrales*, *Oxalidales*, and *Malpighiales*) clade placement in rosids being challenged by different datasets [73]. Otherwise, the support for consensus topology is strong and the relationships, in particular the topology of the legume clade, support the earlier consensus [74,75].

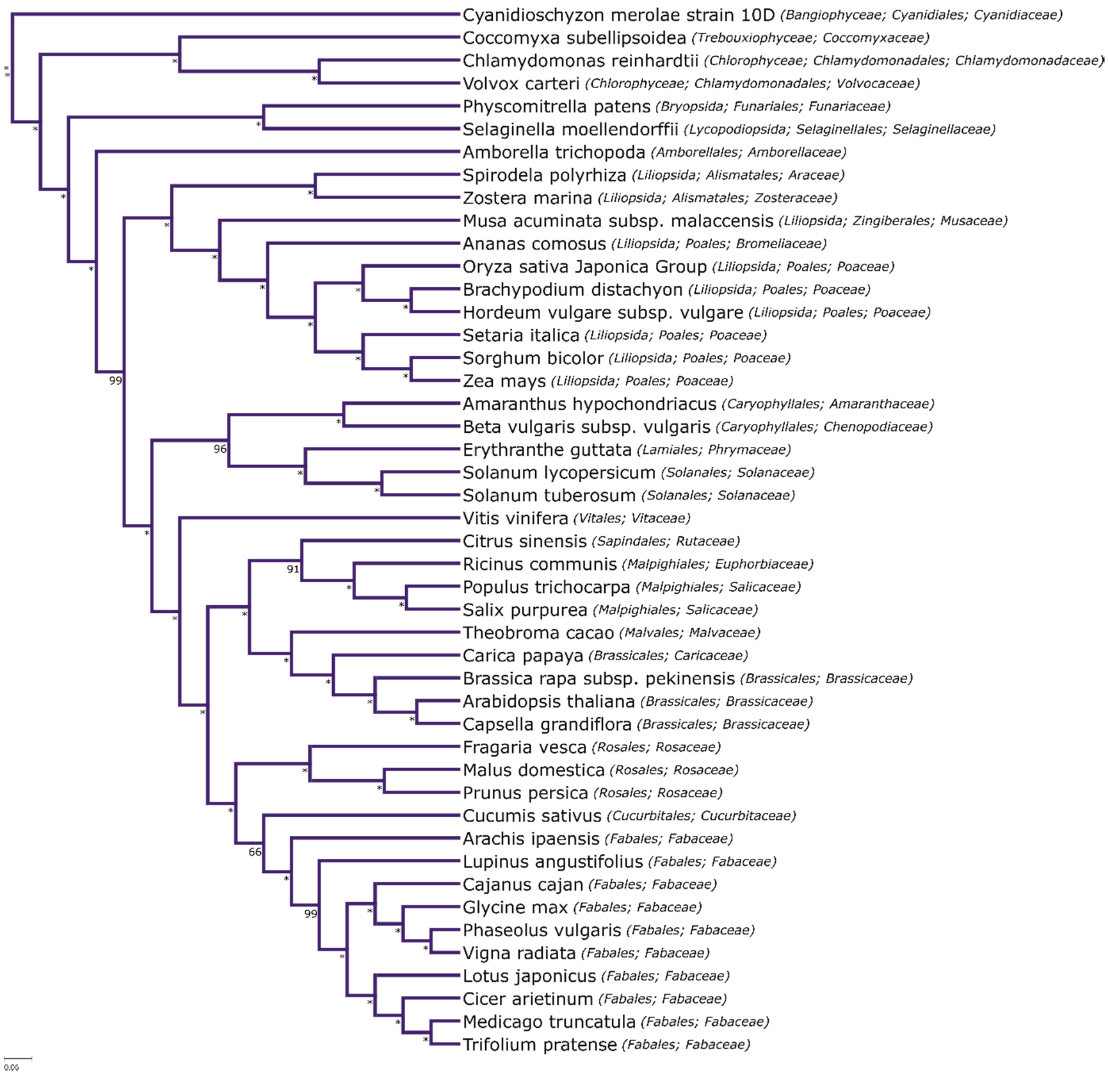

Cyanidioschyzon merolae strain 10D (Bangiophyceae; Cyanidiales; Cyanidiaceae)
Coccomyxa subellipsoidea (Trebouxiophyceae; Coccomyxaceae)
Chlamydomonas reinhardtii (Chlorophyceae; Chlamydomonadales; Chlamydomonadaceae)
Volvox carteri (Chlorophyceae; Chlamydomonadales; Volvocaceae)
Physcomitrella patens (Bryopsida; Funariales; Funariaceae)
Selaginella moellendorffii (Lycopodiopsida; Selaginellales; Selaginellaceae)
Amborella trichopoda (Amborellales; Amborellaceae)
Spirodela polyrhiza (Liliopsida; Alismatales; Araceae)
Zostera marina (Liliopsida; Alismatales; Zosteraceae)
Musa acuminata subsp. malaccensis (Liliopsida; Zingiberales; Musaceae)
Ananas comosus (Liliopsida; Poales; Bromeliaceae)
Oryza sativa Japonica Group (Liliopsida; Poales; Poaceae)
Brachypodium distachyon (Liliopsida; Poales; Poaceae)
Hordeum vulgare subsp. vulgare (Liliopsida; Poales; Poaceae)
Setaria italica (Liliopsida; Poales; Poaceae)
Sorghum bicolor (Liliopsida; Poales; Poaceae)
Zea mays (Liliopsida; Poales; Poaceae)
Amaranthus hypochondriacus (Caryophyllales; Amaranthaceae)
Beta vulgaris subsp. vulgaris (Caryophyllales; Chenopodiaceae)
Erythranthe guttata (Lamiales; Phrymaceae)
Solanum lycopersicum (Solanales; Solanaceae)
Solanum tuberosum (Solanales; Solanaceae)
Vitis vinifera (Vitales; Vitaceae)
Citrus sinensis (Sapindales; Rutaceae)
Ricinus communis (Malpighiales; Euphorbiaceae)
Populus trichocarpa (Malpighiales; Salicaceae)
Salix purpurea (Malpighiales; Salicaceae)
Theobroma cacao (Malvales; Malvaceae)
Carica papaya (Brassicales; Caricaceae)
Brassica rapa subsp. pekinensis (Brassicales; Brassicaceae)
Arabidopsis thaliana (Brassicales; Brassicaceae)
Capsella grandiflora (Brassicales; Brassicaceae)
Fragaria vesca (Rosales; Rosaceae)
Malus domestica (Rosales; Rosaceae)
Prunus persica (Rosales; Rosaceae)
Cucumis sativus (Cucurbitales; Cucurbitaceae)
Arachis ipaensis (Fabales; Fabaceae)
Lupinus angustifolius (Fabales; Fabaceae)
Cajanus cajan (Fabales; Fabaceae)
Glycine max (Fabales; Fabaceae)
Phaseolus vulgaris (Fabales; Fabaceae)
Vigna radiata (Fabales; Fabaceae)
Lotus japonicus (Fabales; Fabaceae)
Cicer arietinum (Fabales; Fabaceae)
Medicago truncatula (Fabales; Fabaceae)
Trifolium pratense (Fabales; Fabaceae)
99
96
91
66
99
0.06

Figure 5. Maximum likelihood codon-based phylogenetic species tree of 46 reference plant genomes, based on 29 putative single-copy orthologs.

Primary metabolism genes were frequently good candidates for molecular taxonomic markers, provided that paralogy was taken into account and suitable low/single copy orthologs were chosen for inference [76]. In this context, the members of *GS* and *PEPC* subfamilies were considered as good candidates in the past. Our results do not fully corroborate these findings.

Contrary to early inquiries [4,77], chloroplastic glutamate synthetases are not particularly good taxonomic markers for legumes. The *GS* phylogeny clearly confirms the existence of multiple, functional copies and the reconstructed ancestry contains both late duplications (*L. angustifolius*, *M. truncatula*, *L. japonicus*, and *G. max*) and traces of earlier events (e.g., positioning *L. angustifolius* sequences, which implies early duplications). From the point of view of future studies, *PEPC* clades provide better candidates for supplementary markers (bacterial-type *PEPC* sequences from clades α and β), as there are less duplications and the phylogenetic signal is strong (as exemplified by the bootstrap support of inner bipartitions). This is supported by past findings demonstrating that WGD may have played a lesser role in the evolution of the *PEPC* family in land plants [78]. However, in all (recent) cases, paralogy should be taken into account (e.g., by targeting UTR regions in order to distinguish paralogs).

More interestingly, the general patterns of lineage-specific duplications suggest that sub-functionalization and/or regulatory rewiring played a large role in shaping the extant carbon and nitrogen primary metabolic pathways in some lineages (*L. angustifolius*, *L. japonicus*, and *G. max*). This is also corroborated by the conserved gene structure and further analyses of selection pressure, which show a lack of changes in core ligand-interacting residues of the encoded proteins. Taken together, the evidence points to regulatory rather than mechanistic changes driving the diversification of both *GS* and *PEPC* family members. Whether this is a result of the differential retention of functional duplicates or different frequency of events, the outcome remains pertinent for future translational/comparative studies of legumes and merits more investigation.

2.6. L. angustifolius Genome Regions Carrying GS and PEPC Genes Arise from Duplication/Triplication with Additional Complex Chromosome Rearrangements

Lupinus angustifolius genome regions carrying all identified variants of *GS* and *PEPC* genes were subjected to comparative mapping to nine well-defined legume genome assemblies. Several patterns of sequence collinearity in these loci were identified. In particular, a high level of microsynteny was observed for the region carrying *GS1a1* and *A. duranensis* chromosome 3 (122.31 Mbp), *A. ipaensis* chromosome 3 (122.88 Mbp), *C. arietinum* chromosome 6 (0.61 Mbp), *C. cajan* chromosome 1 (4.3 Mbp), *G. max* chromosomes 11 (30.88 Mbp) and 18 (3.47 Mbp), *L. japonicus* chromosome 6 (3.75 Mbp), *M. truncatula* chromosome 3 (2.94 Mbp), *P. vulgaris* chromosome 1 (49.04 Mbp), and *V. radiata* chromosome 3 (9.32 Mbp). All these regions carry (at least) one copy of the *GS1* sequence. The narrow-leafed lupin region containing gene *GS1a2* revealed collinearity links to the same regions as those characterized for *GS1a1*, suggesting the occurrence of lineage-specific duplication. A more complex pattern was observed for *GS1b1* and *GS1b2* regions. Well-preserved sequence collinearities of these regions to loci at *A. duranensis* chromosome 7 (14.10 Mbp), *A. ipaensis* chromosome 7 (15.23 Mbp), and *C. cajan* chromosome 2 (8.45 Mbp), which do not carry any (even considerably truncated) *GS* gene sequences, were observed. This may indicate that some *GS1b* gene copies were eliminated during the evolution of these species. Moreover, two *GS1b* sequence variants matched one region of *V. radiata* chromosome 6 (7.14 Mbp), *P. vulgaris* chromosome 8 (55.14 Mbp), and *G. max* chromosomes 2 (43.20 Mbp) and 14 (47.82 Mbp) with a high level of sequence similarity. These regions encode *GS* sequences. *GS1c1* regions did not reveal conserved synteny among any of the species analyzed, only showing alignments between *GS* gene sequences. *GS1c2* regions yielded high collinearity alignments to loci carrying corresponding *GS* sequences at *A. duranensis* chromosome 5 (96.66 Mbp), *A. ipaensis* chromosome 5 (129.41 Mbp), *C. arietinum* chromosome 8 (11.79 Mbp), *C. cajan* scaffold 132405, *G. max* chromosomes 7 (10.08 Mbp) and 9 (39.77 Mbp), *L. japonicus* chromosome 2 (10.53 Mbp), *M. truncatula* chromosome 6 (26.24 Mbp), *P. vulgaris* chromosome 4 (42.89 Mbp), and *V. radiata* chromosome 1 (8.22 Mbp).

In the case of *GS2* regions, clear evidence of sequence collinearity was observed in all analyzed legumes: *A. duranensis* chromosomes 1 (97.70 Mbp) and 4 (3.66 Mbp), *A. ipaensis* chromosomes 1 (128.24 Mbp) and 4 (4.93 Mbp), *C. arietinum* chromosome 1 (4.92 Mbp), *C. cajan* chromosome 2 (8.45 Mbp), *G. max* chromosomes 13 (32.46 Mbp) and 15 (7.96 Mbp), *L. japonicus* chromosome 6 (20.97 Mbp), *M. truncatula* chromosome 2 (7.20 Mbp), *P. vulgaris* chromosome 6 (26.87 Mbp), and *V. radiata* chromosome 10 (16.06 Mbp).

To summarize, all legume regions carrying at least one copy of the *GS* gene revealed shared synteny (Figure 6) to at least one narrow-leafed lupin region carrying a corresponding homologous copy. Some of them matched duplicated regions in the narrow-leafed lupin genome located on different chromosomes and carrying different homologous gene copies, providing clear evidence of ancient duplications of chromosome segments that did not result in the further elimination of additional gene copies.

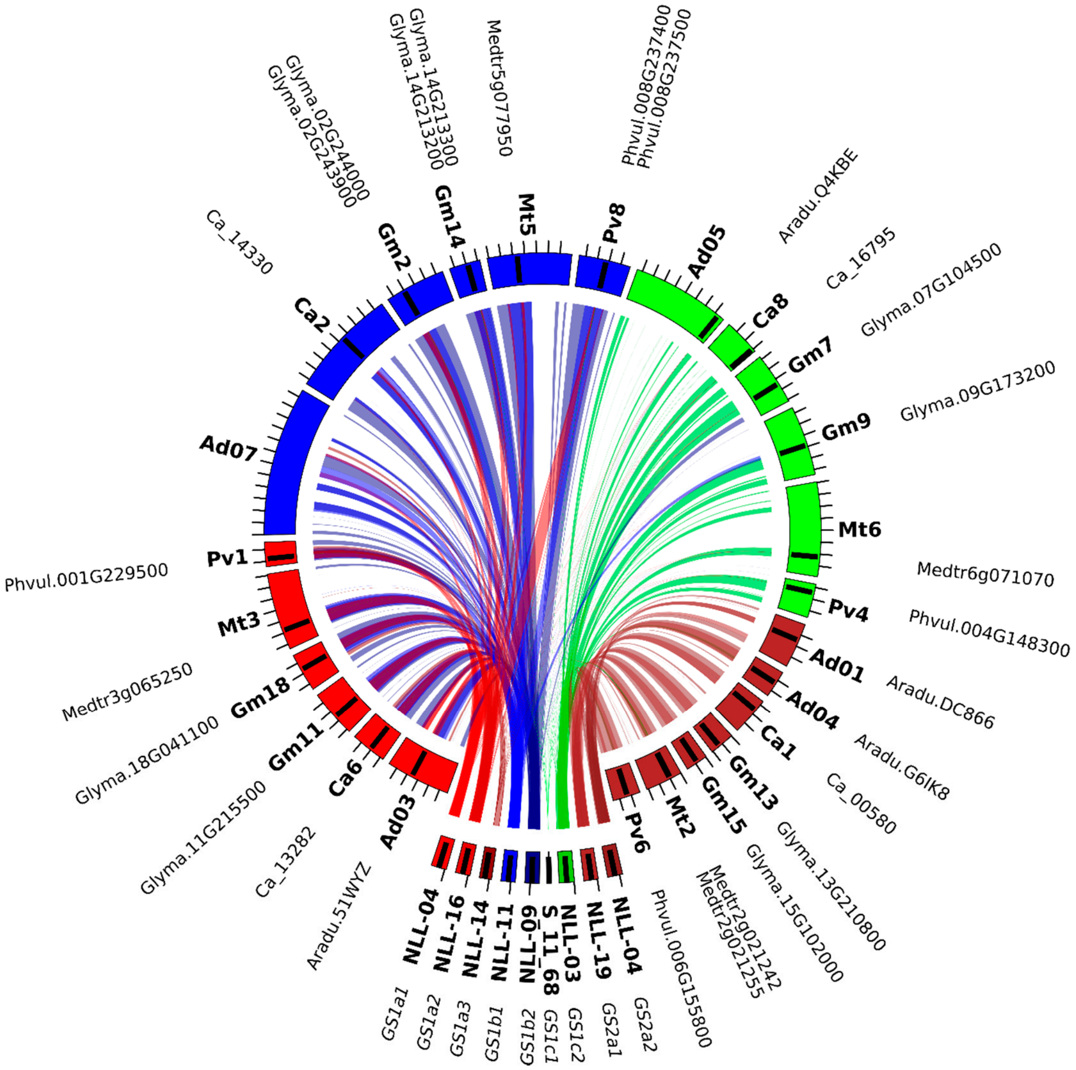


Figure 6. Collinearity links matching narrow-leafed lupin linkage groups and the legume reference genome carrying *GS* genes. NLL—narrow-leafed lupin linkage group, Pv—*P. vulgaris*, Mt—*M. truncatula*, Gm—*G. max*, Ca—*C. arietinum*, and Ad—*A. duranensis*.

The set of legume regions carrying *PEPC* genes had more complex patterns of collinearity links. Two types of syntenic relationship were observed, related to regions carrying a *PEPC* gene and to regions lacking such a gene. Moreover, numerous local duplications in the analyzed data set were revealed. Highly conserved microsynteny, expressed by high values of the total score of sequence alignments, was observed for *PEPC1a*, *PEPC1b*, *PEPC1c*, and *A. duranensis* chromosomes 3 (26.99 Mbp) and 7 (72.62 Mbp); *A. ipaensis* chromosomes 3 (29.54 Mbp) and 8 (27.74 Mbp); *C. arietinum* chromosome 1 (47.88 Mbp) and scaffold 1545; *C. cajan* chromosome 10 (12.46 Mbp) and scaffold 380; *G. max* chromosomes 6 (35.35 Mbp), 12 (29.90 and 38.94 Mbp), and 13 (37.24 Mbp); *L. japonicus* chromosome 3 (3.90 and 14.19 Mbp); *M. truncatula* chromosomes 2 (32.09 Mb) and 8 (22.56 Mbp); *P. vulgaris* chromosomes 5 (10.24 and 19.00 Mbp) and 11 (42.06 Mbp); and *V. radiata* chromosomes 2 (16.95 and 21.00 Mbp) and 5 (35.59 Mbp). All these regions carry *PEPC* gene sequences. *PEPC2a* and *PEPC2b* revealed high collinearity links to the same legume genome regions as *PEPC1a*, *PEPC1b*, and *PEPC1c*. *PEPC3a*, *PEPC3b*, and *PEPC4* genes showed conserved synteny to regions carrying *PEPC* homologs located at *A. duranensis* chromosomes 3 (21.75 Mbp) and 8 (33.03 Mbp); *A. ipaensis* chromosomes 2 (7.18 Mbp) and 3 (24.05 Mbp); *C. arietinum* chromosomes 1 (12.63 Mbp) and 6 (21.32 Mbp); *C. cajan* scaffolds 293 and 330; *G. max* chromosomes 6 (46.94 Mbp), 12 (36.95 Mbp), and 13 (39.10 Mbp); *M. truncatula* chromosome 4 (30.90 Mbp); *P. vulgaris* chromosomes 5 (28.48 Mbp) and 11 (29.34 Mbp); and *V. radiata* scaffold 23. The *PEPC4* region also had highly conserved synteny to some regions lacking *PEPC* sequences, namely *A. ipaensis* chromosome 8 (12.40 Mbp), *C. cajan* chromosome 4 (9.64 Mbp), *G. max* chromosome 12 (13.31 Mbp), *L. japonicus* chromosome 3 (35.10 Mbp), and *V. radiata* scaffold 149. This may suggest that *PEPC4* gene copies were removed from these regions during evolution. In the *PEPC5* region, no microsynteny was found between lupin and other legumes. Nevertheless, several orthologs of *PEPC5* were described. In general, *PEPC* genes revealed complex patterns of microsynteny, indicating both lineage-specific and ancestral duplications, as well as possible deletions of excessive gene copies (Figure 7, Supplementary file 7). The distribution of collinearity links provided a clear line of evidence that both *GS* and *PEPC* gene families have expanded in legumes through segmental duplications, which may be considered as landmarks of two ancient WGD events.

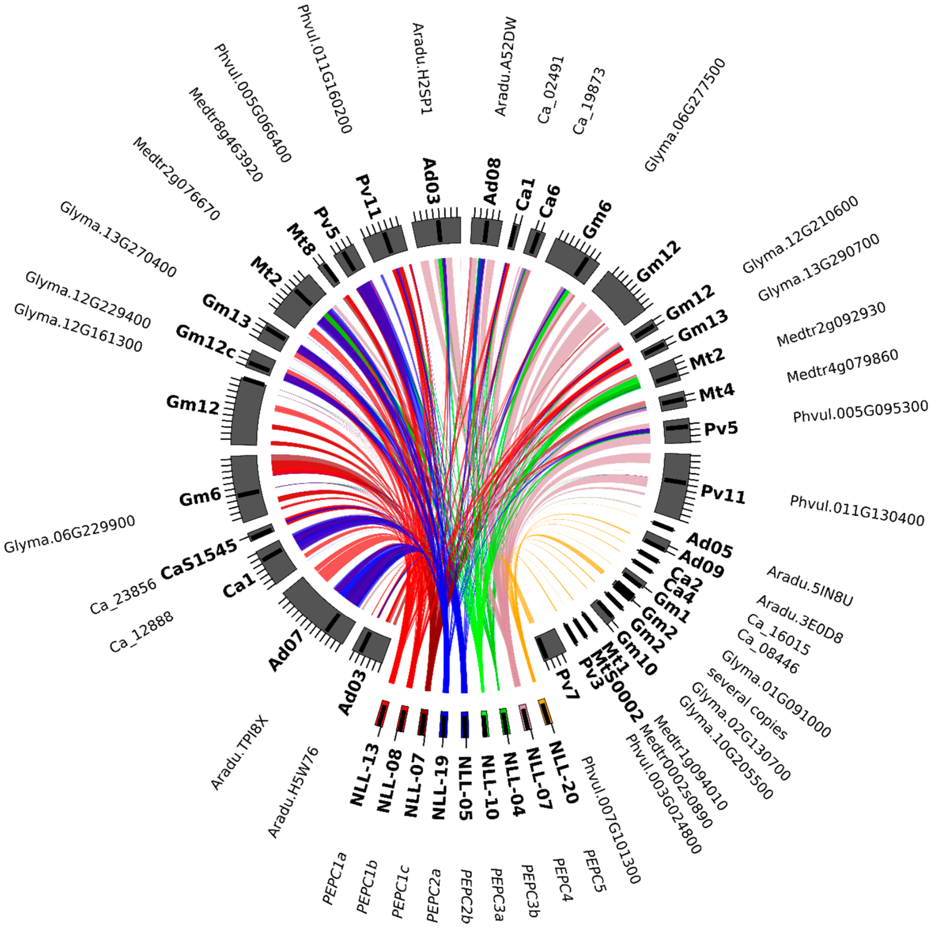


Figure 7. Comparative mapping and phylogenetic inference of legume *PEPC* genes. Syntenic patterns revealed for narrow-leafed lupin genome regions and corresponding regions of legume chromosomes. NLL—narrow-leafed lupin linkage group, Pv—*P. vulgaris*, Mt—*M. truncatula*, Gm—*G. max*, Ca—*C. arietinum*, and Ad—*A. duranensis*.

2.7. The Major Events Promoting the Evolution of GS and PEPC Genes in Legumes were Whole-Genome Duplications

It is a well-accepted hypothesis that the evolution of legumes has been driven by an ancient WGD event which putatively occurred in the progenitor line of Papilionoideae about 50–65 mya, providing the tetraploid ancestor and launching the divergence of ancient lineages of Papilionoideae [3,8,75,79,80]. Traces of that event have been identified in numerous clades spanning the legume tree of life, from Xanthocercis and Cladrastis through dalbergioids (*Arachis* spp.) and genistoids (e.g., *L. angustifolius*), to more recent lineages of millettioids (*P. vulgaris*, *G. max*, *C. cajan*, and *V. radiata*) and galegoids (*M. truncatula*, *L. japonicus*, and *C. arietinum*) [1,3,74,80,81]. Some species have retained relatively large numbers of ancient tetraploid regions (i.e., 309 regions in *M. truncatula* carrying 4198 genes or 343 regions in *G. max* with 9486 genes). Taking into consideration the topology of the legume *GS1c1* tree, this ancestral duplication might have contributed to the origin of β and γ subclades. A similar explanation might be proposed for the emergence of α and β groups of *PEPC1a*, *PEPC1b*, and *PEPC2*, supported by both phylogenetic inference and the synteny-based approach. However, the lack of genome sequencing data for early diverging legumes hampers such a comprehensive comparative analysis and precludes drawing firm conclusions.

During the early divergence of some downstream lineages, dated to roughly ~30–55 mya, additional independent WGD events probably occurred, affecting Mimosoideae-Cassiinae-Caesalpinieae, Detarieae, Cercideae, and *Lupinus* clades [75]. Large-scale duplication and/or triplication in the *L. angustifolius* genome has been well-evidenced by recent studies involving linkage and comparative mapping [17,36] and microsynteny analysis of selected gene families [30,31,34,62,63]. These WGD events apparently contributed to multiplication of the gene copy number of *L. angustifolius GS* and *PEPC* genes because hypothetical duplicates were found in sister branches of the phylogenetic tree and the genome regions harboring these genes shared common collinearity links. Some lineages experienced WGD events relatively recently, including soybean (~13 mya), carrying numerous genes in the duplicated state [3,9]. All *GS* and *PEPC* subclades, except for *PEPC1a-α*, were shown to carry hypothetical survivors of such an event. Hypothetical legume tandem duplicates were only identified in the *GS* family: in *P. vulgaris*, *V. radiata*, and *G. max* for *GS1cs1* and *L. japonicus* and *M. truncatula* for *GS2*. This is an expected outcome, as tandem duplication has been suggested to be a typical mechanism for the expansion of genes, representing flexible steps in the biochemical pathways or located at the end of pathways, where they do not affect many downstream genes [82]. *GS* and *PEPC* are genes encoding key enzymes involved in crucial metabolic pathways. Therefore, the appearance of additional copies without duplication of the whole pathway might have been selected against by evolutionary processes. On the contrary, the WGD event copies the entire molecular machinery, enabling the further evolution and divergence of redundant networks [83]. Moreover, the type of duplication contributes to the further evolutionary fate, demonstrated by different gene expression patterns and the methylation status of duplicates [84]. A recent expression quantitative trait loci mapping study of an *L. angustifolius* recombinant inbred line population (83A:476 x P27255) provided leaf transcriptomic profiles for 30,595 genes, including all *GS* and *PEPC* homologs present in the genome, except *GS2a2* unannotated hitherto [85]. Gene expression values corresponding to *GS* and *PEPC* homologs were extracted from the Supplementary Materials, Table 6, of Plewiński et al. study [85] and are presented here in Table 4 for direct reference. Indeed, that survey highlighted significant differences in leaf expression levels between particular gene duplicates, namely between *GS1a1* and *GS1a2* or *GS1a3* (43.1 ± 16.4 vs. 13.8 ± 5.2 and 11.5 ± 5.0, respectively); *GS1b1* and *GS1b2* (0.3 ± 0.3 vs. 2.6 ± 1.2, respectively); *GS1c1* and *GS1c2* (0.1 ± 0.2 vs. 187.5 ± 52.4, respectively); *PEPC1a*, *PEPC1b*, and *PEPC1c* (17.0 ± 4.0 vs. 0.5 ± 0.5 vs. 65.0 ± 8.7, respectively); and *PEPC3a* and *PEPC3b* (10.5 ± 2.5 vs. 51.0 ± 11.2, respectively) [85]. The observed differences in the gene expression of *L. angustifolius GS* and *PEPC* paralogs support the previously mentioned hypothesis on the expected sub-functionalization of WGD-derived duplicates.

Table 4. Normalized leaf expression level of *GS* and *PEPC* genes in a *L. angustifolius* recombinant inbred line (RIL) mapping population (83A:476 x P27255) [85].

Gene	Accession	Mean Expression in RIL Population	Min Expression Value in RIL Population	Max Expression Value in RIL Population	Expression SD
GS1a1	Lup021297	43.1	20.4	74.1	16.4
GS1a2	Lup001512	13.8	4.9	32.2	5.2
GS1a3	Lup009916	11.5	3.6	43.7	5.0
GS1b1	Lup029429	0.3	0.0	1.4	0.3
GS1b2	Lup032636	2.6	0.4	6.0	1.2
GS1c1	Lup002132	0.1	0.0	0.9	0.2
GS1c2	Lup004581	187.5	117.6	426.6	52.4
GS2a1	Lup023221	516.2	365.3	739.7	80.3
GS2a2	-	-	-	-	-
PEPC1a	Lup022696	17.0	8.1	31.1	4.0
PEPC1b	Lup029825	0.5	0.0	2.4	0.5
PEPC1c	Lup015178	65.0	44.5	87.4	8.7
PEPC2a	Lup002214	0.0	0.0	0.5	0.1
PEPC2b	Lup026946	0.1	0.0	0.9	0.2
PEPC3a	Lup031846	10.5	4.7	16.1	2.5
PEPC3b	Lup016482	51.0	33.0	94.2	11.2
PEPC4	Lup002996	1.7	0.0	4.1	0.9
PEPC5	Lup031638	11.9	1.7	28.7	4.8
HEL	Lup023733	3.0	0.4	7.4	1.2
TUB	Lup021845	78.4	35.3	113.1	15.2

SD—standard deviation; HEL and TUB—reference genes.

2.8. The Majority of Positively Selected GS and PEPC Genes are Duplicates

According to the topology of the majority of consensus trees, 85 pairs of duplicated legume *GS* and *PEPC* sequences were selected, including those located in sister branches and those originating from different subclades (if applicable). The analysis of the nonsynonymous to synonymous substitution rate (Ka/Ks) ratio revealed that all pairs except for Lj6g3v1887800/Lj6g3v1953860 and Lj6g3v1887790/Lj6g3v1953860 were under strong purifying selection, with Ka/Ks values ranging from 0.00 to 0.32 (Supplementary file 8). The two gene pairs mentioned above had a neutral (Ka/Ks) ratio (0.87). The average Ka/Ks ratio was similar in all species except *L. japonicus*: namely 0.09 in *A. ipaensis* and *V. radiata*; 0.10 in *P. vulgaris*; 0.11 in *C. arietinum* and *G. max*; 0.12 in *T. pratense*; and 0.13 in *C. cajan*, *M. truncatula*, and *L. angustifolius*. The outlier value calculated for *L. japonicus* (0.29) resulted from the two sequence pairs with neutral ratios mentioned above. The average Ka/Ks ratio differed between gene clades, from 0.07 to 0.08 in *PEPC1*a and *PEPC1b*, through 0.12 to 0.15 in *GS1_cs2*, *PEPC2*, and *GS1_cs1*, to 0.32 in *GS2* (0.10 in *GS2* without two *L. japonicus* sequence pairs under neutral selection). To address the selection pressure in a wider phylogenetic context, a branch-site test of episodic positive selection was performed for monophyletic clades, as well as all branches, for particular legume species (Supplementary file 9). Of the 163 combinations studied, statistically significant signals of positive selection were revealed for 16 foreground branches; namely, five for *GS1_cs1*, four for *GS2*, three for *PEPC2*, two for *PEPC1a*, and single branches for *GS1_cs2* and *PEPC1b*. *L. japonicus* and *A. ipaensis* revealed the highest number of branches putatively affected by positive selection: four and three, respectively. *C. arietinum* and *T. pratense* revealed two branches with positive selection markers, whereas *C. cajan*, *G. max*, *L. angustifolius*, *M. truncatula*, and *V. radiata* showed only single branches with such residues. Different amino acid positions were altered and no common pattern for any gene clade was observed.

The majority of positively selected genes were duplicates (13 vs. 3). Duplicates revealed common selection patterns for *A. ipaensis* (*GS2* and *PEPC2*) and partially similar patterns for *L. japonicus GS2*. This may indicate that episodic positive selection occurred in these lineages before duplication events. No correlation between the inferred type of duplication (local vs. WGD) and selection pressure parameters was found; remnants of positive selection were found in both types of duplicates.

Amino acid positions altered by relaxed selection constraints did not include known ligand interacting sites (ATP, glutamate, ammonia, and metal coordination sites were evaluated according to [86]). However, few sequences were considerably truncated and lacked several ligand binding sites, namely: *GS1_cs1*, Lj0g3v0335159 and GLYMA02G41106; *GS1cs2*, Lj2g3v0658180; and *GS2*, Lj6g3v1953860.

Calculated Ka/Ks values highlighted the high selection pressure acting on GS and PEPC paralogs. In general, selection constraints are related to the position of the enzyme in metabolic pathways, as well as the contribution of performed enzymatic activity for basic cell metabolic networks. Usually, genes encoding enzymes located at the top of the metabolic pathway are under stronger purifying selection than downstream ones [87]. An association between the selective pressure acting on a gene and the position of an encoded enzyme in the pathway was revealed in a wide metabolic context [88,89], including *L. angustifolius* genes encoding isoflavone synthase and acetyl-coenzyme A carboxylase [63,64]. A higher selection pressure acts on central and highly connected enzymes, enzymes with high metabolic flux, and enzymes catalyzing reactions that are difficult to bypass through alternative pathways [88]. Moreover, enzymes participating in primary metabolism are usually under a constant strong selective pressure, whereas enzymes performing specified metabolism are under weaker negative selection [89]. One of the postulated explanations for the above pattern is that these specified metabolism genes initially experienced positive selection (higher rate than primary metabolism genes) [90].

3. Material and Methods

3.1. Research Material

This study was carried out with the use of *L. angustifolius* cv. Sonet germplasm obtained from the Polish Lupin GenBank in the Breeding Station Wiatrowo (Poznań Plant Breeders Ltd., Wiatrowo, Poland) and the narrow-leafed lupin genome BAC library [28].

3.2. Identifying GS and PEPC in the L. angustifolius Genome

GS and *PEPC* gene models were prepared on the basis of available data on legumes and used as anchors of gene-specific probes. Exon/intron numbers and lengths and elements conserved among several legumes were determined. Accessions AC174349.23 (*M. truncatula*) and L39371.2 (*M. sativa*) served as templates for *GS1* and *PEPC* gene-specific primer design, respectively. The PCR amplification was performed with the use of *L. angustifolius* genomic DNA as a template (25 ng DNA), Taq polymerase (Novazym, Poznan, Poland) supplied with 1× PCR buffer and 2.5 mM Mg^{2+}, 0.16 mM dNTP, 0.25 µM of each primer, and deionized water up to 20 µL. The PCR protocol involved initial denaturation (94 °C, 5 min) and then 40 cycles consisting of steps: denaturation (94 °C, 30 s), annealing (56 and 58 °C, 40 s), elongation (72 °C, 55 s), and final elongation (72 °C, 5 min). The obtained DNA probes were purified with the QIAquick PCR Purification Kit (Qiagen, Hilden, Germany), sequenced, and labeled by random priming with the HexaLabel DNA Labeling Kit (Fermentas, Waltham, MA, USA) and radioisotope 50 µCi [α-32P]-dCTP. Finally, probes were hybridized with the narrow-leafed lupin nuclear genome BAC library, as previously described by Książkiewicz et al. (2013). Verification of positive hybridization signals was performed by PCR and Sanger sequencing with gene-specific primers (Table 5).

Table 5. Gene-specific primers used for the probe amplification and verification of positive hybridization signals.

Probe Name	PCR Primer Sequence	Length (bp)	T*
GS	GS_F: GTTGGTCCCTCTGTTGGAATCTCTG GS_R: ATAAGCAGCAATGTGCTCATTGTGTCTC	571	56
PEPC	PEPC_F: AAAGATGTTAGGAATCTTCACATGCTGCAAGA PEPC_R: GGGGCATATTCACTTGTTGGGTTCAGT	643	58

T*—melting temperature.

3.3. Estimating GS and PEPC Sequence Variant Numbers

To estimate the number of *GS* and *PEPC* sequence variants in the *L. angustifolius* genome, droplet digital PCR (ddPCR) was performed with the use of the Bio-Rad QX200 Droplet Digital PCR System (Bio-Rad, Hercules, CA, USA). The set of *GS* and *PEPC* specific primers was anchored in the most conserved gene regions among legume plants with well-established sequence data. A gene described as a single copy in the narrow-leafed lupin genome, namely aspartate aminotransferase *(AAT*) [31,91], was used as the reference in the ddPCR experiment. A series of *L. angustifolius* genomic DNA dilutions, ranging from 0.125 to 2.0 ng/µL, were used as templates in ddPCR reactions containing 2× QX200 ddPCR EvaGreen Supermix (Bio-Rad, Hercules, CA, USA), 200 nM gene-specific primers, and 50–80 nM AAT-specific primers. The final volumes of ddPCR reactions (20 µL), together with 70 µL of droplet generation oil, were placed in DG8 Cartridges, partitioned into droplets by the QX200 Droplet Generator (Bio-Rad, Hercules, CA, USA) and transferred into 96-well plates. The ddPCR protocol involved initial denaturation (95 °C for 5 min), followed by 40 cycles consisting of steps: denaturation (95 °C, 30 s), annealing (60 and 61 °C, 30s), elongation (72 °C, 45 s), and final elongation (72 °C, 45 s). The fluorescence was read on the QX200 Droplet Reader (Bio-Rad, Hercules, CA, USA). On average, 17,000 droplets were analyzed per 20 µL PCR. The data analysis was performed with QuantaSoft droplet reader software (Bio-Rad, Hercules, CA, USA) that incorporates the Poisson distribution algorithm. Supplementary to this analysis, recently released *L. angustifolius* sequencing data (Lupin Express: annotated gene set cds v1.0 and genome sequence GCA_001865875.1) were screened in order to identify all variants of analyzed genes.

3.4. Characterizing GS1, GS2, and PEPC Gene Variants, as well as Their Corresponding L. angustifolius Genome Regions

Whole BAC insert sequencing was performed by the Miseq platform (Illumina, San Diego, CA, USA) in a paired-end 2 × 250 bp approach (Genomed, Warsaw, Poland).

The narrow-leafed lupin genome scaffold assembly v1.0 (GCA_000338175.1) and genome pseudochromosome assembly v1.0 (GCA_001865875.1) were used to obtain *GS* and *PEPC* gene variant sequences, not represented in BAC clones, and to establish their positions in the genome. The BLAST algorithm was optimized for highly similar sequences: e-value cut-off, 1×10^{-20}; word size, 28; match/mismatch scores, 1/-2; and gap costs, linear.

The obtained BAC clone insert sequences and narrow-leafed lupin scaffold fragments corresponding to the narrow-leafed lupin genome regions carrying *GS1, GS2*, and *PEPC* genes (average length of 100 kb) were subjected to computational characterization of repetitive content and gene coding sequences. Repetitive elements were annotated and masked using RepeatMasker Web Server version 4.0.3 (search engine, cross_match; speed/sensitivity, slow; DNA source, *Arabidopsis thaliana*) and supplemented with the CENSOR tool accessed via the Genetic Information Research Institute (sequence source, Viridiplantae; force translated search; mask pseudogenes).

Gene prediction was performed using FGENESH [92] with *G. max* as a reference species. Functional annotation of predicted coding sequences was performed with the use of the BLAST algorithm (e-value cut-off, 1×10^{-10} word size, 28; match/mismatch scores, 1/-2; and gap costs, linear). The obtained *GS1, GS2*, and *PEPC* gene structures were visualized and compared in Geneious software v 10.1 (http://www.geneious.com). The results of functional annotation were subsequently used for gene density (genes/kbp) calculation.

3.5. Positioning GS1, GS2, and PEPC in NLL Pseudochromosomes

To assign particular *GS* and *PEPC* gene variants to narrow-leafed lupin pseudochromosomes, in silico mapping was performed. *L. angustifolius* genome sequence data (GCA_001865875.1) and the latest version of the species genetic map were used [11,21]. The BLAST algorithm was optimized as follows: e-value cut-off, 1×10^{-20}; word size, 28; match/mismatch scores, 1/-2; and gap costs, linear. Moreover, previously developed molecular markers anchored within *GS1* (036L23_3, 047P22_3, 087N22_2, and 059J08_3) and *PEPC* (064J15_5, 067C07_2, and 131K15_5_3) gene sequences were incorporated into this study [31].

3.6. Describing Local Genome Rearrangements Harboring GS and PEPC Loci

To identify and describe local genome rearrangements and microsynteny patterns in regions carrying *GS* and *PEPC* genes in narrow-leafed lupin and nine Fabaceae species, *L. angustifolius* BAC sequences with a repetitive content were masked by RepeatMasker and Censor [93] and subjected to comparative mapping. The following genome sequences were used: *A. duranensis* (Peanut Genome Project accession V14167, http://www.peanutbase.org*), A. ipaensis* (Peanut Genome Project accession K30076, http://www.peanutbase.org) [6], *C. cajan* [7] (project PRJNA72815, v1.0), *C. arietinum* [8] (v1.0 unmasked, http://comparative-legumes.org), *G. max* [9] (JGI v1.1 unmasked, http://www.phytozome.net), *L. japonicus* [10] (v2.5 unmasked, http://www.kazusa.or.jp), *M. truncatula* [12] (strain A17, JCVI v4.0 unmasked, http://www.jcvi.org/medicago), *P. vulgaris* (v0.9, DOE-JGI, and USDA-NIFA; http://www.phytozome.net) [13], and *V. radiata* [14] (GenBank/EMBL/DDBJ accession JJMO00000000). The CoGe BLAST algorithm [94] was used to perform sequence similarity analyses with the following parameters: e-value cutoff, 1×10^{-20}; word size, 8; gap existence cost, 5; gap elongation cost, 2; and nucleotide match/mismatch scores, 1/−2. Microsyntenic blocks were visualized using the Web-Based Genome Synteny Viewer [95] and Circos [96].

3.7. Phylogenetic Reconstruction of the Plant Species Tree

The reference genome sequences were gathered from Phytozome [97], NCBI/RefSeq [98], and Ensembl/Plants [99] databases. A full list of genomes and respective sources is available in Supplementary file 2.

For species tree reconstruction, a set of conserved homologs were selected with conditional reciprocal BLAST (CRB-BLAST) [100] against the Ensembl/Plants version of the *A. thaliana* representative proteome (longest encoded protein at each coding locus) with default settings. Singular loci with over 95% representation as single-copy orthologs over all the analyzed species were selected for species tree reconstruction, yielding a total of 29 loci. The alignment of representative protein sequences for each orthologous locus was obtained with MAFFT-LINSi v 7.310 [101], and a 70% occupancy threshold was used to filter the alignments with trimal, while simultaneously back translating to underlying codons with the *-backtrans* option provided in trimal [102]. All alignments were concatenated and partitioned analysis was conducted on the basis of this joint supermatrix. The list of all loci (by *A. thaliana* reference locus) and the respective evolutionary models used can be found in Supplementary file 10.

An approximate species tree was reconstructed with IQTREE v 1.5.5 [103]. Optimal model selections [104] were carried out using IQTREE's built-in capabilities (MFP option). Ultrafast bootstrap approximation [105] was used to assess the topology based on a 3000 iteration threshold (convergence was reached in 104 iterations).

3.8. Determining GS1, GS2, and PEPC Gene Families Evolutionary Patterns

Sequences were gathered with independent BLASTP (2.6.0) searches of each included plant genome (including non-legume reference genomes; full list included as Supplementary file 2) and the July 2017 version of the UniProt/SwissProt (The UniProt Consortium 2017) golden standard database. The resulting hits were filtered based on the maximum 1×10^{-20} expectation value threshold and the minimum 40% coverage of at least one of the lupin homologs sequenced during the experimental phase of the project (sequences obtained from sequenced BAC clones: 047P22, 087N22, 036L23, 059J08, 064J15, 067C07, and 131K15 used as queries). Supervised clustering was then conducted in a procedure analogous to that described in our earlier work [46] and the sequences were compared against each other with USEARCH (UBLAST v8.1.1831 search with e-value threshold 1×10^{-10}) [106]. Finally, the pairwise relationships (e-values post log-transformation) were used to cluster the sequences with MCL [107] at multiple inflation threshold values. The optimal value of the inflation threshold was selected as

1.4, based on the averaged values of the silhouette width [108], which is a cluster quality measure independent of predefined class labels. The largest clusters, which contained all of the GS/PEPC hits found in SwissProt, were processed further. SwissProt sequences were initially kept for purposes of alignment/filtering, but were discarded for final phylogenetic tree reconstruction/reconciliation.

In order to filter out assembly errors, heavily truncated partial genes, and/or pseudogenes, additional criteria were used. All accepted sequences were aligned with MAFFT v7.310 and preprocessed with OD-seq [109]. OD-seq uses a gap-based distance metric to filter out outliers with significantly different gap patterns compared to the rest of alignment. Prior to assessment, a round of trimming was carried out with trimal, based on a very permissive 1% gap threshold (parameter choice resulting in retaining sequences longer than average). All discarded sequences can be found in Supplementary file 11. The *PEPC* sequence from *Archaeoglobus fulgidus* and *GS* sequence from *Rhizobium meliloti* were initially used to guide rooting (pruned prior to reconciliation), and both coding sequences were selected on the basis of respective SwissProt records.

During GS analysis, a singular, a previously established [110] sequence for *L. japonicus* was introduced in lieu of seemingly duplicated loci on the sixth pseudochromosome of the draft genome (Lj6g3v0410480/Lj6g3v0410490; both corresponding to C-terminal part of the full coding sequence). A comparison of the *L. japonicus* pseudochromosome and reference sequence of the previously cloned region, has shown that likely misassembly or recombination has affected the region, so the reference UniProt sequence was used in downstream analyses.

During PEPC analyses, sequences from the *Volvox carteri* NCBI/RefSeq genome were used in lieu of Phytozome version due to the higher gene model quality. Additionally, available sequences from *Chlamydomonas reinhardtii* were obtained through UniProt/SwissProt records (and corresponding GenBank entries), as the current reference genome does not contain full-length gene models corresponding to either PEPC1 or PEPC2.

Phylogenetic inference was conducted analogous to the species tree reconstruction described above (IQ-TREE, optimal model selection, ultrafast bootstrap approximation). Codon-based models and coding sequences were used in order to obtain a better resolution of recent bipartitions. The SCHN05 model [111] with a free-rate model of site heterogeneity [112] was selected in both cases (GS:SCHN05+R6, PEPC:SCHN05+R8). Based on the rule of parsimony, reconstructions with the least amount of inferred duplications/losses (minimum cost of optimal reconciliation based on DTL-RANGER [113] reconciliations of species/gene trees, with disabled horizontal transfer events) were chosen. Notably, this resulted in the selection of codon-based nucleotide alignments over protein sequences and the abandonment of alignment trimming for gene tree reconstruction. The visualization of optimal reconciliation was carried out with custom scripts in the Python/ETE2 environment based on the built-in ETE2 reconciliation procedure and DTL-RANGER results [114].

3.9. Selection Pressure Analysis

Pairwise selection pressure parameters, including Ka (the number of nonsynonymous substitutions per nonsynonymous site), Ks (the number of synonymous substitutions per synonymous site), and Ka/Ks ratios, were calculated in DnaSP 5 [115]. To follow the topologies of the trees, the branch-site test of positive selection was performed in PAML4 [116]. Two models were considered: a null model, in which the foreground branch might have different proportions of sites under neutral selection to the background (i.e., relaxed purifying selection), and an alternative model, in which the foreground branch might have a proportion of sites under positive selection. The hypothesis of positive selection was verified by the likelihood ratio test (alternative vs. null model) and *p*-value under a Chi-square distribution and one degree of freedom (maximum *p*-value threshold of 0.05 was used). Sites under positive selection for foreground lineages were predicted by naive empirical Bayes and Bayes empirical Bayes [117] (a minimum posterior probability threshold of 0.95 was used). Both analyses were based on the same alignments as those used for phylogenetic inference; however, codons present in less than 30% of sequences from a particular clade were removed (Supplementary file 12).

4. Conclusions

1. *GS* and *PEPC* genes were shown to have had a complex history, with bacterial-type PEPCs emerging as those best suited for future phylogenetic inquiries into relationships between divergent legumes.
2. Legume *GS* and *PEPC* genes evolved by both ancestral legume-wide and more recent lineage-specific WGDs. Descendants of these duplications have been retained in the majority of lineages and have sustained typical gene structures, implying differences in carbon/nitrogen metabolism due to regulatory rather than mechanistic changes.
3. Legume *PEPC* and *GS* gene sequences were highly conserved by significant purifying selection. Tentative traces of positive selection can only be inferred in several branches and point to single residues, outside of the core set involved in ligand binding.
4. Monocot family members of the *GS* gene family might be more ancient than dicot ones, stemming from the selective culling of duplicates predating the separation of both lineages.
5. The general patterns of lineage-specific duplications suggest that sub-functionalization and/or regulatory rewiring played a large role in shaping the extant carbon and nitrogen primary metabolic pathways in some lineages (*L. angustifolius*, *L. japonicus*, and *G. max*).

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/21/7/2580/s1.

Author Contributions: Conceptualization, K.B.C., M.K., and B.N.; data curation, K.B.C., M.K., and G.K.; formal analysis, K.B.C., M.K., and G.K.; funding acquisition, K.B.C. and B.N.; investigation, K.B.C., M.K., and G.K.; methodology, K.B.C., M.K., and G.K.; project administration, B.N.; resources, K.B.C., M.K., G.K., A.S., and J.P.; software, K.B.C., M.K., and G.K.; supervision, B.N.; validation, M.K. and G.K.; visualization, M.K.; writing—review and editing, K.B.C., M.K., and G.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by grants from the National Science Centre (https://www.ncn.gov.pl) N N301 391,939 and 2013/08/T/NZ2/00796. Bioinformatical comparative analyses were conducted under the project 2016/21/D/NZ8/01300. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments: The authors would like to thank Matthew Nelson (CSIRO Perth) for providing expertise in narrow-leafed lupin genetic mapping and phylogeny studies.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

ACCase	cytosolic and plastid acetyl-coenzyme A carboxylases
AFLP	amplified fragment length polymorphism
BAC	bacterial artificial chromosome
BTPC	bacterial-type PEPC
CHI	chalcone isomerase
EST	expressed sequence tag
GS	glutamine synthetase
IFSs	isoflavone synthetases
ITAP	intron targeted amplified polymorphism
LTRs	long terminal repeats
MFLP	molecular fragment length polymorphism
NLL	narrow-leafed lupin linkage group
PEBPs	phosphatidylethanolamine binding proteins
PEPC	phosphoenolpyruvate carboxylase
PTPC	plant-type PEPC
RFLP	restriction fragment length polymorphism
RADs	restriction site associated DNA markers
SSR	single sequence repeat
TEs	transposable elements
WGD	whole genome duplication

References

1. Bertioli, D.J.; Moretzsohn, M.C.; Madsen, L.H.; Sandal, N.; Leal-Bertioli, S.C.; Guimaraes, P.M.; Hougaard, B.K.; Fredslund, J.; Schauser, L.; Nielsen, A.M.; et al. An analysis of synteny of *Arachis* with *Lotus* and *Medicago* sheds new light on the structure stability and evolution of legume genomes. *BMC Genom.* **2009**, *10*, 45. [CrossRef] [PubMed]
2. Cardoso, D.; de Queiroz, L.P.; Pennington, R.T.; de Lima, H.C.; Fonty, E.; Wojciechowski, M.F.; Lavin, M. Revisiting the phylogeny of papilionoid legumes: New insights from comprehensively sampled early-branching lineages. *Am. J. Bot.* **2012**, *99*, 1991–2013. [CrossRef]
3. Cannon, S.B.; McKain, M.R.; Harkess, A.; Nelson, M.N.; Dash, S.; Deyholos, M.K.; Peng, Y.; Joyce, B.; Stewart, C.N., Jr.; Rolf, M.; et al. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Mol. Biol. Evol.* **2015**, *32*, 193–210. [CrossRef] [PubMed]
4. Doyle, J.J.; Luckow, M.A. The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. *Plant Physiol.* **2003**, *131*, 900–910. [CrossRef]
5. Lewis, G.; Schrire, B.; Mackind, B.; Lock, M. *Legumes of the World*; Royal Botanic Gardens Kew: London, UK, 2005.
6. Bertioli, D.J.; Cannon, S.B.; Froenicke, L.; Huang, G.; Farmer, A.D.; Cannon, E.K.; Liu, X.; Gao, D.; Clevenger, J.; Dash, S.; et al. The genome sequences of *Arachis duranensis* and *Arachis ipaensis* the diploid ancestors of cultivated peanut. *Nat. Genet.* **2016**, *48*, 438–446. [CrossRef]
7. Varshney, R.K.; Chen, W.; Li, Y.; Bharti, A.K.; Saxena, R.K.; Schlueter, J.A.; Donoghue, M.T.A.; Azam, S.; Fan, G.; Whaley, A.M.; et al. Draft genome sequence of pigeonpea (*Cajanus cajan*) an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **2012**, *30*, 83–89. [CrossRef]
8. Varshney, R.K.; Song, C.; Saxena, R.K.; Azam, S.; Yu, S.; Sharpe, A.G.; Cannon, S.; Baek, J.; Rosen, B.D.; Tar'an, B.; et al. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol.* **2013**, *31*, 240–246. [CrossRef] [PubMed]
9. Schmutz, J.; Cannon, S.B.; Schlueter, J.; Ma, J.; Mitros, T.; Nelson, W.; Hyten, D.L.; Song, Q.; Thelen, J.J.; Cheng, J.; et al. Genome sequence of the palaeopolyploid soybean. *Nature* **2010**, *463*, 178–183. [CrossRef]
10. Sato, S.; Nakamura, Y.; Kaneko, T.; Asamizu, E.; Kato, T.; Nakao, M.; Sasamoto, S.; Watanabe, A.; Ono, A.; Kawashima, K.; et al. Genome structure of the legume *Lotus japonicus*. *DNA Res.* **2008**, *15*, 227–239. [CrossRef]
11. Hane, J.K.; Ming, Y.; Kamphuis, L.G.; Nelson, M.N.; Garg, G.; Atkins, C.A.; Bayer, P.E.; Bravo, A.; Bringans, S.; Cannon, S.; et al. A comprehensive draft genome sequence for lupin (*Lupinus angustifolius*) an emerging health food: insights into plant-microbe interactions and legume evolution. *Plant Biotechnol. J.* **2017**, *15*, 318–330. [CrossRef]
12. Young, N.D.; Debelle, F.; Oldroyd, G.E.; Geurts, R.; Cannon, S.B.; Udvardi, M.K.; Benedito, V.A.; Mayer, K.F.; Gouzy, J.; Schoof, H.; et al. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* **2011**, *480*, 520–524. [CrossRef] [PubMed]
13. Schmutz, J.; McClean, P.E.; Mamidi, S.; Wu, G.A.; Cannon, S.B.; Grimwood, J.; Jenkins, J.; Shu, S.; Song, Q.; Chavarro, C.; et al. A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* **2014**, *46*, 707–713. [CrossRef] [PubMed]
14. Kang, Y.J.; Kim, S.K.; Kim, M.Y.; Lestari, P.; Kim, K.H.; Ha, B.K.; Jun, T.H.; Hwang, W.J.; Lee, T.; Lee, J.; et al. Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nat. Commun.* **2014**, *5*, 5443. [CrossRef]
15. Abbo, S.; Miller, T.E.; Reader, S.M.; Dunford, R.P.; King, I.P. Detection of ribosomal DNA sites in lentil and chickpea by fluorescent in situ hybridization. *Genome* **1994**, *37*, 713–716. [CrossRef] [PubMed]
16. Yang, H.; Tao, Y.; Zheng, Z.; Shao, D.; Li, Z.; Sweetingham, M.W.; Buirchell, B.J.; Li, C. Rapid development of molecular markers by next-generation sequencing linked to a gene conferring phomopsis stem blight disease resistance for marker-assisted selection in lupin (*Lupinus angustifolius* L.) breeding. *Theor. Appl. Genet.* **2013**, *126*, 511–522. [CrossRef] [PubMed]
17. Nelson, M.N.; Moolhuijzen, P.M.; Boersma, J.G.; Chudy, M.; Lesniewska, K.; Bellgard, M.; Oliver, R.P.; Swiecicki, W.; Wolko, B.; Cowling, W.A.; et al. Aligning a new reference genetic map of *Lupinus angustifolius* with the genome sequence of the model legume *Lotus japonicus*. *DNA Res.* **2010**, *17*, 73–83. [CrossRef] [PubMed]

18. Yang, H.; Boersma, J.G.; You, M.; Buirchell, B.J.; Sweetingham, M.W. Development and implementation of a sequence-specific PCR marker linked to a gene conferring resistance to anthracnose disease in narrow-leafed lupin (*Lupinus angustifolius* L.). *Mol. Breed.* **2004**, *14*, 145–151. [CrossRef]
19. Książkiewicz, M.; Wyrwa, K.; Szczepaniak, A.; Rychel, S.; Majcherkiewicz, K.; Przysiecka, L.; Karlowski, W.; Wolko, B.; Naganowska, B. Comparative genomics of *Lupinus angustifolius* gene-rich regions: BAC library exploration genetic mapping and cytogenetics. *BMC Genom.* **2013**, *14*, 79. [CrossRef]
20. Nelson, M.N.; Phan, H.T.; Ellwood, S.R.; Moolhuijzen, P.M.; Hane, J.; Williams, A.; O'Lone, C.E.; Fosu-Nyarko, J.; Scobie, M.; Cakir, M.; et al. The first gene-based map of *Lupinus angustifolius* L.—location of domestication genes and conserved synteny with *Medicago truncatula*. *Theor. Appl. Genet.* **2006**, *113*, 225–238. [CrossRef]
21. Kamphuis, L.G.; Hane, J.K.; Nelson, M.N.; Gao, L.; Atkins, C.A.; Singh, K.B. Transcriptome sequencing of different narrow-leafed lupin tissue types provides a comprehensive uni-gene assembly and extensive gene-based molecular markers. *Plant Biotechnol. J.* **2015**, *13*, 14–25. [CrossRef]
22. Zhou, G.; Jian, J.; Wang, P.; Li, C.; Tao, Y.; Li, X.; Renshaw, D.; Clements, J.; Sweetingham, M.W.; Yang, H. Construction of an ultra-high density consensus genetic map and enhancement of the physical map from genome sequencing in *Lupinus angustifolius*. *Theor. Appl. Genet.* **2018**, *131*, 209–223. [CrossRef]
23. Li, H.; Renshaw, D.; Yang, H.; Yan, G. Development of a co-dominant DNA marker tightly linked to gene tardus conferring reduced pod shattering in narrow-leafed lupin (*Lupinus angustifolius* L.). *Euphytica* **2010**, *176*, 49–58. [CrossRef]
24. Boersma, J.G.; Buirchel, B.J.; Sivasithamparam, K.; Yang, H. Development of two sequence-specific PCR markers linked to the le gene that reduces pod shattering in narrow-leafed Lupin (*Lupinus angustifolius* L.). *Genet. Mol. Biol.* **2007**, *30*, 623–629. [CrossRef]
25. Nelson, M.N.; Ksiazkiewicz, M.; Rychel, S.; Besharat, N.; Taylor, C.M.; Wyrwa, K.; Jost, R.; Erskine, W.; Cowling, W.A.; Berger, J.D.; et al. The loss of vernalization requirement in narrow-leafed lupin is associated with a deletion in the promoter and de-repressed expression of a Flowering Locus T (FT) homologue. *New Phytol.* **2017**, *213*, 220–232. [CrossRef] [PubMed]
26. You, M.; Boersma, J.G.; Buirchell, B.J.; Sweetingham, M.W.; Siddique, K.H.; Yang, H. A PCR-based molecular marker applicable for marker-assisted selection for anthracnose disease resistance in lupin breeding. *Cell. Mol. Biol. Lett.* **2005**, *10*, 123–134. [PubMed]
27. Książkiewicz, M.; Yang, H. Molecular Marker Resources Supporting the Australian Lupin Breeding Program. In *Compendium of Plant Genomes, The Lupin Genome*; Karam, S., Kamphuis, L., Nelson, M., Eds.; Springer Nature Switzerland AG: Cham, Switzerland, 2020.
28. Kasprzak, A.; Safar, J.; Janda, J.; Dolezel, J.; Wolko, B.; Naganowska, B. The bacterial artificial chromosome (BAC) library of the narrow-leafed lupin (*Lupinus angustifolius* L.). *Cell. Mol. Biol. Lett.* **2006**, *11*, 396–407. [CrossRef] [PubMed]
29. Gao, L.L.; Hane, J.K.; Kamphuis, L.G.; Foley, R.; Shi, B.J.; Atkins, C.A.; Singh, K.B. Development of genomic resources for the narrow-leafed lupin (*Lupinus angustifolius*): construction of a bacterial artificial chromosome (BAC) library and BAC-end sequencing. *BMC Genom.* **2011**, *12*, 521. [CrossRef]
30. Książkiewicz, M.; Rychel, S.; Nelson, M.N.; Wyrwa, K.; Naganowska, B.; Wolko, B. Expansion of the phosphatidylethanolamine binding protein family in legumes: a case study of *Lupinus angustifolius* L. FLOWERING LOCUS T homologs LanFTc1 and LanFTc2. *BMC Genom.* **2016**, *17*, 820. [CrossRef]
31. Wyrwa, K.; Ksiazkiewicz, M.; Szczepaniak, A.; Susek, K.; Podkowinski, J.; Naganowska, B. Integration of *Lupinus angustifolius* L. (narrow-leafed lupin) genome maps and comparative mapping within legumes. *Chromosome Res.* **2016**, *24*, 355–378. [CrossRef]
32. Susek, K.; Bielski, W.K.; Hasterok, R.; Naganowska, B.; Wolko, B. A First Glimpse of Wild Lupin Karyotype Variation As Revealed by Comparative Cytogenetic Mapping. *Front. Plant Sci.* **2016**, *7*, 1152. [CrossRef]
33. Susek, K.; Naganowska, B. Cytomolecular Insight Into *Lupinus* Genomes. In *Compendium of Plant Genomes, The Lupin Genome*; Karam, S., Kamphuis, L., Nelson, M., Eds.; Springer Nature Switzerland AG: Cham, Switzerland, 2020.
34. Książkiewicz, M.; Zielezinski, A.; Wyrwa, K.; Szczepaniak, A.; Rychel, S.; Karlowski, W.; Wolko, B.; Naganowska, B. Remnants of the Legume Ancestral Genome Preserved in Gene-Rich Regions: Insights from Physical Genetic and Comparative Mapping. *Plant Mol. Biol. Rep.* **2015**, *33*, 84–101. [CrossRef] [PubMed]

35. Cannon, S.B. Chromosomal Structure History and Genomic Synteny Relationships in Lupinus. In *Compendium of Plant Genomes, The Lupin Genome*; Karam, S., Kamphuis, L., Nelson, M., Eds.; Springer Nature Switzerland AG: Cham, Switzerland, 2020.
36. Kroc, M.; Koczyk, G.; Święcicki, W.; Kilian, A.; Nelson, M.N. New evidence of ancestral polyploidy in the Genistoid legume *Lupinus angustifolius* L. (narrow-leafed lupin). *Theor. Appl. Genet.* **2014**, *127*, 1237–1249. [CrossRef] [PubMed]
37. Wang, Z.; Zhou, Z.; Liu, Y.; Liu, T.; Li, Q.; Ji, Y.; Li, C.; Fang, C.; Wang, M.; Wu, M.; et al. Functional evolution of phosphatidylethanolamine binding proteins in soybean and *Arabidopsis*. *Plant Cell* **2015**, *27*, 323–336. [CrossRef] [PubMed]
38. De Bodt, S.; Theissen, G.; Van de Peer, Y. Promoter analysis of MADS-box genes in eudicots through phylogenetic footprinting. *Mol. Biol. Evol.* **2006**, *23*, 1293–1303. [CrossRef] [PubMed]
39. Maere, S.; De Bodt, S.; Raes, J.; Casneuf, T.; Van Montagu, M.; Kuiper, M.; Van de Peer, Y. Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 5454–5459. [CrossRef] [PubMed]
40. Hahn, M.W. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol.* **2007**, *8*, R141. [CrossRef]
41. Freeling, M. Bias in plant gene content following different sorts of duplication: tandem whole-genome segmental or by transposition. *Annu. Rev. Plant Biol.* **2009**, *60*, 433–453. [CrossRef]
42. Tautz, D.; Domazet-Loso, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **2011**, *12*, 692–702. [CrossRef]
43. De Smet, R.; Van de Peer, Y. Redundancy and rewiring of genetic networks following genome-wide duplication events. *Curr. Opin. Plant Biol.* **2012**, *15*, 168–176. [CrossRef]
44. Gladieux, P.; Ropars, J.; Badouin, H.; Branca, A.; Aguileta, G.; de Vienne, D.M.; Rodriguez de la Vega, R.C.; Branco, S.; Giraud, T. Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. *Mol. Ecol.* **2014**, *23*, 753–773. [CrossRef]
45. Page, R.D.; Charleston, M.A. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol.* **1997**, *7*, 231–240. [CrossRef]
46. Koczyk, G.; Dawidziuk, A.; Popiel, D. The Distant Siblings-A Phylogenomic Roadmap Illuminates the Origins of Extant Diversity in Fungal Aromatic Polyketide Biosynthesis. *Genome Biol. Evol.* **2015**, *7*, 3132–3154. [CrossRef] [PubMed]
47. Fedorowicz-Strońska, O.; Koczyk, G.; Kaczmarek, M.; Krajewski, P.; Sadowski, J. Genome-wide identification, characterisation and expression profiles of calcium-dependent protein kinase genes in barley (Hordeum vulgare L.). *J. Appl. Genet.* **2017**, *58*, 11–22. [CrossRef] [PubMed]
48. Kumada, Y.; Benson, D.R.; Hillemann, D.; Hosted, T.J.; Rochefort, D.A.; Thompson, C.J.; Wohlleben, W.; Tateno, Y. Evolution of the glutamine synthetase gene one of the oldest existing and functioning genes. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 3009–3013. [CrossRef] [PubMed]
49. Betti, M.; Garcia-Calderon, M.; Perez-Delgado, C.M.; Credali, A.; Estivill, G.; Galvan, F.; Vega, J.M.; Marquez, A.J. Glutamine synthetase in legumes: recent advances in enzyme structure and functional genomics. *Int. J. Mol. Sci.* **2012**, *13*, 7994–8024. [CrossRef] [PubMed]
50. O'Leary, B.; Park, J.; Plaxton, W.C. The remarkable diversity of plant PEPC (phosphoenolpyruvate carboxylase): recent insights into the physiological functions and post-translational controls of non-photosynthetic PEPCs. *Biochem. J.* **2011**, *436*, 15–34. [CrossRef]
51. Sheen, J. C4 Gene Expression. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **1999**, *50*, 187–217. [CrossRef]
52. Yang, H.; Tao, Y.; Zheng, Z.; Zhang, Q.; Zhou, G.; Sweetingham, M.W.; Howieson, J.G.; Li, C. Draft genome sequence, and a sequence-defined genetic linkage map of the legume crop species Lupinus angustifolius L. *PLoS ONE* **2013**, *8*, e64799. [CrossRef]
53. Choi, S.; Creelman, R.A.; Mullet, J.E.; Wing, R.A. Construction and characterization of a bacterial artificial chromosome library of *Arabidopsis thaliana*. *Plant Mol. Biol. Rep.* **1995**, *13*, 124–128. [CrossRef]
54. Schulte, D.; Ariyadasa, R.; Shi, B.; Fleury, D.; Saski, C.; Atkins, M.; deJong, P.; Wu, C.C.; Graner, A.; Langridge, P.; et al. BAC library resources for map-based cloning and physical map construction in barley (*Hordeum vulgare* L.). *BMC Genom.* **2011**, *12*, 247. [CrossRef]
55. Yang, X.; Makaroff, C.A.; Ma, H. The Arabidopsis MALE MEIOCYTE DEATH1 gene encodes a PHD-finger protein that is required for male meiosis. *Plant Cell.* **2003**, *15*, 1281–1295. [CrossRef] [PubMed]

56. Gebhardt, C.; Oliver, J.E.; Forde, B.G.; Saarelainen, R.; Miflin, B.J. Primary structure and differential expression of glutamine synthetase genes in nodules roots and leaves of *Phaseolus vulgaris*. *EMBO J.* **1986**, *5*, 1429–1435. [CrossRef] [PubMed]
57. Tingey, S.V.; Walker, E.L.; Coruzzi, G.M. Glutamine synthetase genes of pea encode distinct polypeptides which are differentially expressed in leaves roots and nodules. *EMBO J.* **1987**, *6*, 1–9. [CrossRef] [PubMed]
58. Stanford, A.C.; Larsen, K.; Barker, D.G.; Cullimore, J.V. Differential expression within the glutamine synthetase gene family of the model legume *Medicago truncatula*. *Plant Physiol.* **1993**, *103*, 73–81. [CrossRef]
59. Temple, S.J.; Heard, J.; Ganter, G.; Dunn, K.; Sengupta-Gopalan, C. Characterization of a nodule-enhanced glutamine synthetase from alfalfa: nucleotide sequence in situ localization and transcript analysis. *Mol. Plant Microbe Interact.* **1995**, *8*, 218–227. [CrossRef]
60. Morey, K.J.; Ortega, J.L.; Sengupta-Gopalan, C. Cytosolic glutamine synthetase in soybean is encoded by a multigene family and the members are regulated in an organ-specific and developmental manner. *Plant Physiol.* **2002**, *128*, 182–193. [CrossRef]
61. Susek, K.; Bielski, W.; Czyz, K.B.; Hasterok, R.; Jackson, S.A.; Wolko, B.; Naganowska, B. Impact of Chromosomal Rearrangements on the Interpretation of Lupin Karyotype Evolution. *Genes* **2019**, *10*. [CrossRef]
62. Przysiecka, L.; Ksiazkiewicz, M.; Wolko, B.; Naganowska, B. Structure expression profile and phylogenetic inference of chalcone isomerase-like genes from the narrow-leafed lupin (*Lupinus angustifolius* L.) genome. *Front. Plant Sci.* **2015**, *6*, 268. [CrossRef]
63. Narożna, D.; Ksiazkiewicz, M.; Przysiecka, L.; Kroliczak, J.; Wolko, B.; Naganowska, B.; Madrzak, C.J. Legume isoflavone synthase genes have evolved by whole-genome and local duplications yielding transcriptionally active paralogs. *Plant Sci.* **2017**, *264*, 149–167. [CrossRef]
64. Szczepaniak, A.; Książkiewicz, M.; Podkowiński, J.; Czyż, K.B.; Figlerowicz, M.; Naganowska, B. Legume Cytosolic and Plastid Acetyl-Coenzyme-A Carboxylase Genes Differ by Evolutionary Patterns and Selection Pressure Schemes Acting before and after Whole-Genome Duplications. *Genes* **2018**, *9*. [CrossRef]
65. Ainouche, A. The Repetitive Content in Lupin Genomes. In *Compendium of Plant Genomes, The Lupin Genome*; Karam, S., Kamphuis, L., Nelson, M., Eds.; Springer Nature Switzerland AG: Cham, Switzerland, 2020.
66. Ma, B.; Kuang, L.; Xin, Y.; He, N. New Insights into Long Terminal Repeat Retrotransposons in Mulberry Species. *Genes* **2019**, *10*. [CrossRef]
67. Lockton, S.; Gaut, B. The evolution of transposable elements in natural populations of self-fertilizing *Arabidopsis thaliana* and its outcrossing relative *Arabidopsis lyrata*. *BMC Evol. Biol.* **2010**, *10*. [CrossRef] [PubMed]
68. Nakashima, K.; Abe, J.; Kanazawa, A. Chromosomal distribution of soybean retrotransposon SORE-1 suggests its recent preferential insertion into euchromatic regions. *Chromosome Res.* **2018**, *26*, 199–210. [CrossRef] [PubMed]
69. Gonzalez, L.G.; Deyholos, M.K. Identification characterization and distribution of transposable elements in the flax (*Linum usitatissimum* L.) genome. *BMC Genom.* **2012**, *13*, 644. [CrossRef] [PubMed]
70. Seabra, A.R.; Vieira, C.P.; Cullimore, J.V.; Carvalho, H.G. Medicago truncatula contains a second gene encoding a plastid located glutamine synthetase exclusively expressed in developing seeds. *BMC Plant Biol.* **2010**, *10*, 183. [CrossRef] [PubMed]
71. Forde, B.G.; Day, H.M.; Turton, J.F.; Shen, W.J.; Cullimore, J.V.; Oliver, J.E. Two glutamine synthetase genes from *Phaseolus vulgaris* L. display contrasting developmental and spatial patterns of expression in transgenic *Lotus corniculatus* plants. *Plant Cell* **1989**, *1*, 391–401. [CrossRef]
72. Tang, H.; Wang, X.; Bowers, J.E.; Ming, R.; Alam, M.; Paterson, A.H. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **2008**, *18*, 1944–1954. [CrossRef]
73. Wang, B.; Zhang, Y.; Wei, P.; Sun, M.; Ma, X.; Zhu, X. Identification of nuclear low-copy genes and their phylogenetic utility in rosids. *Genome* **2014**, *57*, 547–554. [CrossRef]
74. Cannon, S.B.; Sterck, L.; Rombauts, S.; Sato, S.; Cheung, F.; Gouzy, J.; Wang, X.; Mudge, J.; Vasdewani, J.; Schiex, T.; et al. Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 14959–14964. [CrossRef]
75. Cannon, S.B.; Ilut, D.; Farmer, A.D.; Maki, S.L.; May, G.D.; Singer, S.R.; Doyle, J.J. Polyploidy did not predate the evolution of nodulation in all legumes. *PLoS ONE* **2010**, *5*, e11630. [CrossRef]

76. Zimmer, E.A.; Wen, J. Using nuclear gene data for plant phylogenetics: progress and prospects. *Mol. Phylogenetics Evol.* **2012**, *65*, 774–785. [CrossRef] [PubMed]
77. Doyle, J.J. Evolution of higher plant glutamine synthetase genes: tissue specifity as a criterior for predicting orhology. *Mol. Biol. Evol.* **1991**, *8*, 366–377.
78. Deng, H.; Zhang, L.S.; Zhang, G.Q.; Zheng, B.Q.; Liu, Z.J.; Wang, Y. Evolutionary history of PEPC genes in green plants: Implications for the evolution of CAM in orchids. *Mol. Phylogenetics Evol.* **2016**, *94*, 559–564. [CrossRef] [PubMed]
79. Lavin, M.; Herendeen, P.S.; Wojciechowski, M.F. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. *Syst. Biol.* **2005**, *54*, 575–594. [CrossRef] [PubMed]
80. Schlueter, J.A.; Dixon, P.; Granger, C.; Grant, D.; Clark, L.; Doyle, J.J.; Shoemaker, R.C. Mining EST databases to resolve evolutionary events in major crop species. *Genome* **2004**, *47*, 868–876. [CrossRef]
81. Pfeil, B.E.; Schlueter, J.A.; Shoemaker, R.C.; Doyle, J.J. Placing paleopolyploidy in relation to taxon divergence: A phylogenetic analysis in legumes using 39 gene families. *Syst. Biol.* **2005**, *54*, 441–454. [CrossRef]
82. Rizzon, C.; Ponger, L.; Gaut, B.S. Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. *PLoS Comput. Biol.* **2006**, *2*, e115. [CrossRef]
83. Blanc, G.; Wolfe, K.H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **2004**, *16*, 1667–1678. [CrossRef]
84. Xu, C.; Nadon, B.D.; Kim, K.D.; Jackson, S.A. Genetic and epigenetic divergence of duplicate genes in two legume species. *Plant Cell Environ.* **2018**, *41*, 2033–2044. [CrossRef]
85. Plewiński, P.; Książkiewicz, M.; Rychel-Bielska, S.; Rudy, E.; Wolko, B. Candidate domestication-related genes revealed by expression quantitative trait loci mapping of narrow-leafed lupin (*Lupinus angustifolius* L.). *Int. J. Mol. Sci.* **2019**, *20*, 5670. [CrossRef]
86. Torreira, E.; Seabra, A.R.; Marriott, H.; Zhou, M.; Llorca, O.; Robinson, C.V.; Carvalho, H.G.; Fernandez-Tornero, C.; Pereira, P.J. The structures of cytosolic and plastid-located glutamine synthetases from *Medicago truncatula* reveal a common and dynamic architecture. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2014**, *70*, 981–993. [CrossRef]
87. Cork, J.M.; Purugganan, M.D. The evolution of molecular genetic pathways and networks. *Bioessays* **2004**, *26*, 479–484. [CrossRef] [PubMed]
88. Aguilar-Rodriguez, J.; Wagner, A. Metabolic Determinants of Enzyme Evolution in a Genome-Scale Bacterial Metabolic Network. *Genome Biol. Evol.* **2018**, *10*, 3076–3088. [CrossRef] [PubMed]
89. Maeda, H.A. Evolutionary Diversification of Primary Metabolism and Its Contribution to Plant Chemical Diversity. *Front. Plant Sci.* **2019**, *10*, 881. [CrossRef] [PubMed]
90. Moore, B.M.; Wang, P.; Fan, P.; Leong, B.; Schenck, C.A.; Lloyd, J.P.; Lehti-Shiu, M.D.; Last, R.L.; Pichersky, E.; Shiu, S.H. Robust predictions of specialized metabolism genes through machine learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 2344–2353. [CrossRef] [PubMed]
91. Mett, V.; Mett, V.L.; Reynolds, P.H. The aspartate aminotransferase-P2 gene from *Lupinus angustifolius*. *Plant Physiol.* **1994**, *106*, 1683–1684. [CrossRef]
92. Salamov, A.A.; Solovyev, V.V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **2000**, *10*, 516–522. [CrossRef]
93. Kohany, O.; Gentles, A.J.; Hankus, L.; Jurka, J. Annotation submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinform.* **2006**, *7*, 474. [CrossRef]
94. Lyons, E.; Pedersen, B.; Kane, J.; Alam, M.; Ming, R.; Tang, H.; Wang, X.; Bowers, J.; Paterson, A.; Lisch, D.; et al. Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya poplar and grape: CoGe with rosids. *Plant Physiol.* **2008**, *148*, 1772–1781. [CrossRef]
95. Revanna, K.V.; Chiu, C.C.; Bierschank, E.; Dong, Q. GSV: a web-based genome synteny viewer for customized data. *BMC Bioinform.* **2011**, *12*, 316. [CrossRef]
96. Krzywinski, M.; Schein, J.; Birol, I.; Connors, J.; Gascoyne, R.; Horsman, D.; Jones, S.J.; Marra, M.A. Circos: an information aesthetic for comparative genomics. *Genome Res.* **2009**, *19*, 1639–1645. [CrossRef] [PubMed]
97. Goodstein, D.M.; Shu, S.; Howson, R.; Neupane, R.; Hayes, R.D.; Fazo, J.; Mitros, T.; Dirks, W.; Hellsten, U.; Putnam, N.; et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **2012**, *40*, D1178–D1186. [CrossRef] [PubMed]

98. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference sequence (RefSeq) database at NCBI: current status taxonomic expansion and functional annotation. *Nucleic Acids Res.* **2016**, *44*, D733–D745. [CrossRef]
99. Bolser, D.M.; Staines, D.M.; Perry, E.; Kersey, P.J. Ensembl Plants: Integrating Tools for Visualizing Mining and Analyzing Plant Genomic Data. *Methods Mol. Biol.* **2017**, *1533*, 1–31. [CrossRef] [PubMed]
100. Aubry, S.; Kelly, S.; Kumpers, B.M.; Smith-Unna, R.D.; Hibberd, J.M. Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C4 photosynthesis. *PLoS Genet.* **2014**, *10*, e1004365. [CrossRef]
101. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef]
102. Capella-Gutierrez, S.; Silla-Martinez, J.M.; Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **2009**, *25*, 1972–1973. [CrossRef]
103. Nguyen, L.T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274. [CrossRef] [PubMed]
104. Kalyaanamoorthy, S.; Minh, B.Q.; Wong, T.K.F.; von Haeseler, A.; Jermiin, L.S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **2017**, *14*, 587–589. [CrossRef] [PubMed]
105. Minh, B.Q.; Nguyen, M.A.; von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **2013**, *30*, 1188–1195. [CrossRef]
106. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **2010**, *26*, 2460–2461. [CrossRef] [PubMed]
107. Enright, A.J.; Van Dongen, S.; Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **2002**, *30*, 1575–1584. [CrossRef] [PubMed]
108. Rousseeuw, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
109. Jehl, P.; Sievers, F.; Higgins, D.G. OD-seq: outlier detection in multiple sequence alignments. *BMC Bioinform.* **2015**, *16*, 269. [CrossRef] [PubMed]
110. Thykjaer, T.; Danielsen, D.; She, Q.; Stougaard, J. Organization and expression of genes in the genomic region surrounding the glutamine synthetase gene Gln1 from Lotus japonicus. *Mol. Genet. Genom.* **1997**, *255*, 628–636. [CrossRef]
111. Schneider, A.; Cannarozzi, G.M.; Gonnet, G.H. Empirical codon substitution matrix. *BMC Bioinform.* **2005**, *6*, 134. [CrossRef] [PubMed]
112. Soubrier, J.; Steel, M.; Lee, M.S.; Der Sarkissian, C.; Guindon, S.; Ho, S.Y.; Cooper, A. The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Mol. Biol. Evol.* **2012**, *29*, 3345–3358. [CrossRef]
113. Bansal, M.S.; Kellis, M.; Kordi, M.; Kundu, S. RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication transfer and loss. *Bioinformatics* **2018**, *34*, 3214–3216. [CrossRef]
114. Huerta-Cepas, J.; Dopazo, J.; Gabaldon, T. ETE: a python Environment for Tree Exploration. *BMC Bioinformatics* **2010**, *11*, 24. [CrossRef]
115. Librado, P.; Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **2009**, *25*, 1451–1452. [CrossRef]
116. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **2007**, *24*, 1586–1591. [CrossRef] [PubMed]
117. Yang, Z.; Wong, W.S.; Nielsen, R. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **2005**, *22*, 1107–1118. [CrossRef] [PubMed]

International Journal of
Molecular Sciences

MDPI

Article

Integrated Metabolomic and Transcriptomic Analysis to Characterize Cutin Biosynthesis between Low- and High-Cutin Genotypes of *Capsicum chinense* Jacq

Purushothaman Natarajan [1,2,†], Tolulope Abodunrin Akinmoju [1,†], Padma Nimmakayala [1,*], Carlos Lopez-Ortiz [1], Marleny Garcia-Lozano [1], Benjamin J. Thompson [1], John Stommel [3] and Umesh K. Reddy [1,*]

[1] Department of Biology and Gus R. Douglass Institute, West Virginia State University, Institute, WV 25112, USA; pnatarajan@wvstateu.edu (P.N.); takinmoju@wvstateu.edu (T.A.A.); carlos.ortiz@wvstateu.edu (C.L.-O.); mgarcialozano@wvstateu.edu (M.G.-L.); thompson504@live.marshall.edu (B.J.T.)

[2] Department of Genetic Engineering, SRM Institute of Science and Technology, Chennai 603203, TN, India

[3] Genetic Improvement of Fruits and Vegetables Laboratory, U.S. Department of Agriculture, Agricultural Research Service, Beltsville Agricultural Research Center, Beltsville, MD 20705-2325, USA; john.stommel@usda.gov

* Correspondence: padma@wvstateu.edu (P.N.); ureddy@wvstateu.edu (U.K.R.)

† These authors contributed equally to this work.

Received: 25 December 2019; Accepted: 17 February 2020; Published: 19 February 2020

Abstract: Habanero peppers constantly face biotic and abiotic stresses such as pathogen/pest infections, extreme temperature, drought and UV radiation. In addition, the fruit cutin lipid composition plays an important role in post-harvest water loss rates, which in turn causes shriveling and reduced fruit quality and storage. In this study, we integrated metabolome and transcriptome profiling pertaining to cutin in two habanero genotypes: PI 224448 and PI 257145. The fruits were selected by the waxy or glossy phenotype on their surfaces. Metabolomics analysis showed a significant variation in cutin composition, with about 6-fold higher cutin in PI 257145 than PI 224448. It also revealed that 10,16-dihydroxy hexadecanoic acid is the most abundant monomer in PI 257145. Transcriptomic analysis of high-cutin PI 257145 and low-cutin PI 224448 resulted in the identification of 2703 statistically significant differentially expressed genes, including 1693 genes upregulated and 1010 downregulated in high-cutin PI 257145. Genes and transcription factors such as GDSL lipase, glycerol-3 phosphate acyltransferase 6, long-chain acyltransferase 2, cytochrome P450 86A/77A, SHN1, ANL2 and HDG1 highly contributed to the high cutin content in PI 257145. We predicted a putative cutin biosynthetic pathway for habanero peppers based on deep transcriptome analysis. This is the first study of the transcriptome and metabolome pertaining to cutin in habanero peppers. These analyses improve our knowledge of the molecular mechanisms regulating the accumulation of cutin in habanero pepper fruits. These resources can be built on for developing cultivars with high cutin content that show resistance to biotic and abiotic stresses with superior postharvest appearance.

Keywords: GDSL lipase; GPAT6; cutin; habaneros; *Capsicum chinense*; fruit; RNA-Seq

1. Introduction

Hot peppers (*Capsicum chinense* Jacq.), popularly known as habaneros, are domesticated from tropical regions of Central America and have great economic significance in terms of culinary, pharmaceutical and ornamental perspectives. Their fruits are a good source of vitamins, antioxidants and other phytonutrients, including major important alkaloids such as phenolics, carotenoids, flavonoids and capsaicinoids [1,2]. Habanero pepper fruits are subjected to desiccation and postharvest wilting because of their hollow shape and limited water-holding capacity. The abundance of cuticle in the pericarp can resist desiccation, but pepper cuticle development is not yet well understood [3,4]. Cuticle is known to play a critical role in plant survival because its primary physiological function is as a sealant around plant tissues to protect against drought conditions and prevent desiccation by reducing nonstomatal water loss [5–7]. The cuticle structure is diverse among different species but is made up of a polyester cutin that is covered with waxes (intracuticular and epicuticular). Cuticle, a hydrophobic polymer synthesized by the epidermal cells, is a major physiological trait acquired by plants during evolution for survival in dehydrated conditions. It also coordinates the interaction between a plant and its environment by limiting UV radiation and mechanical damage and is a defense against pathogen entry. In terms of chemical composition, the cuticle is a polyester matrix of cutin embedded with waxes [8].

Cutin is the major constituent of the plant cuticle and makes up about 80% of the plant cuticle. It is an insoluble, covalently cross-linked polymer that is synthesized by epidermal cells in higher plants [5,6,9]. Cutin is made up of organic chemicals that include glycerol, hydroxylated fatty acids and hydroxylated epoxy compounds with carbon atom chains of lengths 16 and 18 and phenolic compounds [5,10–13]. Cutin composition and its genetic basis have been studied in model plants such as *Arabidopsis*, tomato and rice [14–17]. In *Arabidopsis*, several genes including GPAT6, GDSL lipase, LACS, CYP86A, CYP77A, ABCG32 and ABCG11 involved in cutin initiation and development have been identified [7,18,19]. In peppers, Parsons et al. [4] reported that the cuticle of *Capsicum annum* fruit show variations in composition among species and cultivars [4,7,20]. Additionally, cuticle composition varies across pepper cultivars, which in turn affects the response to biotic and abiotic stresses [21]. However, a better understanding of the molecular basis of this monomer composition is important for using cutin for crop improvement in pepper [4].

Recent progress in "omics" approaches is being utilized for tracking the metabolites and genes involved in cutin biosynthesis, transport and regulation in plant tissues [4,19,22]. Owing to the widely proposed significance of cuticle in plant physiology and metabolism, the metabolite profile of cutin has been explored extensively in the model plant *Arabidopsis* and other crops such as barley, tomato, rice and maize [15,17,23]. Different aspects of cuticle biosynthesis have been considered in *Arabidopsis* and tomato fruits [5,12,13], however, there are no studies reported in habanero peppers in terms of whole fruit transcriptome and metabolome to understand cutin accumulation and metabolism. Hence, the current study aimed to understand cutin biosynthesis in habanero peppers by taking advantage of integrated RNA-Seq and metabolome analysis to study cutin biosynthesis in fruit tissues of diverse *C. chinense* genotypes. Here, we used gas chromatography–mass spectrometry (GC-MS) of two different habanero peppers, PI 224448 from Costa Rica and PI 257145 from Peru, for metabolome analysis to study cutin composition across genotypes. We also performed deep paired-end RNA-Seq of the two samples by using the Illumina Nextseq 500 platform to identify differentially expressed genes (DEGs) and pathways associated with cutin and other traits by comparing PI 257145 and PI 224448. This is the first study to generate transcriptome and metabolome data pertaining to cutin in habanero peppers. These results can be used by plant breeders for hot pepper fruit quality improvement via biotechnological modifications and can also serve as a model for the other Solanaceae crops.

2. Results and Discussion

2.1. Metabolomic Analysis of Cutin Monomers

Raw cutin from two habanero genotypes, PI 224448 and PI 257145, were depolymerized in 3N methanolic hydrochloride (Me-OH-HCl), and cutin composition was analyzed and quantified by using GC-MS. The compositions of cutin monomers identified from the two habanero cultivars are given in Table 1. The cutin monomers from these genotypes mostly consisted of long-chain aliphatic ω-hydroxy acids, especially dihydroxy hexadecanoic acids, considered the most important component of most plant cutin materials, especially in fruits [7,24]. Parsons et al. [4] showed 16-fold differences in cutin monomer amounts between the most extreme accessions studied. Similar to this report, the metabolic analysis of cutin composition between our selected genotypes revealed significant variations in both total cutin monomer content and relative proportion of cutin. PI 257145 had the most abundant cutin content, with about 1284 mg/g dry weight (DW), and PI 224448 had the lowest cutin content, 232.4 mg/g DW. Total cutin composition and relative proportion of individual monomers varied between the two cultivars, with about 6-fold higher cutin content in PI 257145 versus PI 224448. Reports by Kissinger et al. [21] and Parsons et al. [4], showed that 10,16-dihydroxy hexadecanoic acid was the dominant cutin monomer with portions from 50% to 82% total cutin in *Capsicum annum*. Of note, our study showed a similar pattern, with PI 224448 having the lowest amount of dihydroxy hexadecanoic acid, about 114 mg/g DW (49%), and genotype PI 257145 showing the highest amount, 1060mg/g DW (83%). Among the octadecanoic acids, 9,10,12,13,18-pentahydroxy octadecanoic acid was dominant, with PI 257145 showing the highest amount, 35.3 mg/g DW. This compound was detected only in fruits, which suggests that they might play a major role in cutin composition of plants. Levels of p-coumaric acid, a phenolic compound, also showed significant variations between the two pepper genotypes. These variations between the samples provided a good background to investigate the cutin biosynthesis mechanisms by examining variations in expression of the some of the key players in this pathway.

Table 1. Cutin monomers identified from habanero pepper fruits quantified by GC-MS.

Cutin Monomers	**PI 224448**		**PI 257145**	
	Mean ± SD	**%**	**Mean ± SD**	**%**
Hexadecanoic acid	11.2 ± 2.6	4.8	16.1 ± 6.0	1.3
10,16-Dihydroxy hexadecanoic acid	114.1 ± 19.7	49.1	1060.1 ± 495.4	82.6
16-Hydroxy hexadecanoic acid	39.7 ± 15.8	17.1	77.3 ± 11.4	6.0
Octadecanoic acid	4.7 ± 0.8	2.0	6.3 ± 2.1	0.5
9,10,12,13,18-Pentahydroxy octadecanoic acid	18.1 ± 4.4	7.8	35.3 ± 11.5	2.8
9,10,18-Trihydroxy octadecanoic acid	2.1 ± 0.8	0.9	2.8 ± 1.8	0.2
Octadecenoic acid	0.6 ± 0.4	0.3	4.1 ± 0.3	0.3
Octadec-9-enoic acid	0.6 ± 0.3	0.2	0.7 ± 0.4	0.1
18-Hydroxy octadecenoic acid	1.4 ± 0.5	0.6	5.3 ± 1.0	0.4
Octadecadienoic acid	5.0 ± 2.3	2.2	14.8 ± 1.4	1.2
18-Hydroxy octadecadienoic acid	6.0 ± 2.6	2.6	7.1 ± 2.4	0.6
p-Coumaric acid	28.7 ± 7.6	12.3	54.0 ± 37.4	4.2
Total cutin	232.4 ± 57.8	100.0	1284.0 ± 571.0	100.0

Data are mean ± standard deviation mg/g dry weight from three independent biological replications.

2.2. Fruit Transcriptome Sequencing and Analysis

Total RNA was isolated from the green fruit tissues from the two habanero pepper genotypes, PI 257145 (high cutin) and PI 224448 (low cutin). An RNA-Seq library was prepared for each genotype separately by using total RNA pooled from three biological replicates. The library was subjected to paired-end sequencing (2 × 75 bp) with Illumina NextSeq 500 platform (Illumina, California, USA). A total of 22,550,145 and 24,056,689 reads were generated for PI 257145 and PI 224448, respectively (Table 2). The raw RNA sequencing data for the two genotypes were deposited in the Short Read Archive (SRA) database of NCBI with the accession numbers SRX6761116 and SRX6761113 for PI 257145 and PI 224448, respectively, under the bioproject PRJNA562491. The raw reads were subjected to stringent quality filtering, which resulted in 21,411,561 and 19,981,360 high-quality reads for PI 257145 and PI 224448, respectively. The Q30 percentage of reads in each library was ≥95%. The reads from the two genotypes were aligned to the *C. chinense* reference genome [25] by using the STAR universal RNA-Seq alignment tool with default parameters [26]. A total of 21,021,870 (98.18%) and 19,649,669 (98.34%) quality-filtered reads were mapped to the reference genomes for PI 257145 and PI 224448, respectively; 2% of the reads remained unmapped.

Table 2. Summary of RNA-Seq and reference genome alignment in fruit tissue of *Capsicum chinense* Jacq.

Particulars	PI 257145	PI 224448
Total raw reads	22,550,145	24,056,689
Total valid paired-end reads	21,411,561	19,981,360
Read length	75	75
GC content (%)	41	43
Q30 (%)	95.2	95.7
Mapped reads	21,021,870(98.18%)	19,649,669 (98.34%)
Unmapped reads	389,691 (1.82%)	331,691 (1.66%)
Unique mapped reads	19,802,144 (92.48%)	17,391,436 (87.03%)
Multiple mapped reads	1,100,712 (5.14%)	1,569,061 (7.85%)

2.3. DEGs Between PI 257145 and PI 224448

The individual read count tables across genes for the two genotypes were created by genome alignment with the HTSeq R package and RSEM [27] with RPKM normalization. DEGs were identified by pair-wise combinations by comparing PI 257145 and PI 224448 with the use of NOISeq R/Bioc package [28] with three simulated replicates having a variability of 0.02 and CPM value 1. The DEGs were filtered based on the minimum Log2FC of 1 and *p*-value 0.9 as per the NOISeq R/Bioc package. A total of 2703 statistically significant DEGs were identified including 1693 upregulated and 1010 downregulated genes in PI 257145 versus PI 224448 (Figure 1). The top 20 upregulated genes in PI 257145 versus PI 224448 are in Table 3. The top 50 differentially expressed genes between PI 224448 and PI 257145 based on the FPKM-normalized-Log10 transformed counts are in Figure 1.

Table 3. Top 20 upregulated genes in PI 257145 versus PI 224448.

Name	Annotation	Log2FC	PI 257145 (FPKM)	PI 224448 (FPKM)
TC.CC.CCv1.2.scaffold1153.2	Glycine-rich protein-like	11.68	4066.11	1.24
TC.CC.CCv1.2.scaffold403.5	BURP domain protein USPL1-like	11.59	3119.46	1.01
TC.CC.CCv1.2.scaffold917.27	Uncharacterized mitochondrial protein AtMg00810-like	11.58	141.33	0.05
TC.CC.CCv1.2.scaffold543.19	Nonspecific lipid-transfer protein A-like	11.57	655.87	0.22
TC.CC.CCv1.2.scaffold177.64	Wound-induced protein	11.50	8286.74	2.85
TC.CC.CCv1.2.scaffold1131.20	Protein EXORDIUM-like 2	11.36	1497.42	0.57
TC.CC.CCv1.2.scaffold260.25	Probable cellulose synthase A catalytic subunit 3]	10.41	47.52	0.04
TC.CC.CCv1.2.scaffold123.73	Proteinase inhibitor PSI-1.2-like	10.34	1359.44	1.05
TC.CC.CCv1.2.scaffold552.63	Proline-rich receptor-like protein kinase PERK13	10.02	488.72	0.47
TC.CC.CCv1.2.scaffold327.13	Haloacid dehalogenase-like hydrolase domain-containing protein 3	9.99	13.39	0.01
TC.CC.CCv1.2.scaffold217.2	Patatin group D-3-like	9.87	57.32	0.06

Table 3. *Cont.*

Name	Annotation	Log2FC	PI 257145 (FPKM)	PI 224448 (FPKM)
TC.CC.CCv1.2.scaffold543.15	Nonspecific lipid-transfer protein A-like	9.79	442.52	0.00
TC.CC.CCv1.2.scaffold200.84	Proline-rich extensin-like protein EPR1	9.73	1784.23	2.10
TC.CC.CCv1.2.scaffold726.49	Em protein H5	9.69	2630.41	3.17
TC.CC.CCv1.2.scaffold1580.5	Probable polyamine oxidase 4	9.66	26.49	0.03
TC.CC.CCv1.2.scaffold260.9	Chlorophyll a-b binding protein 3C, chloroplastic	9.56	198.21	0.26
TC.CC.CCv1.2.scaffold223.6	GDSL esterase/lipase At4g01130-like	9.55	168.16	0.22
TC.CC.CCv1.2.scaffold600.23	NADPH-dependent aldehyde reductase 1, chloroplastic-like	9.41	301.84	0.44
TC.CC.CCv1.2.scaffold323.10	Neutral ceramidase-like	9.38	13.37	0.02
TC.CC.CCv1.2.scaffold161.6	Zinc finger CCCH domain-containing protein 32-like isoform X1	9.13	14.35	0.03

FC, fold change; FPKM, fragments per kilobase of transcripts per million.

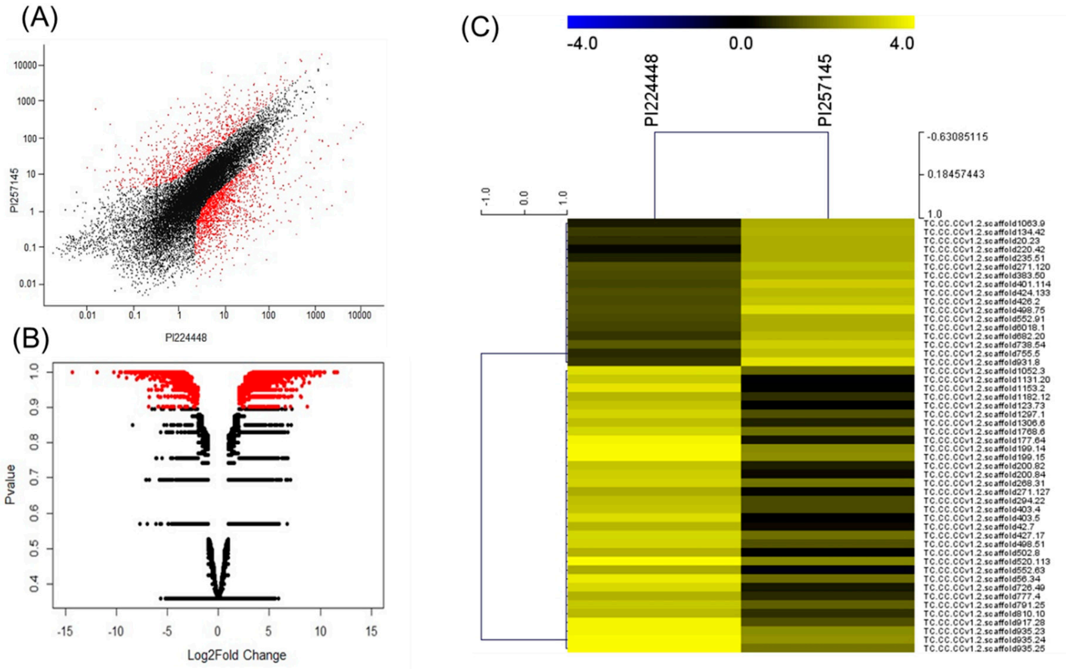


Figure 1. (**A**) Summary plot of expression values for the genotypes PI 257145 and PI 224448. The red points represent the genes with significant *p*-value of ≥ 0.9. (**B**) Volcano plot showing the Log2 fold change (FC) of differentially expressed genes (DEGs) in PI 257145 versus PI 224448. The Log2FC is plotted on the *x*-axis and the *p*-value is plotted on the y-axis. The red points in the scatter-plot show the DEGs with *p*-value ≥ 0.9 and the black points are less significant with *p*-value > 0.9. (**C**) Top 50 differentially expressed genes between the genotypes PI 224448 and PI 257145 based on the fragments per kilobase of transcripts per million (FPKM) normalized Log10-transformed counts. The color key yellow represents high expression and blue represents low expression.

2.4. Functional Annotation and Classification of DEGs

DEGs were annotated by using the BLASTx algorithm and nonredundant protein database at NCBI. Gene annotation and gene ontology (GO) enrichment analysis was performed with BLAST2GO (https://www.blast2go.com/). The DEGs were classified under the three major GO terms such as biological process, molecular function and cellular components. GO classification showed significant functions of the identified DEGs in PI 257145 versus PI 224448. A total of 1071 upregulated genes in PI 257145 were classified under top ten categories of biological process. The significant categories upregulated in PI 257145 included "cell wall biogenesis", "polysaccharide biosynthetic process", "carbohydrate metabolic process", "cell wall biogenesis", "cell wall organization", "response to hormone" and "external encapsulating structure organization". These categories are important for the structural stability of the fruits. The molecular function category included 941 DEGs with enriched terms of "oxidoreductase activity", "transferase activity" and "transmembrane receptor protein kinase activity". All these enriched molecular functions are important for fruit quality and plant defense. The major cellular components enriched in DEGs upregulated in PI 257145 included "cell periphery", "plasma membrane", "cell wall", "cell–cell junction", "plasmodesma" and "external encapsulating structure". All these cellular component terms enriched in DEGs upregulated in high-cutin PI 257145 are essentially involved in maintenance of cellular structure and fruit quality. The statistically enriched GO terms (false discovery rate (FDR) < 0.05) among the DEGs upregulated in PI 257145 versus PI 224448 are shown in Figure 2.

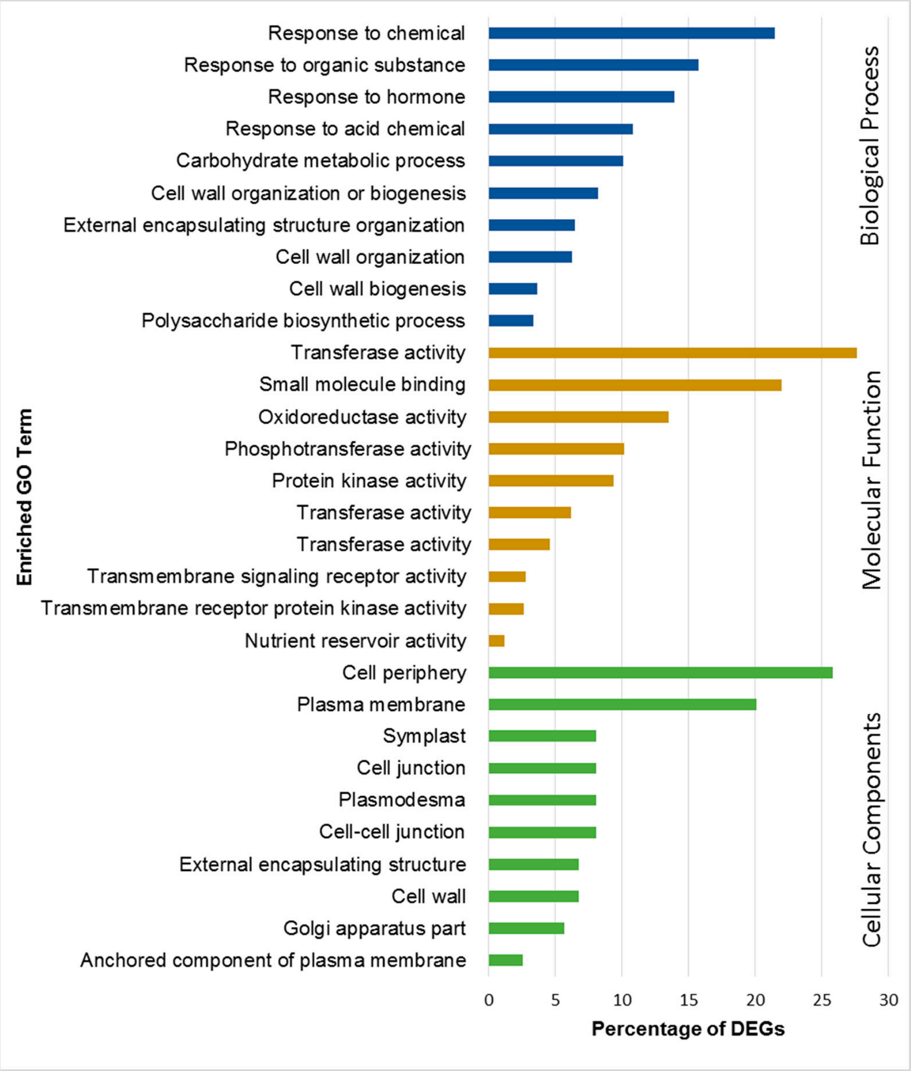


Figure 2. Top 10 gene ontology terms under biological process, molecular function and cellular components enriched among the DEGs upregulated in PI 257145 versus PI 224448.

2.5. Pathway Analysis of DEGs

Pathway analysis of DEGs involved using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database with KOBAS and MapMan. The DEGs upregulated (1693) and downregulated (1010) in PI 257145 versus PI 224448 were assigned to 100 and 97 pathways, respectively. KEGG pathway analysis shown that many of the upregulated genes are enriched in pathways relevant to cutin biosynthesis and its regulation [11]. The pathways enriched in upregulated genes of high-cutin PI 257145 were "phenylpropanoid biosynthesis", "plant hormone signal transduction", "oxidative phosphorylation", "biosynthesis of secondary metabolites", "linoleic acid metabolism", "cutin, suberine and wax biosynthesis", "fatty acid biosynthesis", "sesquiterpenoid and triterpenoid biosynthesis", "alpha-linolenic acid metabolism" and "brassinosteroid biosynthesis. The top 20 enriched KEGG pathways among upregulated and downregulated DEGs in PI 257145 compared to PI 224448 are shown in Figure 3. Pathway analysis using MapMan [29] showed differences in the activity of different cellular metabolisms between PI 257145 and PI 224448. Many of the DEGs involved in lipid metabolism and secondary metabolism were highly upregulated in PI 257145 (Figure 4). Cutin composition and its genetic basis have been studied in model plants such as *Arabidopsis*, tomato and rice [14–17,23]. Studies have shown that cuticle composition varies across pepper cultivars and this variation in turn affects their responses to biotic and abiotic stresses [21,30]. The genes involved in management of biotic and abiotic stresses are highly upregulated in the high-cutin PI 257145 versus low-cutin PI 224448. The genes regulating redox state and TFs involved in regulating defense genes were upregulated in the high-cutin PI 257145 (Figure 5).

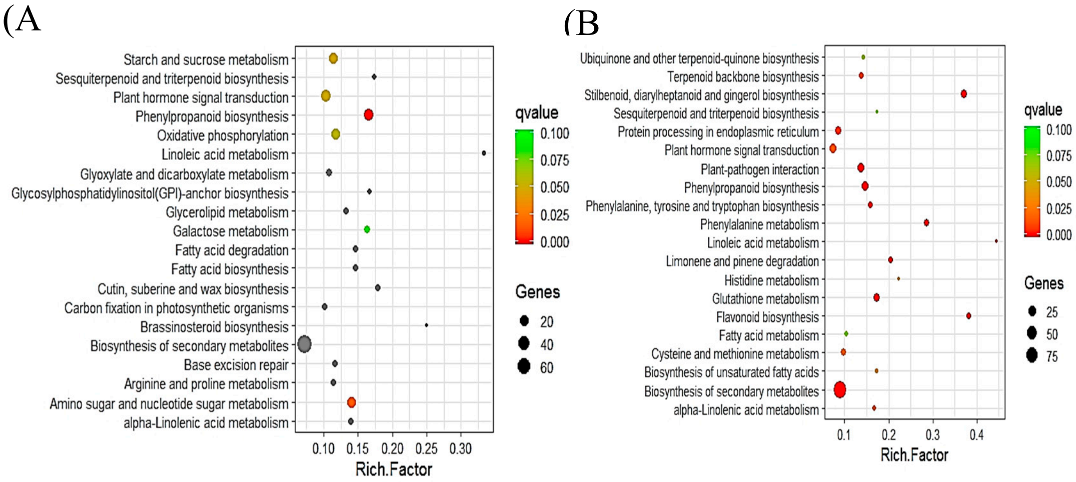


Figure 3. Top 20 enriched KEGG pathways among upregulated (**A**) and downregulated (**B**) DEGs in PI 257145 versus PI 224448. Rich factor is the ratio of the number of DEGs to the total gene number in a pathway. Here, q-value is a corrected *p*-value. The color and size of the dots represent the range of q-value and the number of DEGs mapped to the indicated pathways, respectively.

2.6. Functional Network Analysis of DEGs

Ten functional network clusters were obtained, including response to organic substance, signal transduction, multicellular organism development, cell wall biogenesis, lipid metabolic process, cell surface receptor signaling pathway, phenylpropanoid metabolic process, monocaboxylic acid metabolic process, seed oil body biosynthesis and lipid localization (Figure 6A). Most of the functional network groups were well connected to the enzymes and proteins involved in cutin biosynthesis and its regulation. Figure 6B shows the network representing an interaction between cutin genes and TFs. Bar plots are used to denote the gene expression profiles between the two genotypes and show DEGs between high- and low-cutin habanero pepper based on FPKM.

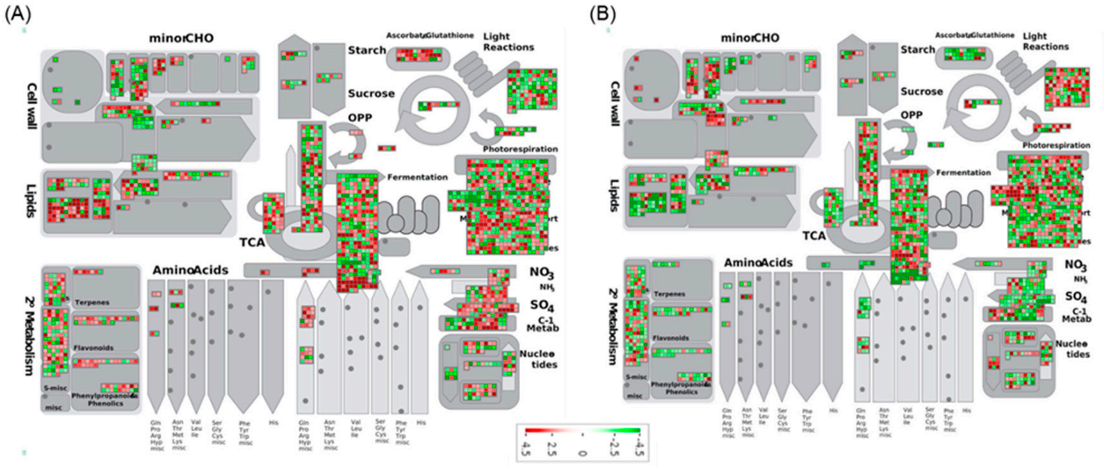


Figure 4. MapMan pathway analysis shows the differences in activity of different cellular metabolisms between the genotypes (**A**) PI 257145 and (**B**) PI 224448.

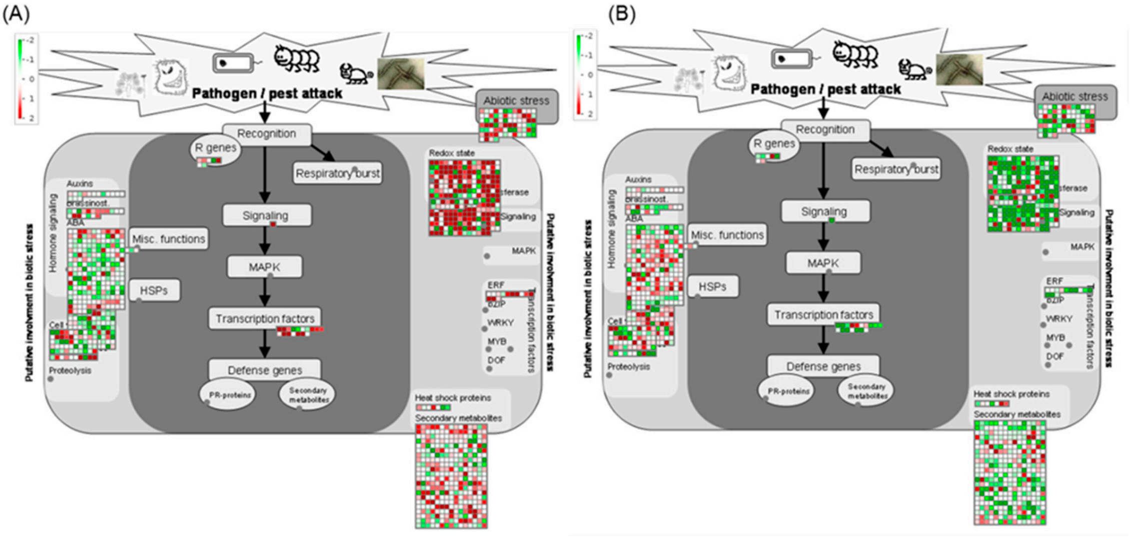


Figure 5. MapMan pathway analysis shows the expression of genes in biotic and abiotic stress-related pathways between the genotypes (**A**) PI 257145 and (**B**) PI 224448.

2.7. Analysis of DEGs for Transcription Factors (TFs)

TFs plays a major role in regulating genes for cutin biosynthesis and genes involved in biotic and abiotic stress-related pathways. TFs enriched in the DEGs were analyzed by using the Plant Transcription Factor Database (http://planttfdb.cbi.pku.edu.cn/). Among DEGs coding for TFs, 71 were upregulated and 43 were downregulated in PI 257145 versus PI 224448. The upregulated TFs represented 27 families and major TFs upregulated in PI 257145 included ERF (14), GRAS (8), MYB (4), ZF-HD (4), B3 (3), bZIP (3), C2H2 (3), C3H (3), MADS (3), NAC (3), ANL2 (3), NF-YB (2), NF-YC (2), SHN1 (2) and HDG1. Similarly, the downregulated TFs represented 16 families and major TFs downregulated in PI 257145 included ERF (13), C2H2 (7), NAC (3), WRKY (3), bHLH (2), bZIP (2), C3H (2), Dof (2), GRAS (2), HD-ZIP (2) and CFL1. Among the TF families differentially expressed, many TF families were significantly upregulated in high-cutin PI 257145. Different genes of the same TF family showed differential expression between the two genotypes. Among the TFs, 14 ERFs, 8 GRASs, 3 NACs, 3bZIPs and 3 C3Hs, were upregulated in PI 257145 and 13 ERFs, 2 GRASs, 3 NACs, 2 bZIPs and 2 C3Hs were downregulated. Furthermore, 4 MYBs, 3 B3s, 3 MADSs, 3 ANL2s, 2 SHN1s and HDG1 are uniquely upregulated in PI 257145. Among the TFs upregulated, 3 ANL2s, 2 SHN1s and HDG1 were considered important positive regulators of cutin biosynthesis [11,31,32]. These TFs playing an important role in regulating cutin biosynthesis were highly upregulated in high-cutin genotype PI 257145 versus low-cutin genotype PI 224448.

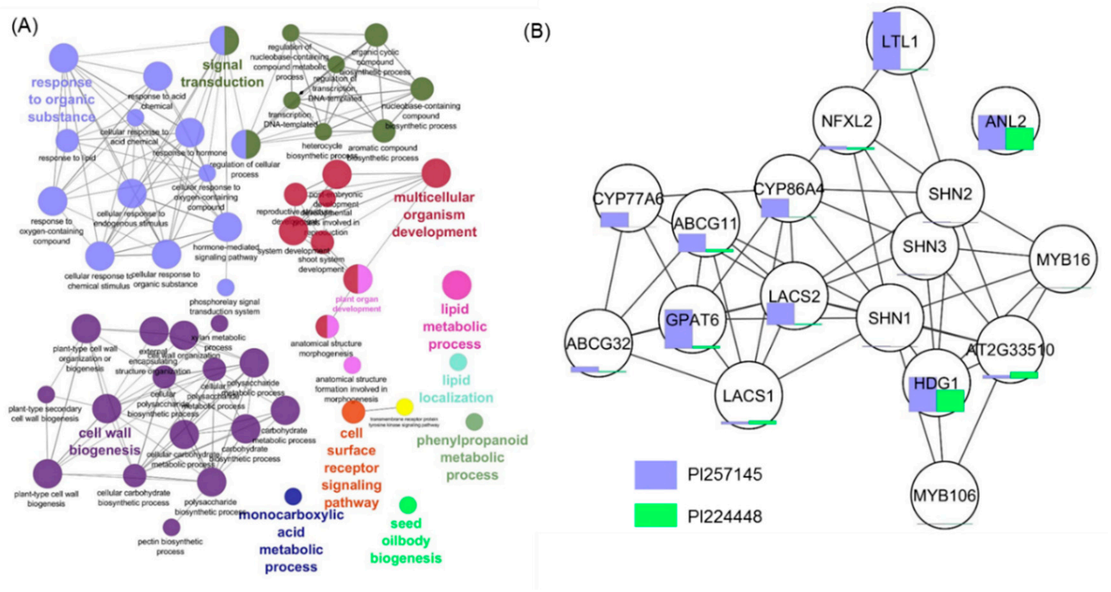

(A)
response to organic substance
signal transduction
regulation of nucleobase-containing compound metabolic process
organic cyclic compound biosynthetic process
regulation of transcription, DNA-templated
response to acid chemical
transcription, DNA-templated
nucleobase-containing compound biosynthetic process
cellular response to acid chemical
response to lipid
response to hormone
regulation of cellular process
heterocycle biosynthetic process
aromatic compound biosynthetic process
cellular response to oxygen-containing compound
post-embryonic development
multicellular organism development
reproductive developmental process
developmental process involved in reproduction
response to oxygen-containing compound
cellular response to endogenous stimulus
hormone-mediated signaling pathway
system development
shoot system development
cellular response to chemical stimulus
cellular response to organic substance
phosphorelay signal transduction system
plant organ development
lipid metabolic process
plant-type cell wall organization or biogenesis
external encapsulating structure organization
cell wall organization
xylan metabolic process
anatomical structure morphogenesis
anatomical structure formation involved in morphogenesis
lipid localization
cellular polysaccharide metabolic process
polysaccharide metabolic process
cellular polysaccharide biosynthetic process
plant-type secondary cell wall biogenesis
transmembrane receptor protein tyrosine kinase signaling pathway
carbohydrate metabolic process
cell wall biogenesis
cellular carbohydrate metabolic process
carbohydrate biosynthetic process
cell surface receptor signaling pathway
phenylpropanoid metabolic process
plant-type cell wall biogenesis
cellular carbohydrate biosynthetic process
polysaccharide biosynthetic process
monocarboxylic acid metabolic process
seed oilbody biogenesis
pectin biosynthetic process
(B)
LTL1
NFXL2
ANL2
CYP77A6
CYP86A4
SHN2
ABCG11
SHN3
MYB16
LACS2
SHN1
GPAT6
ABCG32
AT2G33510
HDG1
LACS1
MYB106
PI257145
PI224448

Figure 6. (**A**) Functional network analysis of upregulated genes in PI 257145 showing the functionally grouped terms with nodes linked based on their kappa score level (≥ 0.3), where only the label of the most significant term per group is shown. The node size represents the enrichment significance of the term. (**B**) Network analysis of genes involved in cutin biosynthesis in *Capsicum chinense*. Bar chart associated with the nodes shows the expression value fragments per kilobase of transcripts per million (FPKM) between PI 257145 and PI 224448.

2.8. Genes Involved in Cutin Biosynthesis

The cutin monomers are mostly composed of long-chain aliphatic ω-hydroxy acids, especially dihydroxy hexadecanoic acids, and have been considered the most important component of most plant cutin materials, especially in fruits [7,24]. Cutin is synthesized by epidermal cells in higher plants and is an insoluble, covalently cross-linked polymer consisting of organic chemicals including glycerol, hydroxylated fatty acids and hydroxylated epoxy compounds with carbon atom chains of lengths 16 and 18 [5,12,33]. Their monomers consist of C16 or C18 aliphatic fatty acids, their derivatives and glycerol and phenolic compounds. These monomers are generated from fatty acyl-CoA by a series of hydroxylation and epoxidation reactions that are catalyzed primarily by cytochrome-P450-dependent enzymes [5,34]. Cutin polymers are essential for plant development and are synthesized via the cutin biosynthetic pathway [11]. C16/C18 fatty acid precursors are initially catalyzed by long-chain acyl-CoA synthetase (LACS) genes, and further catalysis by downstream genes yields various monomers along the cutin pathway. Several enzymes for the biosynthesis of cutin polymer have been identified in *Arabidopsis*, involving cascade of activities from long-chain acyl-CoA synthetase (LACS1/LACS2) to cutin synthase/GDSL lipase [11,35]. *Arabidopsis* homologs for genes involved in cutin biosynthesis were used to identify the corresponding homologs from *C. chinense*, and their differential expression between the two genotypes in terms of fold change were calculated (Table 4).

Table 4. Expression of genes and transcriptions factors involved in cutin biosynthesis identified from *Capsicum chinense* genotypes.

SeqName	Gene Name	Annotation	Function	Arabidopsis Ortholog	PI 257145 (FPKM)	PI 224448 (FPKM)	Fold Change (FC)	Log2FC
Biosynthesis								
TC.CC.CCv1.2.scaffold339.9	*LACS1*	Long chain acyl-CoA synthetase 1	Attachment of CoA to free fatty acids	AT2G47240	1.21	1.532	0.78982	−0.3404
TC.CC.CCv1.2.scaffold383.59	*LACS2*	Long chain acyl-CoA synthetase 2		AT1G49430	19.408	2.595	7.479	2.90285
TC.CC.CCv1.2.scaffold383.57	*LACS2*	Long chain acyl-CoA synthetase 2		AT1G49430	14.165	0.722	19.6191	4.29419
TC.CC.CCv1.2.scaffold383.60	*LACS2*	Long chain acyl-CoA synthetase 2		AT1G49430	5.567	0.484	11.5021	3.52382
TC.CC.CCv1.2.scaffold449.40	*CYP86A8*	Cytochrome P450 86A	ω-Hydroxylase	AT2G45970	61.423	1.179	52.0975	5.70314
TC.CC.CCv1.2.scaffold419.19	*CYP86A8*	Cytochrome P450 86A		AT2G45970	12.48	0.547	22.8154	4.51193
TC.CC.CCv1.2.scaffold1130.1	*CYP77A6*	Cytochrome P450 77A	Midchain hydroxylase	AT3G10570	10.376	0.225	46.1156	5.52718
TC.CC.CCv1.2.scaffold159.143	*CYP77A4*	Cytochrome P450 77A	Epoxidase	AT5G04660	90.595	12.342	7.34038	2.87586
TC.CC.CCv1.2.scaffold419.22	*GPAT4*	Glycerol-3-phosphate 2-O-acyltransferase 4	Synthesis of 2-monoacylglycerols	AT1G01610	50.493	18.174	2.77831	1.47421
TC.CC.CCv1.2.scaffold29.10	*GPAT6*	Glycerol-3-phosphate 2-O-acyltransferase 6		AT2G38110	23.967	1.422	16.8544	4.07506
TC.CC.CCv1.2.scaffold387.10	*CUS1*	GDSL esterase/lipase	Polymerization of 2-monoacylglycerols monomers	AT3G04290	84.079	7.607	11.0528	3.46635
TC.CC.CCv1.2.scaffold120.8	*CUS1*	GDSL esterase/lipase		AT3G04290	32.9	0.203	162.069	7.34046
Transport								
TC.CC.CCv1.2.scaffold236.42	*LTPG2*	Lipid transfer protein	Transport of lipids through the cell wall	AT3G43720	5.923	2.107	2.81111	1.49114
TC.CC.CCv1.2.scaffold810.2	*ABCG11*	ABC transporter G family member 11	Export of monoacylglycerols	AT1G17840	10.092	1.494	6.75502	2.75596
TC.CC.CCv1.2.scaffold791.2	*ABCG11*	ABC transporter G family member 11		AT1G17840	36.434	6.913	5.27036	2.3979
TC.CC.CCv1.2.scaffold814.31	*ABCG11*	ABC transporter G family member 11		AT1G17840	0.457	0.134	3.41045	1.76996
TC.CC.CCv1.2.scaffold877.25	*ABCG32*	ABC transporter G family member 32		AT2G26910	2.759	0.746	3.69839	1.8869
Regulation								
TC.CC.CCv1.2.scaffold498.34	*SHN1*	AP2 transcription factor	Positive regulator	AT1G15360	0.275	0.001	275	8.10329
TC.CC.CCv1.2.scaffold772.31	*SHN1*	AP2 transcription factor		AT1G15360	1.043	0.399	2.61404	1.38628
TC.CC.CCv1.2.scaffold680.25	*MYB16*	MYB transcription factor		AT5G15310	0.191	0.108	1.76852	0.82254
TC.CC.CCv1.2.scaffold101.83	*ANL2*	HD-ZIP IV transcription factor		AT4G00730	20.477	12.96	1.58002	0.65994
TC.CC.CCv1.2.scaffold191.43	*ANL2*	HD-ZIP IV transcription factor		AT4G00730	1.863	0.731	2.54856	1.34968
TC.CC.CCv1.2.scaffold449.31	*ANL2*	HD-ZIP IV transcription factor		AT4G00730	0.087	0.001	87	6.44294
TC.CC.CCv1.2.scaffold449.30	*HDG1*	HD-ZIP IV transcription factor		AT3G61150	0.068	0.001	68	6.08746
TC.CC.CCv1.2.scaffold23.22	*NFXL2*	Zinc-finger transcription factor	Negative regulator	AT5G05660	1.998	1.094	1.82633	0.86894
TC.CC.CCv1.2.scaffold662.10	*CFL1*	WW domain protein		AT2G33510	1.859	2.924	0.63577	-0.6534
TC.CC.CCv1.2.scaffold1560.12	*BDG*	α/β-Hydrolase	Unknown	AT1G64670	4.569	0.17	26.8765	4.74827
TC.CC.CCv1.2.scaffold366.17	*BDG*	BAHD acyltransferase		AT1G64670	0.174	0.072	2.41667	1.27302

The CoAs are esterified to fatty acids by long-chain acyl-CoA synthetase (LACS1 and LACS2) to give acyl-CoA [35,36]. Mutations in LACS2 showed a reduced amount of cutin monomers and slight reduction in amount of cuticular wax in *Arabidopsis* [37]. LACS2 is important for the biosynthesis of cutin monomer, and in our study the expression of LACS2 was highly correlated with cutin content for the two genotypes. LACS2 was expressed three-fold higher in PI 257145 than PI 224448. We located the expression of three LACS2 genes in habanero genotypes and all three were significantly upregulated in high-cutin PI 257145 versus low-cutin PI 224448. The Log2FC values for the three LACS2 genes were 2.9, 4.2 and 3.5. Cytochrome-P450-dependent enzymes (particularly members of the CYP86A family and CYP77A) catalyze a series of hydroxylation and epoxidation reactions in epidermal cells in plants [5]. In the cutin biosynthetic pathway, CYP86A encodes a ω-hydroxylase that incorporates a hydroxyl group to give 16-hydroxy or 18-hydroxy fatty acids, whereas CYP77A carries a midchain hydroxylase or epoxidase activity for the synthesis of dihydroxy fatty acids [11,15,38]. Both enzyme families were upregulated in PI 257145 with fold-change range from 2.8 to 5.5 (Table 4).

Another enzyme in the pathway encodes the activity of an acyltransferase, glycerol-3-phospahate acyl transferase 6 (GPAT6), which adds the glycerol moieties into cutin. GPAT6 enzymes are involved in the transfer of fatty acids from acyl-CoA to glycerol-3-phosphate [11,24,39]. A gpat6-a mutant showed a striking phenotype in tomato fruit, with greatly altered cuticle thickness, composition and properties [40,41]. GPAT6 gene was expressed four-fold higher in high-cutin PI 257145 than low-cutin PI 224448. The enzyme GDSL esterase or lipase/cutin deficient 1 (CD1) encodes α-hydroxylase that is involved in the polymerization of various acyl-glycerols to give the cutin polymers. Previous reports have clearly demonstrated the role of this enzyme in cutin biosynthesis and showed a marked reduction of cutin content in GDSL lipase mutant tomato genotypes [12,42–44]. GDSL is considered one of the major rate-limiting enzymes for cutin biosynthesis. We have found two genes for GDSL esterase or lipase in *C. chinense*, and both were highly expressed in high-cutin PI 257145. The FC ranged from 3 to 7 in PI 257145 compared with low-cutin PI 224448. Certain ATP binding cassette (ABC) transporters, the ABCG subfamily (ABCG11 and ABCG32), have also been associated with cutin biosynthesis and are involved in the export of cutin precursors across the plasma membrane in plants [45–48]. These transporter genes are important for cutin biosynthesis. All are highly expressed in PI 257145 versus PI 224448, which agreed with the cutin content.

The transcriptional regulators in the cutin biosynthesis pathway play major roles in regulating biosynthetic genes. The WIN/SHN TFs were first identified in *Arabidopsis*, and there are three major SHN genes for cuticle biosynthesis (SHN1, SHN2 and SHN3) [31]. These sets of genes belong to the *Arabidopsis* APETALA 2 (AP2) family TFs and they regulate cutin and epidermal cells. WIN1/SHN1 is an activator of the promoter region of several cutin genes, and in tomato, SISHN3 has been reported to upregulate multiple genes involved in cutin metabolism, e.g., CYP86A gene of the cytochrome P450 [17,18,33,38,49]. Hence, the SHN1 TF is considered a strong positive regulator of cutin biosynthesis. Of note, SHN1 was expressed at a higher level in PI 257145 than in PI 224448, with 8-fold difference. Another set of TFs, the homeodomain leucine zipper IV (HD-Zip IV) TFs, were identified in *Arabidopsis*. They are highly expressed in epidermal cells and their functions are epidermis-related. One of these TFs, nuclear factor X-like 2 (NFXL2), has been identified as a negative repressor for all SHN genes, ultimately leading to negative alterations in cutin composition [18]. The expression of NFXL2 in PI 257145 was not significant, which agrees with high cutin content in this genotype. Another member of the class IV homeodomain–leucine-zipper proteins TFs regulating cutin biosynthesis discovered in *Arabidopsis* was anthocyaninless2 (ANL2). In [50], the leaf cutin composition in the ANL2 mutant was 40% less than in the *Arabidopsis* wild-type. Supporting this, in our study, ANL2, a positive regulator of cutin biosynthesis was expressed at higher level in PI 257145 than PI 224448, with about 6-fold difference. Overexpression of MYB30 in Arabidopsis was also reported to stimulate the synthesis of long chain fatty acids and cutin [51]. In our study, MYB protein had higher expression in high-cutin PI 257145 than low-cutin PI 224448, which further strengthens its role as a candidate regulatory factor in cutin metabolism. The putative cutin biosynthetic pathway genes predicted for habanero peppers based on RNA-Seq data are shown in Figure 7A; their expression based on FPKM-normalized Log10-transformed counts is shown in Figure 7B. Analysis of all genes for the cutin biosynthesis pathway revealed that all the genes experimentally validated to positively regulate cutin biosynthesis were significantly upregulated in high-cutin PI 257145 versus low-cutin PI 224448. The RNA-Seq based gene expression data and metabolic data showed significant correlation in cutin content and gene expression between the two habanero genotypes, which in turn identified the important genes and TFs contributing to the increased cutin content in PI 257145.

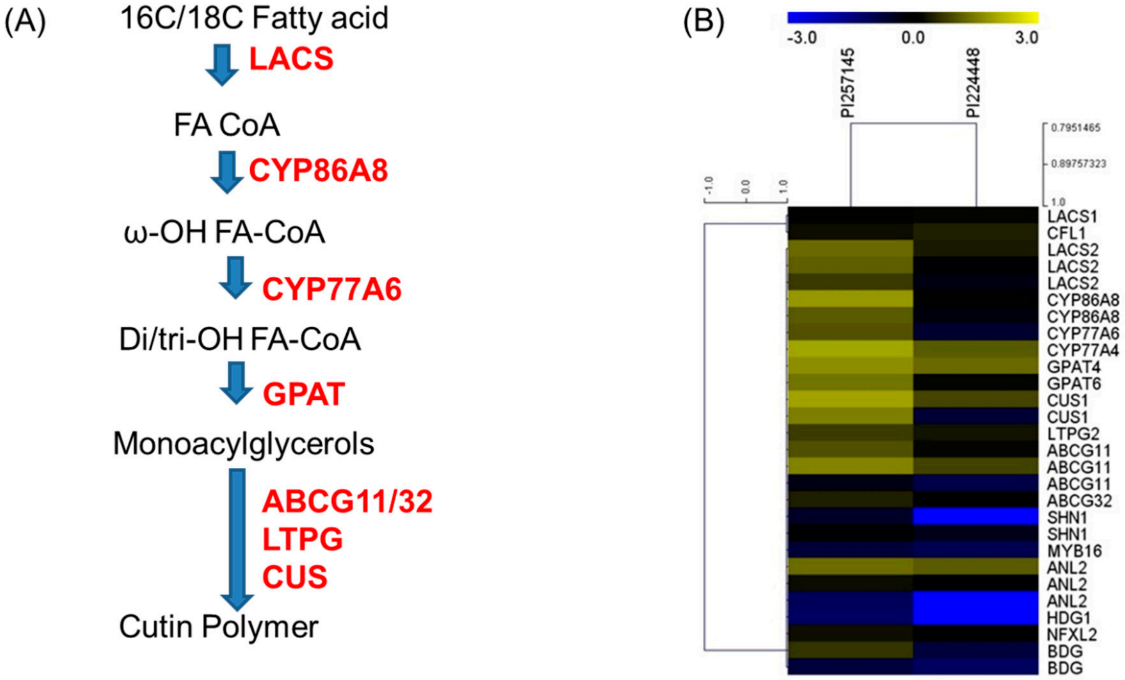


Figure 7. (**A**) Putative cutin biosynthetic pathway predicted in *Capsicum chinense* based on RNA-Seq data. (**B**) Expression of genes involved in cutin biosynthesis, transport, and regulation between the two genotypes PI 224448 and PI 257145 based on FPKM-normalized Log10-transformed counts. The color yellow represents high expression and blue represents low expression.

2.9. RNA-Seq Gene Expression Validation by RT-qPCR

To validate the RNA-Seq data, randomly selected genes involved in the cutin biosynthetic pathway with significant expression difference between PI 257145 and PI 222448 were chosen for RT-qPCR. The selected genes were GDSL esterase/lipase (CUS), glycerol-3-phosphate 2-O-acyltransferase 6 (GPAT6), long chain acyl-CoA synthetase 2 (LACS2), HD-ZIP IV transcription factor (ANL2) and cytochrome P450 86A (CYP86A4). All the five genes were significantly upregulated in high-cutin PI 257145 versus low-cutin PI 224448. The overall results from RT-qPCR were consistent with RNA-Seq data (Figure 8).

Integrating metabolomic and transcriptomic analysis revealed significant differences in cutin biosynthesis between the habanero genotypes PI 257145 and PI 224448. Metabolomics analysis revealed about 6-fold higher cutin content in PI 257145 versus PI 224448. Transcriptomic analysis revealed several significant DEGs between the high- and low-cutin genotypes. Genes such as GDSL lipase, glycerol-3 phosphate acyltransferase 6, long-chain acyltransferase 2 and cytochrome P450 86A/77A were found to be important for cutin biosynthesis. TFs such as SHN1, ANL2 and HDG1 are found to be the key regulators of the cutin biosynthetic pathway.

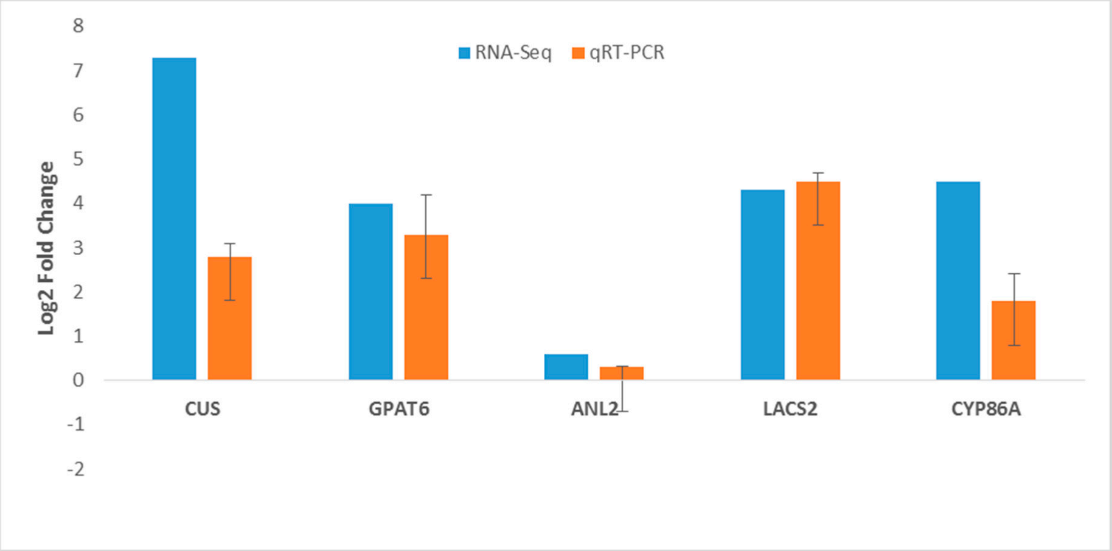


Figure 8. Relative gene expression of selected genes involved in cutin biosynthesis pathway by using RNA-Seq and RT-qPCR. Data represent Log2FC from high-cutin PI 257145 versus low-cutin PI 224448.

3. Materials and Methods

3.1. Collection of Plant Material

Seeds from two different habanero pepper genotypes (PI 224448 from Costa Rica and PI 257145 from Peru) from a worldwide collection of habanero peppers were obtained from USDA GRIN. Ten plants for each line were started in the greenhouse as surface-sterilized seeds in pots. The seeds were sown in moderately wet soil and covered with black paper bags in the dark for about 3 to 4 days to germinate in a temperature- and humidity-controlled incubator. After 4 days of germination in darkness, the pots were removed from the incubator, uncovered and left to grow under controlled conditions in the greenhouse, watered daily and finally transplanted to the Sissonville field plots. The plants were allowed to mature, and the appearance of waxes or glossiness guided our selection for the fruit sample collection. Mature green fruit tissues from each of the genotypes flowered at the same time were collected, frozen in liquid nitrogen and stored at −80 °C.

3.2. Cutin Isolation and GC-MS Analysis

Detailed metabolite profiling involved GC-MS. Cutin composition of the fruit tissues of the genotypes were examined with three replications according to the protocol reported by Parsons et al. [4] with slight modifications. Cuticle was isolated from 50 mg frozen fruit tissue powder obtained from lyophilized matured green fruits. Enzymatic digestion of the powdered samples involved using 2% pectinase and 0.1% cellulase in 0.2 mM citrate buffer, 3.7 pH (using 0.001% phenylmercuric nitrate as an antimicrobial agent). An incubator–shaker was set at 35 °C and 100 rpm for several days until the discs had little or no debris on them. Acetone with 50 mg L^{-1} butylated hydroxytoluene was used to rinse the isolated cuticles three times, followed by refluxing delipidation of the discs in chloroform:methanol (1:1, *v*/*v*). Depolymerization in 3N methanolic hydrochloride (Me-OH-HCl) was then performed by using a protocol by [52] with 6.5 mL of 3 N Me-OH-HCl for each depolymerization reaction and left for 16 h at 60 °C. The reaction vials were cooled to room temperature, and 6 mL saturated aqueous NaCl was added to stop the depolymerization reaction. The individual cutin monomers were removed as methyl esters in two different extractions by using distilled dichloromethane [53] Centrifugation at 3000 rpm for 3 min was used to separate the different phases, followed by washing the organic phase with 0.9% aqueous NaCl three times and incubation with 2,2-dimethoxy propane at 60 °C to remove dissolved water in the organic phase and then drying under nitrogen gas. BSTFA was used for derivatization followed by GC-FID analysis as previously described [54]. An Agilent 5975C GC-MS instrument with an HP-5 MS column (30 m, 0.25 mmID, 0.25 µm film) was used, and methyl heptadecanoate and methyl tricosanoate were used as internal standards. Published mass spectra of methyl ester and trimethyl silyl derivatives were used to identify the monomers ([55]; http://lipidlibrary.aocs.org/). The amount of individual cutin monomers was expressed in milligrams/gram dry weight.

3.3. RNA Isolation, Library Preparation and Transcriptome Sequencing

Total RNA was isolated from 100 mg matured green fruit tissues of the two genotypes PI 257145 and PI 224448 with three biological replicates by using the Nucleospin RNA plant kit (Macherey Nagel). Total RNA was treated with DNAseI (Qiagen) to remove co-isolated genomic DNA and purified by using the RNeasy MinElute Cleanup Kit (Qiagen). The Qubit 4 Fluorometer (Invitrogen) and Agilent 2100 Bioanalyzer were used to detect the concentration and integrity of total RNA. Total RNA from three replicates was pooled for each genotype before RNA-Seq library preparation. Libraries for the RNA-Seq of the two habanero genotypes were prepared by using the NEBNext Ultra II RNA Library Prep Kit according to the manufacturer's specification. Taking 1 µg total RNA, mRNA enrichment for poly-A involved using magnetic beads with Oligo (dT) with NEBNext Poly (A) mRNA Magnetic Isolation Module (NEB, E7490) followed by fragmentation into shorter fragments by using fragmentation buffer. Oligo dT primers were used for synthesis of first-strand cDNA. Sequencing adapters were added to the resulting cDNA followed by amplification of the library using sequencing primers. After constructing the RNA-seq library, the Agilent 2100 Bioanalyzer (Invitrogen) was used to analyze the library insert size, and the Qubit 4 Fluorometer (Invitrogen) was used to quantify the library concentration. RNA-Seq for each of the samples involved using the Illumina NextSeq 500 platform with a paired-end sequencing protocol. The resulting image files in the bcl format were converted to FASTQ with 2 × 75 bp reads with the bcl2fastq tool (Illumina).

3.4. Transcriptome Analysis

The quality of raw reads was ascertained by checking the adapter, GC distribution, average base content and quality score of the distribution by using fastqc (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The adapter sequences and low-quality reads (Phred score QV < 30) were removed and the clean reads were filtered from the raw data by using the software cutadapt (https://cutadapt.readthedocs.io/en/stable/guide.html) and sickle (https://github.com/najoshi/sickle), respectively. The quality-filtered reads were mapped to the *C. chinense* reference genome v1.2 (http://peppergenome.snu.ac.kr/) by using the STAR universal RNA-Seq alignment tool with default parameters [26] to generate BAM alignment. The read count tables for the genes across all the samples were created by using BAM alignment and the general feature format (GFF) of genome annotation with the HTSeq R package [27] and RSEM (https://deweylab.github.io/RSEM/). The counts were normalized by using reads per kilobase of transcripts per million (RPKM). The gene expression based on the read counts were studied using fragments per kilobase of transcripts per million (FPKM). The FPKM values for each of the genes were calculated based on the read count table, the total number of reads per sample and gene length in kb.

The DEGs resulting from the comparison of PI 257145 and PI 224448 were identified using the NOISeq R/Bioc package [28] with three simulated replicates having variability of 0.02 and counts per million (CPM) of 1. The DEGs were filtered based on the minimum Log2FC of 1 and *p*-value of 0.9 as per the NOISeq R/Bioc package. Gene annotation, gene ontology (GO) enrichment analysis was performed with BLAST2GO (https://www.blast2go.com/). Transcription factor (TF) prediction, and TF enrichment analysis was involved using the Plant Transcription Factor Database (http://planttfdb.cbi.pku.edu.cn/). Heatmaps were generated by using mev (http://mev.tm4.org/). Gene network analysis involved using Cytoscape (https://cytoscape.org/) and the STRING database (https://string-db.org/) with *Arabidopsis* as the reference to retrieve protein–protein interactions. Functional networks for DEGs were derived by using the ClueGO plugin [56] available in Cytoscape. Pathway mapping involved using KOBAS [57] and MapMan (https://mapman.gabipd.org/). Sequences of genes involved in cutin biosynthesis were identified by using the Arabidopsis homolog and *C. chinense* mRNA sequences [11].

3.5. RT-Quantitative PCR (RT-qPCR)

Total RNA was isolated from frozen matured green fruit tissues of habanero pepper by using the Plant RNA mini spin kit (Macherey-Nagel). The NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific, MA, USA) was used to measure RNA concentrations. The Super Script First-Strand Synthesis system (Invitrogen) was used for first-strand cDNA synthesis with 6 µg total RNA per sample. An amount of 1 µL cDNA diluted 1:6 was used for RT-qPCR analysis. In a final volume of 20 µL, diluted cDNA was mixed with 10 µL SYBR Green PCR master mix (Applied Biosystems, Foster City, CA, USA) and 10 pmol each of forward and reverse primers and completed with nuclease free water. Primer3Plus software (http://www.primer3plus.com/) was used to design gene-specific primers for the randomly selected genes involved in cutin biosynthesis. Details of the genes with primer sequences are available in Supplementary Table S1. Semiquantitative RT-PCR amplification to test primers was performed in a total reaction volume of 20 µL containing 1 µL cDNA, 10 µL colorless GoTaq and 10 pmol each of forward and reverse primers and completed with nuclease free water. Thermocycling conditions were an initial denaturing step of 95 °C for 1 min, followed by 25 cycles of 95 °C for 15 s, corresponding annealing temperature 60 °C for 70 s and 72 °C for 30 s, with a final extension step of 72 °C for 25 min. An amount of 1% agarose gel pre-stained with ethidium bromide was used to confirm the amplified fragments by visualization under UV light. Transcript-level expression was detected by RT-qPCR with SYBR Green PCR Master mix (ROX) (Roche, Shanghai) on a StepOnePlus Real-Time PCR System (Applied Biosystems, Foster City, USA). PCR involved a total reaction volume of 20 µL containing 1 µL cDNA, 1 µL of the forward and reverse primers (10 µM), 10 µL of SYBR Green PCR Master mix (ROX) (Roche, Shanghai, China) and 8 µL sterile distilled water. Amplification conditions were 95 °C for 10 min, followed by 40 cycles of 95 °C for 15 s, and 60 °C for 1 min. The reactions were performed in three technical replications and three biological replicates to compute the average Ct values. The gene expression for each gene was normalized against beta-tubulin expression and data analysis for the relative gene expression was computed with the $2^{-\Delta\Delta CT}$ method. The results are expressed as Log2foldchange (Log2FC) ± mean standard error (SEM).

4. Conclusions

Integrating metabolomic and transcriptomic analysis revealed significant differences in cutin biosynthesis between the habanero genotypes PI 257145 and PI 224448. Metabolomics analysis revealed significant variations in cutin composition between the two genotypes, with about 6-fold higher cutin content in PI 257145 versus PI 224448. Cutin monomer 10,16-dihydroxy hexadecanoic acid was present at the highest percentage (82.6%) in PI 257145. Transcriptomic analysis with RNA-Seq revealed significant gene expression differences between the high- and low-cutin genotypes. In this study, we report transcriptome and metabolome data pertaining to cutin in habanero peppers along with the predicted putative cutin biosynthetic pathway for habanero peppers. Genes such as GDSL lipase, glycerol-3 phosphate acyltransferase 6, long-chain acyltransferase 2 and cytochrome P450 86A/77A and TFs such as SHN1, ANL2 and HDG1 are found to be the key genes highly contributing to the high cutin content in PI 257145. These genes previously showed a similar pattern of regulation in tomato and *Arabidopsis*. These analyses advance our knowledge on the molecular mechanisms regulating the accumulation of cutin in habanero pepper fruits. These resources can be built on for developing habanero fruit cultivars with high cutin content that show resistance to biotic and abiotic stresses.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/21/4/1397/s1.

Author Contributions: Conceptualization, U.K.R., P.N. (Padma Nimmakayala) and J.S.; methodology, P.N. (Purushothaman Natarajan), T.A.A., C.L.-O., M.G.-L., B.J.T., U.K.R., P.N. (Padma Nimmakayala); formal analysis, P.N. (Purushothaman Natarajan), T.A.A., C.L.-O., M.G.-L., B.J.T.; data curation, P.N. (Purushothaman Natarajan),

T.A.A.; writing—original draft preparation, P.N. (Purushothaman Natarajan), T.A.A., U.K.R.; writing—review and editing, P.N. (Purushothaman Natarajan), T.A.A., U.K.R.; supervision, U.K.R., P.N. (Padma Nimmakayala); project administration, U.K.R., P.N. (Padma Nimmakayala); funding acquisition, U.K.R., P.N. (Padma Nimmakayala). All authors have read and agreed to the published version of the manuscript.

Funding: We greatly acknowledge the funding support for the current study from the USDA NIFA (grant nos. 2019-38821-29064, 2017-38821-26434 and Evans-Allen project (1008515)) and Department of Defense award (Agreement Number W911NF-16-1-0423) for next-generation sequencing instrument.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Reddy, U.K.; Almeida, A.; Abburi, V.L.; Alaparthi, S.B.; Unselt, D.; Hankins, G.; Park, M.; Choi, D.; Nimmakayala, P. Identification of gene-specific polymorphisms and association with capsaicin pathway metabolites in *Capsicum annuum* L. collections. *PLoS ONE* **2014**, *9*, e86393. [CrossRef]
2. Tripodi, P.; Cardi, T.; Bianchi, G.; Migliori, C.A.; Schiavi, M.; Rotino, G.L.; Scalzo, R.L. Genetic and environmental factors underlying variation in yield performance and bioactive compound content of hot pepper varieties (*Capsicum annuum*) cultivated in two contrasting Italian locations. *Eur. Food Res. Technol.* **2018**, *244*, 1555–1567. [CrossRef]
3. Lownds, N.; Banaras, M.; Bosland, P. Relationships between postharvest water loss and physical properties of pepper fruit (*Capsicum annuum* L.). *HortScience* **1993**, *28*, 1182–1184. [CrossRef]
4. Parsons, E.P.; Popopvsky, S.; Lohrey, G.T.; Alkalai-Tuvia, S.; Perzelan, Y.; Bosland, P.; Bebeli, P.J.; Paran, I.; Fallik, E.; Jenks, M.A. Fruit cuticle lipid composition and water loss in a diverse collection of pepper (Capsicum). *Physiol. Plan.* **2013**, *149*, 160–174. [CrossRef] [PubMed]
5. Mintz-Oron, S.; Mandel, T.; Rogachev, I.; Feldberg, L.; Lotan, O.; Yativ, M.; Wang, Z.; Jetter, R.; Venger, I.; Adato, A. Gene expression and metabolism in tomato fruit surface tissues. *Plant Physiol.* **2008**, *147*, 823–851. [CrossRef] [PubMed]
6. Leide, J.; de Souza, A.X.; Papp, I.; Riederer, M. Specific characteristics of the apple fruit cuticle: investigation of early and late season cultivars 'Prima'and 'Florina'(*Malus domestica* Borkh.). *Sci. Hortic.* **2018**, *229*, 137–147. [CrossRef]
7. Yeats, T.H.; Rose, J.K. The formation and function of plant cuticles. *Plant Physiol.* **2013**, *163*, 5–20. [CrossRef]
8. Fernández, V.; Guzmán-Delgado, P.; Graça, J.; Santos, S.; Gil, L. Cuticle structure in relation to chemical composition: re-assessing the prevailing model. *Front. Plant Sci.* **2016**, *7*, 427. [CrossRef]
9. Pollard, M.; Beisson, F.; Li, Y.; Ohlrogge, J.B. Building lipid barriers: biosynthesis of cutin and suberin. *Trends Plant Sci.* **2008**, *13*, 236–246. [CrossRef]
10. Domínguez, E.; Heredia-Guerrero, J.A.; Heredia, A. Plant cutin genesis: unanswered questions. *Trends Plant Sci.* **2015**, *20*, 551–558. [CrossRef]
11. Fich, E.A.; Segerson, N.A.; Rose, J.K. The plant polyester cutin: biosynthesis, structure, and biological roles. *Annu. Rev. Plant Biol.* **2016**, *67*, 207–233. [CrossRef] [PubMed]
12. Girard, A.-L.; Mounet, F.; Lemaire-Chamley, M.; Gaillard, C.; Elmorjani, K.; Vivancos, J.; Runavot, J.-L.; Quemener, B.; Petit, J.; Germain, V. Tomato GDSL1 is required for cutin deposition in the fruit cuticle. *Plant Cell* **2012**, *24*, 3119–3134. [CrossRef] [PubMed]
13. Yang, X.; Zhao, H.; Kosma, D.K.; Tomasi, P.; Dyer, J.M.; Li, R.; Liu, X.; Wang, Z.; Parsons, E.P.; Jenks, M.A. The acyl desaturase CER17 is involved in producing wax unsaturated primary alcohols and cutin monomers. *Plant Physiol.* **2017**, *173*, 1109–1124. [CrossRef] [PubMed]
14. Leide, J.; Hildebrandt, U.; Reussing, K.; Riederer, M.; Vogg, G. The developmental pattern of tomato fruit wax accumulation and its impact on cuticular transpiration barrier properties: effects of a deficiency in a β-ketoacyl-coenzyme A synthase (LeCER6). *Plant Physiol.* **2007**, *144*, 1667–1679. [CrossRef] [PubMed]
15. Parsons, E.P.; Popopvsky, S.; Lohrey, G.T.; Lü, S.; Alkalai-Tuvia, S.; Perzelan, Y.; Paran, I.; Fallik, E.; Jenks, M.A. Fruit cuticle lipid composition and fruit post-harvest water loss in an advanced backcross generation of pepper (Capsicum sp.). *Physiol. Plant.* **2012**, *146*, 15–25. [CrossRef]
16. Qin, B.-X.; Tang, D.; Huang, J.; Li, M.; Wu, X.-R.; Lu, L.-L.; Wang, K.-J.; Yu, H.-X.; Chen, J.-M.; Gu, M.-H. Rice OsGL1-1 is involved in leaf cuticular wax and cuticle membrane. *Mol. Plant* **2011**, *4*, 985–995. [CrossRef]

17. Shi, J.X.; Adato, A.; Alkan, N.; He, Y.; Lashbrooke, J.; Matas, A.J.; Meir, S.; Malitsky, S.; Isaacson, T.; Prusky, D. The tomato S l SHINE 3 transcription factor regulates fruit cuticle formation and epidermal patterning. *New Phytol.* **2013**, *197*, 468–480. [CrossRef]
18. Borisjuk, N.; Hrmova, M.; Lopato, S. Transcriptional regulation of cuticle biosynthesis. *Biotechnol. Adv.* **2014**, *32*, 526–540. [CrossRef]
19. Trivedi, P.; Nguyen, N.; Hykkerud, A.L.; Häggman, H.; Martinussen, I.; Jaakola, L.; Karppinen, K. Developmental and environmental regulation of cuticular wax biosynthesis in fleshy fruits. *Front. Plant Sci.* **2019**, *10*, 431. [CrossRef] [PubMed]
20. Popovsky-Sarid, S.; Borovsky, Y.; Faigenboim, A.; Parsons, E.P.; Lohrey, G.T.; Alkalai-Tuvia, S.; Fallik, E.; Jenks, M.A.; Paran, I. Genetic and biochemical analysis reveals linked QTLs determining natural variation for fruit post-harvest water loss in pepper (Capsicum). *Theor. Appl. Genet.* **2017**, *130*, 445–459. [CrossRef]
21. Kissinger, M.; Tuvia-Alkalai, S.; Shalom, Y.; Fallik, E.; Elkind, Y.; Jenks, M.A.; Goodwin, M.S. Characterization of physiological and biochemical factors associated with postharvest water loss in ripe pepper fruit during storage. *J. Am. Soc. Hortic. Sci.* **2005**, *130*, 735–741. [CrossRef]
22. Cohen, H.; Szymanski, J.; Aharoni, A. Assimilation of 'omics' strategies to study the cuticle layer and suberin lamellae in plants. *J. Exp. Bot.* **2017**, *68*, 5389–5400. [CrossRef] [PubMed]
23. Wu, R.; Li, S.; He, S.; Waßmann, F.; Yu, C.; Qin, G.; Schreiber, L.; Qu, L.-J.; Gu, H. CFL1, a WW domain protein, regulates cuticle development by modulating the function of HDG1, a class IV homeodomain transcription factor, in rice and Arabidopsis. *Plant Cell* **2011**, *23*, 3392–3411. [CrossRef]
24. Yeats, T.H.; Buda, G.J.; Wang, Z.; Chehanovsky, N.; Moyle, L.C.; Jetter, R.; Schaffer, A.A.; Rose, J.K. The fruit cuticles of wild tomato species exhibit architectural and chemical diversity, providing a new model for studying the evolution of cuticle function. *Plant J.* **2012**, *69*, 655–666. [CrossRef] [PubMed]
25. Kim, S.; Park, M.; Yeom, S.-I.; Kim, Y.-M.; Lee, J.M.; Lee, H.-A.; Seo, E.; Choi, J.; Cheong, K.; Kim, K.-T. Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nat. Genet.* **2014**, *46*, 270. [CrossRef] [PubMed]
26. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [CrossRef]
27. Anders, S.; Pyl, P.T.; Huber, W. HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics* **2015**, *31*, 166–169. [CrossRef]
28. Tarazona, S.; Furió-Tarí, P.; Turra, D.; Pietro, A.D.; Nueda, M.J.; Ferrer, A.; Conesa, A. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.* **2015**, *43*, e140. [CrossRef]
29. Thimm, O.; Bläsing, O.; Gibon, Y.; Nagel, A.; Meyer, S.; Krüger, P.; Selbig, J.; Müller, L.A.; Rhee, S.Y.; Stitt, M. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **2004**, *37*, 914–939. [CrossRef]
30. Smith, D.L.; Stommel, J.R.; Fung, R.W.; Wang, C.Y.; Whitaker, B.D. Influence of cultivar and harvest method on postharvest storage quality of pepper (*Capsicum annuum* L.) fruit. *Postharvest Biol. Technol.* **2006**, *42*, 243–247. [CrossRef]
31. Kannangara, R.; Branigan, C.; Liu, Y.; Penfield, T.; Rao, V.; Mouille, G.; Höfte, H.; Pauly, M.; Riechmann, J.L.; Broun, P. The transcription factor WIN1/SHN1 regulates cutin biosynthesis in *Arabidopsis thaliana*. *Plant Cell* **2007**, *19*, 1278–1294. [CrossRef] [PubMed]
32. Oshima, Y.; Shikata, M.; Koyama, T.; Ohtsubo, N.; Mitsuda, N.; Ohme-Takagi, M. MIXTA-like transcription factors and WAX INDUCER1/SHINE1 coordinately regulate cuticle development in Arabidopsis and Torenia fournieri. *Plant Cell* **2013**, *25*, 1609–1624. [CrossRef] [PubMed]
33. Lara, I.; Belge, B.; Goulao, L.F. A focus on the biosynthesis and composition of cuticle in fruits. *J. Agric. Food Chem.* **2015**, *63*, 4005–4019. [CrossRef] [PubMed]
34. Aharoni, A.; Dixit, S.; Jetter, R.; Thoenes, E.; van Arkel, G.; Pereira, A. The SHINE clade of AP2 domain transcription factors activates wax biosynthesis, alters cuticle properties, and confers drought tolerance when overexpressed in Arabidopsis. *Plant Cell* **2004**, *16*, 2463–2480. [CrossRef] [PubMed]
35. Li, Y.; Beisson, F.; Koo, A.J.; Molina, I.; Pollard, M.; Ohlrogge, J. Identification of acyltransferases required for cutin biosynthesis and production of cutin with suberin-like monomers. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 18339–18344. [CrossRef] [PubMed]

36. Lü, S.; Song, T.; Kosma, D.K.; Parsons, E.P.; Rowland, O.; Jenks, M.A. Arabidopsis CER8 encodes LONG-CHAIN ACYL-COA SYNTHETASE 1 (LACS1) that has overlapping functions with LACS2 in plant wax and cutin synthesis. *Plant J.* **2009**, *59*, 553–564. [CrossRef] [PubMed]
37. Schnurr, J.; Shockey, J. The acyl-CoA synthetase encoded by LACS2 is essential for normal cuticle development in Arabidopsis. *Plant Cell* **2004**, *16*, 629–642. [CrossRef]
38. Wellesen, K.; Durst, F.; Pinot, F.; Benveniste, I.; Nettesheim, K.; Wisman, E.; Steiner-Lange, S.; Saedler, H.; Yephremov, A. Functional analysis of the LACERATA gene of Arabidopsis provides evidence for different roles of fatty acid ω-hydroxylation in development. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 9694–9699. [CrossRef]
39. Yang, W.; Simpson, J.P.; Li-Beisson, Y.; Beisson, F.; Pollard, M.; Ohlrogge, J.B. A land-plant-specific glycerol-3-phosphate acyltransferase family in Arabidopsis: substrate specificity, sn-2 preference, and evolution. *Plant Physiol.* **2012**, *160*, 638–652. [CrossRef]
40. Fawke, S.; Torode, T.A.; Gogleva, A.; Fich, E.A.; Sørensen, I.; Yunusov, T.; Rose, J.K.; Schornack, S. Glycerol-3-phosphate acyltransferase 6 controls filamentous pathogen interactions and cell wall properties of the tomato and Nicotiana benthamiana leaf epidermis. *New Phytol.* **2019**, *223*, 1547–1559. [CrossRef]
41. Petit, J.; Bres, C.; Mauxion, J.-P.; Tai, F.W.J.; Martin, L.B.; Fich, E.A.; Joubès, J.; Rose, J.K.; Domergue, F.; Rothan, C. The glycerol-3-phosphate acyltransferase GPAT6 from tomato plays a central role in fruit cutin biosynthesis. *Plant Physiol.* **2016**, *171*, 894–913. [CrossRef]
42. Lara, I.; Belge, B.; Goulao, L.F. The fruit cuticle as a modulator of postharvest quality. *Postharvest Biol. Technol.* **2014**, *87*, 103–112. [CrossRef]
43. Yeats, T.H.; Howe, K.J.; Matas, A.J.; Buda, G.J.; Thannhauser, T.W.; Rose, J.K. Mining the surface proteome of tomato (*Solanum lycopersicum*) fruit for proteins associated with cuticle biogenesis. *J. Exp. Bot.* **2010**, *61*, 3759–3771. [CrossRef]
44. Yeats, T.H.; Huang, W.; Chatterjee, S.; Viart, H.M.F.; Clausen, M.H.; Stark, R.E.; Rose, J.K. Tomato Cutin Deficient 1 (CD 1) and putative orthologs comprise an ancient family of cutin synthase-like (CUS) proteins that are conserved among land plants. *Plant J.* **2014**, *77*, 667–675. [CrossRef]
45. Buda, G.J.; Barnes, W.J.; Fich, E.A.; Park, S.; Yeats, T.H.; Zhao, L.; Domozych, D.S.; Rose, J.K. An ATP binding cassette transporter is required for cuticular wax deposition and desiccation tolerance in the moss Physcomitrella patens. *Plant Cell* **2013**, *25*, 4000–4013. [CrossRef]
46. Chen, G.; Komatsuda, T.; Ma, J.F.; Nawrath, C.; Pourkheirandish, M.; Tagiri, A.; Hu, Y.-G.; Sameri, M.; Li, X.; Zhao, X. An ATP-binding cassette subfamily G full transporter is essential for the retention of leaf water in both wild barley and rice. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 12354–12359. [CrossRef]
47. Fabre, G.; Garroum, I.; Mazurek, S.; Daraspe, J.; Mucciolo, A.; Sankar, M.; Humbel, B.M.; Nawrath, C. The ABCG transporter PEC1/ABCG32 is required for the formation of the developing leaf cuticle in Arabidopsis. *New Phytol.* **2016**, *209*, 192–201. [CrossRef]
48. Panikashvili, D.; Shi, J.X.; Bocobza, S.; Franke, R.B.; Schreiber, L.; Aharoni, A. The Arabidopsis DSO/ABCG11 transporter affects cutin metabolism in reproductive organs and suberin in roots. *Mol. Plant* **2010**, *3*, 563–575. [CrossRef]
49. Broun, P.; Poindexter, P.; Osborne, E.; Jiang, C.-Z.; Riechmann, J.L. WIN1, a transcriptional activator of epidermal wax accumulation in Arabidopsis. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 4706–4711. [CrossRef]
50. Nadakuduti, S.S.; Pollard, M.; Kosma, D.K.; Allen, C.; Ohlrogge, J.B.; Barry, C.S. Pleiotropic phenotypes of the sticky peel mutant provide new insight into the role of CUTIN DEFICIENT2 in epidermal cell function in tomato. *Plant Physiol.* **2012**, *159*, 945–960. [CrossRef]
51. Raffaele, S.; Vailleau, F.; Léger, A.; Joubès, J.; Miersch, O.; Huard, C.; Blée, E.; Mongrand, S.; Domergue, F.; Roby, D. A MYB transcription factor regulates very-long-chain fatty acid biosynthesis for activation of the hypersensitive cell death response in Arabidopsis. *Plant Cell* **2008**, *20*, 752–767. [CrossRef]
52. Kosma, D.K.; Bourdenx, B.; Bernard, A.; Parsons, E.P.; Lü, S.; Joubès, J.; Jenks, M.A. The impact of water deficiency on leaf cuticle lipids of Arabidopsis. *Plant Physiol.* **2009**, *151*, 1918–1929. [CrossRef]
53. Bonaventure, G.; Beisson, F.; Ohlrogge, J.; Pollard, M. Analysis of the aliphatic monomer composition of polyesters associated with Arabidopsis epidermis: occurrence of octadeca-cis-6, cis-9-diene-1, 18-dioate as the major component. *Plant J.* **2004**, *40*, 920–930. [CrossRef]

54. Saladié, M.; Matas, A.J.; Isaacson, T.; Jenks, M.A.; Goodwin, S.M.; Niklas, K.J.; Xiaolin, R.; Labavitch, J.M.; Shackel, K.A.; Fernie, A.R. A reevaluation of the key factors that influence tomato fruit softening and integrity. *Plant Physiol.* **2007**, *144*, 1012–1028. [CrossRef]
55. Holloway, P. Structure and histochemistry of plant cuticular membranes: an overview. *Plant Cuticle* **1982**, *36*, 1–13.
56. Bindea, G.; Mlecnik, B.; Hackl, H.; Charoentong, P.; Tosolini, M.; Kirilovsky, A.; Fridman, W.-H.; Pagès, F.; Trajanoski, Z.; Galon, J. ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **2009**, *25*, 1091–1093. [CrossRef]
57. Xie, C.; Mao, X.; Huang, J.; Ding, Y.; Wu, J.; Dong, S.; Kong, L.; Gao, G.; Li, C.-Y.; Wei, L. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* **2011**, *39*, W316–W322. [CrossRef]

International Journal of
Molecular Sciences

Article

Genomics-Enabled Analysis of *Puroindoline b2* Genes Identifies New Alleles in Wheat and Related *Triticeae* Species

Xiaoyan Li [1,†], Yin Li [2,†], Xiaofen Yu [1,†], Fusheng Sun [1], Guangxiao Yang [1,*] and Guangyuan He [1,*]

[1] The Genetic Engineering International Cooperation Base of Chinese Ministry of Science and Technology, Key Laboratory of Molecular Biophysics of Chinese Ministry of Education, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China; yanziahnu@163.com (X.L.); yuixf@hust.edu.cn (X.Y.); fufu4567@126.com (F.S.)

[2] Waksman Institute of Microbiology, Rutgers, the State University of New Jersey, 190 Frelinghuysen Road, Piscataway, NJ 08854, USA; yl737@waksman.rutgers.edu

* Correspondence: ygx@hust.edu.cn (G.Y.); hegy@hust.edu.cn (G.H.)

† These authors contributed equally to this work.

Received: 17 January 2020; Accepted: 11 February 2020; Published: 14 February 2020

Abstract: Kernel hardness is a key trait of wheat seeds, largely controlled by two tightly linked genes *Puroindoline a* and *b* (*Pina* and *Pinb*). Genes homologous to *Pinb*, namely *Pinb2*, have been studied. Whether these genes contribute to kernel hardness and other important seed traits remains inconclusive. Using the high-quality bread wheat reference genome, we show that PINB2 are encoded by three homoeologous loci *Pinb2* not syntenic to the *Hardness* locus, with *Pinb2-7A* locus containing three tandem copies. PINB2 proteins have several features conserved for the *Pin*/*Pinb2* phylogenetic cluster but lack a structural basis of significant impact on kernel hardness. *Pinb2* are seed-specifically expressed with varied expression levels between the homoeologous copies and among wheat varieties. Using the high-quality genome information, we developed new *Pinb2* allele specific markers and demonstrated their usefulness by 1) identifying new *Pinb2* alleles in *Triticeae* species; and 2) performing an association analysis of *Pinb2* with kernel hardness. The association result suggests that *Pinb2* genes may have no substantial contribution to kernel hardness. Our results provide new insights into *Pinb2* evolution and expression and the new allele-specific markers are useful to further explore *Pinb2*'s contribution to seed traits in wheat.

Keywords: wheat; wheat genome; kernel hardness; *Puroindoline*; *Puroindoline b-2* variants; genotype-to-phenotype association; synteny; phylogenetic analysis

1. Introduction

Wheat (*Triticum aestivum* L.) is one of the major stable crops on Earth, feeding around 40% of the world population. Kernel hardness directly affects a set of physical and chemical properties of wheat seeds, such as water absorption, starch damage and flour particles [1,2]. Kernel hardness is one of the key traits in wheat seeds, largely determining milling quality and influencing the end-use qualities [3,4]. The major genetic determinant of wheat kernel hardness is the *Hardness* locus (*Ha*) on chromosome 5DS, harboring three closely linked genes: *Grain Softness Protein-1* (*GSP-1*), *Puroindoline a* (*Pina*) and *Puroindoline b* (*Pinb*). The starch granule-associated protein, Friabilin, determines kernel hardness and is composed of PINA and PINB, encoded by the *Pina* and *Pinb* genes, respectively [5–7]. Genetic studies and transgenic studies show that wildtype PINA and PINB proteins confers a soft kernel phenotype [8–16]. Certain *Pina* or *Pinb* mutations and their combinations lead to hard kernel phenotypes, with slight variation in kernel hardness between the alleles of *Pina* and/or *Pinb* [17]. As the major causal genes of kernel hardness, the allelic diversity of *Pina*/*Pinb* has been extensively investigated in a wide range of wheat germplasm, revealing 26 alleles of *Pina* and 33 of *Pinb*, as well as a few double null alleles [18–25]. Genotype–phenotype association studies show that the diverse *Pin* alleles are related to the phenotypic variations in kernel hardness.

Due to the functional importance of PINs in wheat, efforts have been made to search for genes that are potentially homologous to *Puroindolines* using the wheat Expressed Sequence Tag (EST) database, resulting in the identification of several transcripts known as *Pinb*-like genes (also known as *Pinb-2v*) [26,27]. Due to their high sequence similarity to *Pinb* (~60%), the molecular characterization of these *Pinb*-like genes and genotype–phenotype correlations have drawn research attentions. Physical mapping of the *Pinb-2v* genes, using several wheat genetic stocks, proves that *Pinb-2v1* is located on chromosome 7D, and *Pinb-2v2* and *Pinb-2v3* are allelic and located on 7B, with *Pinb-2v4* or *Pinb-2v5* located on 7A [27–30], indicating that these *Pinb-2v* genes might be homoeologous.

Mining for *Pinb-2v* genotypic diversity revealed 23 variants, suggesting less sequence diversity compared to *Pina* or *Pinb* [29,31]. With the development of *Pinb-2v* genotyping primers specific for *Pinb-2v1*, *2v2*, *2v3*, and *2v4* genotype–phenotype association has been studied in several collections of wheat varieties, from which different results were reported [28,29,32,33]. Chen et al. found that *Pinb-2v3* was associated with preferable grain traits and higher kernel hardness compared to those possessing *Pinb-2v2* in soft wheat varieties [26]. The results of *Pinb-2v*'s impacts on kernel hardness have been supported by association analysis using other wheat populations [34]. Another association mapping study that considered population structure and kinship showed that *Pinb-2v* variants were associated with semolina extraction but not kernel hardness [32]. By contrast, a genotype–phenotype study surveying representative U.S. wheat accessions did not support an ascertained role of *Pinb-2v* in kernel hardness [33]. An explanation for this could be that the expression levels of *Pinb-2v* variants (*2v1*, *2v2*, *2v3* and *2v4*) were much lower compared to that of *Pinb*, while varied expression levels were also observed among the *Pinb-2v* variants [35].

It is still inconclusive whether the *Pinb-2v* loci or particular alleles could influence kernel hardness or other kernel traits, even at a lesser extent than *Pina*/*Pinb*. Nevertheless, the prevalent *Pinb-2v* variants identified thus far have limited sequence polymorphisms within open reading frames (ORFs). The *Pinb-2v* genotyping primers were specific for variants *2v1*, *2v2*, *2v3* and *2v4*, with *Pinb-2v2* and *Pinb-2v3* being the most frequently genotyped. Previously, the genomic resources were unavailable to address these critical issues of *Pinb-2v* variants. More recently, owing to the advance in new genomics technologies, such as NRGene deNovoMagic, PacBio single-molecule sequencing, and optical mapping, the contiguous and well-ordered genome assemblies with high-quality annotations are now available in several *Triticeae* species, including bread wheat (*Triticum aestivum*), wild emmer wheat (*Triticum turgidum* spp. *dicoccoides*), the progenitor of the wheat D genome, *Aegilops tauschii,* and that of the wheat A genome *Triticum urartu* [36–40]. These advanced genomics technologies have also substantially contributed to several unambiguously reconstructed, contiguous monocot genomes (maize, broomcorn millet, sugarcane, etc.), as well as to the chromosomal regions with tandem gene clusters and extensive haplotype variations in the gene family [41–44]. With the bread wheat genome (International Wheat Genome Sequencing Consortium (IWGSC) RefSeq v1.0 for the cultivar Chinese Spring (CS42)), we show that the *Pinb-2v* genes are encoded by five gene models at three homoeologous loci, *Pinb2-7A*, *Pinb2-7B* and *Pinb2-7D*, with the *Pinb2-7A* locus containing three tandemly duplicated copies, *Pinb2-7A1*, *Pinb2-7A2*, and *Pinb2-7A3*. We propose the new nomenclature of *Pinb-2v* for *Pinb2*, so as to be consistent with the IWGSC RefSeq v1.0 annotation (Table S1) [36]. We establish a series of new PCR-based genotyping primers for *Pinb2* genes. Genome-wide analysis of *Pinb2* genes and applications of the *Pinb2* genotyping method provide new insights into this gene family.

2. Results

2.1. Genomic and Phylogenetic Analyses of Pinb2 Genes

The DNA and predicted protein sequences of known *Pinb2* variants were used to identify gene models that could be *Pinb*-like genes, annotated in the IWGSC bread wheat genome RefSeq v1.0. Five gene models were identified, two on the B and D genomes (TraesCS7B02G431200 and TraesCS7D02G504800, respectively), with the other three on the A genome in a tandemly manner (TraesCS7A02G514400, TraesCS7A02G514500 and TraesCS7A02G514505). Sequence alignments confirmed their high sequence similarity with the kernel hardness-related genes *Pinb* and *Pina*, as well as the gene encoding *GSP-1* on the *Ha* locus (Figures S1 and S2).

We further expanded our BLAST search to several recently published reference genomes of the *Triticeae* species and identified the *Pinb2* homolog, AET0Gv20021600, in the *Ae. tauschii* genome of accession AL8/78; TuG1812G0700005557.01 in the *T. urartu* genome of accession G1812; and TRIDC7BG068420.1 in the B genome of wild emmer wheat accession Zavitan; as well as a tandem cluster of three *Pinb2* homologous sequences in the A genome of wild emmer wheat, which lack annotations as gene models. The nomenclature of *Pinb2* alleles and the corresponding variants, reported previously, are shown in Table S1 with the suffixes 7A, 7B, and 7D used to indicate the chromosomal locations. The gene models corresponding to *Pinb2* genes from the *Triticeae* species are listed in Table S2.

Due to the high sequence similarity between *Pinb* and *Pinb2*, it has long been considered that *Pinb2* might be associated with kernel hardness. Therefore, we sought to address its relationship with *Pinb* by using synteny alignment. The gene order at the *Ha* locus in the IWGSC RefSeq v1.0 assembly of the bread wheat genome is consistent with previous analyses of the Bacterial Artificial Chromosome (BAC) library containing the *Ha* locus. The *BGGP* gene, encoding b-1-3-galactosyl-O-glycosyl-glycoprotein, and the *GSP-1* gene, encoding Grain Softness Protein-1, are located at the 5′ end of *Ha* (these genes indicated as 'BG' and 'Gs', respectively, in Figure 1, Table S3) [45–47].

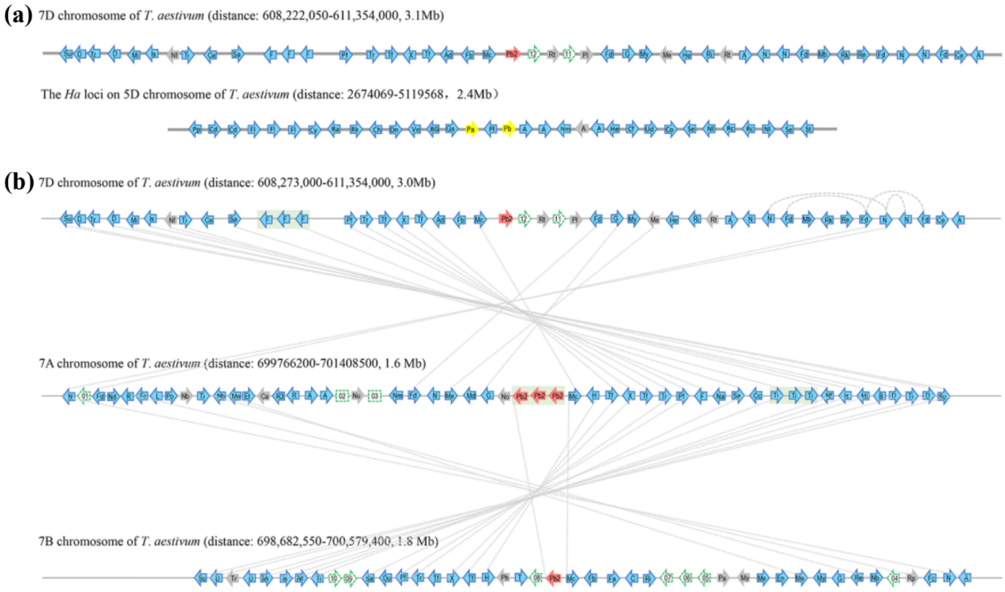

(a)
7D chromosome of T. aestivum (distance: 608,222,050-611,354,000, 3.1Mb)
The Ha loci on 5D chromosome of T. aestivum (distance: 2674069-5119568，2.4Mb)
(b)
7D chromosome of T. aestivum (distance: 608,273,000-611,354,000, 3.0Mb)
7A chromosome of T. aestivum (distance: 699766200-701408500, 1.6 Mb)
7B chromosome of T. aestivum (distance: 698,682,550-700,579,400, 1.8 Mb)

Figure 1. Syntenic alignments of bread wheat chromosomal segments that contains *Pinb2-7D* locus at 7DL and *Ha* locus at 5DS (**a**), and those that contains *Pinb2-7D*, *Pinb2-7A*, *Pinb2-B* at 7DL, 7AL and 7BL, respectively (**b**). Wheat chromosomes are represented as grey horizontal lines, whereas genes along the chromosome are represented as arrowheads. *Pina-D1* and *Pinb-D1* are represented as yellow arrowheads. *Pinb2* are represented as red arrowheads, while the other protein-coding genes with high confidence annotations in the IWGSC RefSeq v1.0 bread wheat genome are represented as blue arrowheads. Homoeologous gene pairs are indicated as dotted lines connecting arrowheads. Tandem duplication of the three *Pinb2-7A* genes is shaded in the grey box. To simplify visualization, only high-confidence protein-coding genes are shown with their orders and orientations along the chromosomal segments, shown as the same in wheat genome assembly. The intergenic regions are not in proportion to the wheat genome assembly. For visualization, only abbreviations of the genes are labeled on the arrowheads, with their full names provided in Table S3.

Two clusters of *ATPase* genes with a *Nodulin* gene in between are located at the 3′ downstream of *Pina* and *Pinb* (these genes are indicated as 'A' and 'Nm', respectively, in Figure 1). Comparison of the genes flanking *Pina*/*Pinb* and *Pinb2* failed to identify clear synteny between the *Ha* and the *Pinb2-7D* locus (Figure 1, Table S3). We then analyzed the homologous regions containing the *Pinb2* genes at 7AL, 7BL, and 7DL. Alignment of the *Pinb2* flanking genes revealed a clear collinearity between the segments containing *Pinb2-7A*, *Pinb2-7B*, and *Pinb2-7D* (Figure 1, Table S3). Interestingly, three copies of *Pinb2-7A* with their flanking sequences were arranged in a cluster, suggesting their origin from tandem duplication.

Next, we investigated the collinearity of *Pinb2*-containing segments among the sequenced *Triticeae* species. A comparison of the orthologous segments on chromosome 7A showed collinearity at this region between bread wheat, wild emmer wheat, and *T. urartu*, while only bread and wild emmer wheats have three copies of *Pinb2-7A* (Figure S3a, Table S3). Comparison of the segments on chromosome 7B, between bread and wild emmer wheats, showed a good collinearity between the regions (Figure S3b, Table S3). Furthermore, *TaPinb2-7D1* is contained within a segment that is collinear with an orthologous region of *Ae. tauschii* containing *AtPinb2-7D1* (Figure S3c, Table S3). The *Pinb2-7A* cluster is only detected in polyploidy wheat species, but not in the A-genome and D-genome donor species, *T.urartu* and *Ae. tauschii*, suggesting that the tandem duplication of *Pinb2-7A* could emerge after the polyploidization of wheat.

While there is a lack of synteny between the genomic segments harboring *Pinb2* and the *Pina*/*Pinb* genes, the *Pinb2* genes indeed have a high sequence similarity with *Pinb*, indicating that *Pinb2* might have the sequence basis for kernel hardness or interactions with other seed proteins. Therefore, we analyzed its phylogeny together with other seed proteins, many of which belong to the prolamin superfamily [48]. The phylogenetic results separated wheat globulins and albumins from those in the prolamin superfamily (Figure 2a, Table S4). PINB2 proteins were clustered together with PINA, PINB and GSP. The puroindoline cluster was grouped together with a cluster largely consisting of α-amylase inhibitors (ATIs). As the repetitive sequences exist in many proteins of the prolamin superfamily and could be problematic for sequence alignment or annotation, we therefore used the domain signatures and conserved cysteine residue patterns to help define the phylogenetic clusters of wheat seed proteins. The results showed that almost all the phylogenetically analyzed prolamin proteins, including PINB2, have Gliadin and Tryp_alpha_amyl domains (PF13016 and PF00234, respectively) except that the α-amylase inhibitors only contain domain PF00234 and high-molecular-weight glutenin subunits (HMW-GSs) contain the HMW domain (PF03157). In addition, the conserved cysteine-rich patterns and number of cysteine residuals can be used as characteristics for defining groups of storage proteins within the prolamin superfamily [43]. For example, HMW-glutenin x- and y-type subunits contain four to six cysteines, while α-gliadin and γ-gliadin have six to eight and eight to eleven cysteines, respectively (Figure S4). Type a-, b- and c- avenin-like proteins (ALPs) have 14, 18–20, and 10–12 cysteine residues, respectively, and some ALPs have effects on dough quality, likely due to the number of cysteine residues for interaction with wheat storage proteins (Figure S4) [49–52]. Here, PINB2 identified in the bread wheat genome of CS shares the same cysteine-rich backbone consisting of 10 residues as PINA, PINB, and GSP (Figure S4).

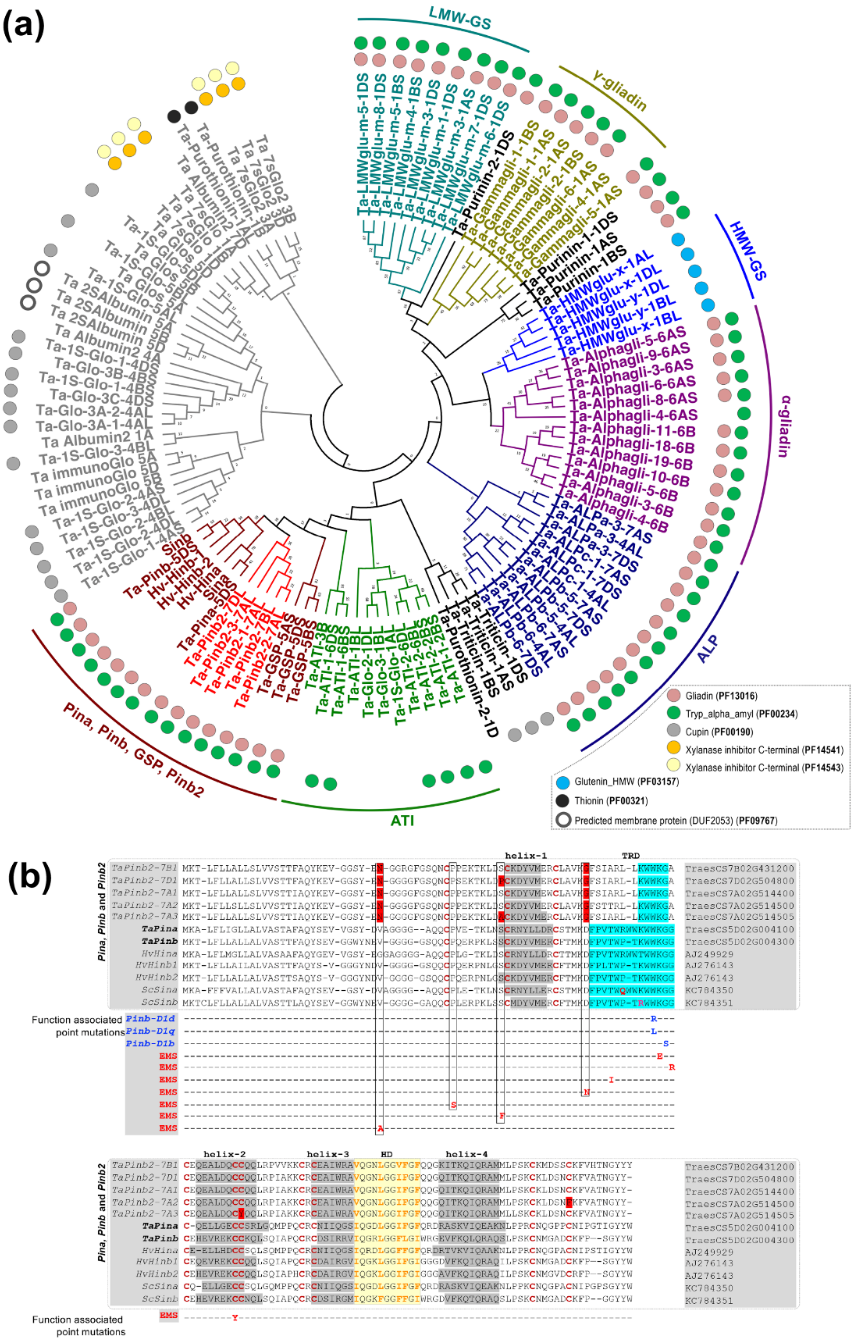


Figure 2. Analyses of phylogeny and protein sequence alignment for PINB2, together with other wheat seed proteins. (**a**) Maximum likelihood (ML) phylogeny of HMW-GS, LMW-GS, α-gliadin, γ-gliadin, purinin, purothionin, avenin-like proteins (ALPs), PINA, PINB and GSP from the bread wheat CS and the homologous proteins of PINA, PINB and GSP from triticale and barley. Results for protein domain analysis using profile hidden Markov Models (HMMER) are shown as color-coded circles. (**b**) Alignment of the amino acid sequences between PINA, PINB, GSP, PINB2, from wheat, barley and triticale, highlights that PINB2 proteins share the cysteine residue backbone, five α-helixes and hydrophobic domain with PINs but lack TRD domain. PINB2 proteins also contain some changes in the conserved amino acids that are important for PINs' function in kernel hardness (those amino acids shaded in red for PINB2 proteins). α-helixes, hydrophobic domain and TRD are shaded in grey, yellow and blue, respectively. The cysteine residues are shown in red.

We performed a detailed sequence comparison between *Pinb2* and *Pina*, *Pinb*, *puroindoline*-like genes in barley and hexaploid triticale (*Hordoindolines*, *Hin*, and *Secaloindolines*, *Sin*), as well as those *Pinb* alleles that naturally exist or are mutagenized by ethylmethanesulfonate (EMS) with functional changes (Figure 2b) [53–58]. The *puroindoline*-like genes in barley include *Hina*, *Hinb1* and *Hinb2* with *Hinb*s, but not *Hina* associated with grain hardness, while the *puroindoline*-like genes in hexaploid triticale (AABBRR), *Sina* and *Sinb*, don't affect grain hardness [53,56–58]. The point mutations identified in these natural and EMS-mutagenized alleles of *Pinb* have been reported to be associated with kernel hardness and were summarized in our previous study [4]. The results showed that PINB2s encoded by all three homoeologous genes had the four helixes and a hydrophobic domain (HD), as with PINA/PINB and their homologs in barley and triticale, but lacked a functional tryptophan-rich domain (TRD), which is key for interactions with polar lipids on starch granule surfaces [59]. Additionally, PINB2s had distinct amino acid residues at several conserved locations that were proved important for PINB's function in kernel hardness by genetic analyses [4,54,55]. Despite these important sequence differences compared to PINs, PINB2s have the conserved HD domain which could be involved in protein–protein interactions in Puroindolines (Figure 2b) [59,60]. In summary, analyses of PINB2s's sequence signatures (domains, cysteine residue patterns, etc.) and its phylogeny emphasize the sequence and structural similarities between PINB2 and PINA/PINB but also point out the sequence distinction, suggesting one of the probable reasons for PINB2 not being associated with obvious kernel hardness differences.

2.2. Expression Levels of *Pinb2*

High expression levels of *Pina* and *Pinb* during seed development are important for their functions in kernel hardness. We analyzed the *Pinb2* expression dynamics in wheat using the many publicly available wheat RNA-seq datasets. The RNA-seq datasets from five different wheat varieties showed that *Pinb2* genes were expressed at much lower levels compared to *Pinb* (Figure 3; RNA-seq datasets, listed in Table S5 and briefly described in Materials and Methods). For example, *Pinb* was expressed at ~20,000 transcripts per million (TPM) in CS, while *Pinb2-7B1* and *Pinb2-7D1* were expressed at less than 1500 TPM (Figure 3d). Here, to visualize the expression levels of *Pinb2* genes across different RNA-seq experiments or libraries, TPM was used as the expression measure, since it has been suggested to be more consistent and accurate in reflecting the quantitative differences of genes across wheat RNA-seq samples with the consideration that *Pinb2* genes are comprised of a single-exon and short [61,62]. The expression levels of *Pinb2-7A1*, *-7A2*, -7A3, *-7B1* and *-7D1* differed from each other, with *Pinb2-7D1* being the one with a relatively higher expression, followed by *Pinb2-7B1*. *Pinb2-7A3* was expressed at a low level, while *Pinb2-7A1* and *Pinb2-7A2* were expressed at extremely low levels around the RNA-seq expression threshold (>=0.5 TPM), except for their expression in cv. Holdfast (Figure 3c). Previous RNA-seq data of the wheat variety Azhurnaya covers a wide range of tissues and developmental stages, representing a comprehensive wheat expression atlas [62]. Azhurnaya RNA-seq data showed that *Pinb2* genes were highly expressed in several samples of developing seeds, particularly during soft and hard dough stages, with *Pinb2-7D1* most highly expressed followed by *Pinb2-7B1* and *Pinb2-7A1* (Figure 3a). The results also exhibited a very low expression of *Pinb2-7A2* in leaf and root tissues. These RNA-seq data highlights that the expression levels of *Pinb2* genes vary between loci (Figure 3). To explore the *Pinb2* expression across wheat varieties, we visualized the *Pinb2* expression levels together with a reference gene, *TaActin* (TraesCS1D02G274400; Figure S5) [63]. The reference gene TaActin showed relative stable expression for the seed samples within each variety except for cv. Holdfast (Figure S5a). To take into account for potential RNA-seq batch effects between varieties, the TPM expression values of Pinb2 genes were normalized to TaActin. The results showed that Pinb2-7D1 was highly expressed in Holdfast, followed by that in CS, then Azhuranaya and Zhou 8425B, while Pinb2-7B1 was highly expressed in CS, followed by that in Holdfast, then Azhurnaya and Zhou 8425B (Figure S5b). Both TPM expression and relative expression data suggest that the expression levels of *Pinb2* genes vary between varieties. Such varied expression levels of *Pinb2* indicates potential regulatory differences between the loci and varieties, consistent with the much lower sequence similarities in the promoter regions than in the ORFs (Figure S1). Quantitative RT-PCR results validated the lower expression levels of *Pinb2* in comparison with *Pinb* (Figure 3f, Table S6). Our qRT-PCR results also confirmed that *Pinb2* genes were specifically expressed in the developing endosperm of CS (Figure 3g) [35].

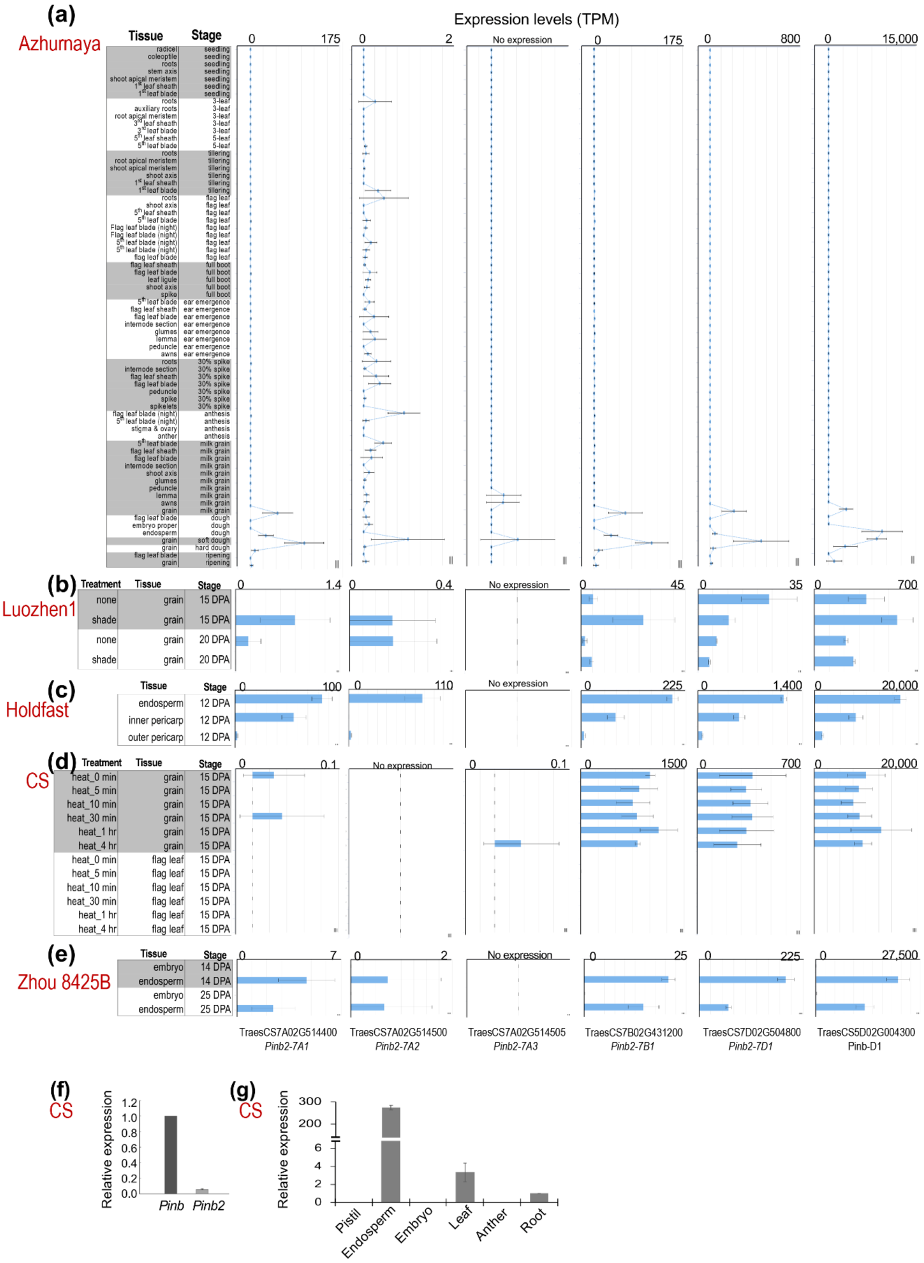


Figure 3. Expression patterns of *Pinb2* genes. Expression levels of *Pinb2* genes and *Pinb-D1* from five publicly available RNA-seq datasets and shown in bar charts (**a**–**e**). The varieties used for RNA-seq analyses are indicated in red, while the tissues, developmental stages and conditions for each RNA-seq experiments are labeled on the left of each panel. Expression levels are shown using TPM and are not normalized across datasets. (**f**) The relative expression levels of *Pinb* and *Pinb2* genes were quantified using qRT-PCR at 28 Days Post Anthesis (DPA) in CS seeds. (**g**) Expression levels of *Pinb2* genes in different tissues from CS proved *Pinb2* genes are expressed specific in seeds. Detailed information of the RNA-seq datasets used are given in Table S5.

2.3. Genotyping of *Pinb2*

PCR methods to amplify the sequences of *Pinb*-like variants in wheat have been established, and these contribute to our understanding of *Pinb*-like variants [26–29]. Particularly, primers of PCR or derived cleaved amplified polymorphic sequence (dCAPS) assays detecting certain *Pinb2-7B1* or *Pinb2-7D1* variants were reported and have been widely used [32–34]. The use of *Pinb2* PCR markers has significantly helped to identify new variants, to study genetic diversity and to facilitate association analysis of *Pinb2* with seed traits, including kernel hardness.

The previous *Pinb2* PCR primers target within the ORF regions, where the sequence variation is much lower compared to the flanking regions, where very few SNPs can be used to distinguish certain variants (Figure S1). With the high-quality reference genomes of several *Triticeae* species, we compared the sequences of *Pinb2* homoeologous genes and designed several new primer pairs for amplifying and detecting *Pinb2*. These PCR primers target *Pinb2*-flanking regions to ensure PCR specificity (Table S7).

Among the newly designed *Pinb2* primers, primer pair D can be used to amplify *Pinb2-7D1*, and primer pair U can be used to amplify all three homoeologous *Pinb2* genes (Figure 4f). Primer pair B1 is designed to specifically amplify the *Pinb2-7B1-v3* allele as a ~1326 bp PCR product (Figure 4c). Nested primer pairs B2 and B3 can be used to detect *Pinb2-7B1-v2* alleles [64]. Further, a pair of dCAPS primers (primer pair C) was designed specifically for the *Pinb2-7B1-v2-1* allele. *Bst*X I digestion, after amplification by primer pair C, distinguishes the *Pinb2-7B1-v2-1* allele (107-bp product) from the *Pinb2-7B1-v2* allele (135-bp product), which only has an SNP difference (Figure 4e). Therefore, primer pairs B1, B2/B3 and the dCAPS primers form a sequential pipeline for detecting the alleles of *Pinb2-7B1*.

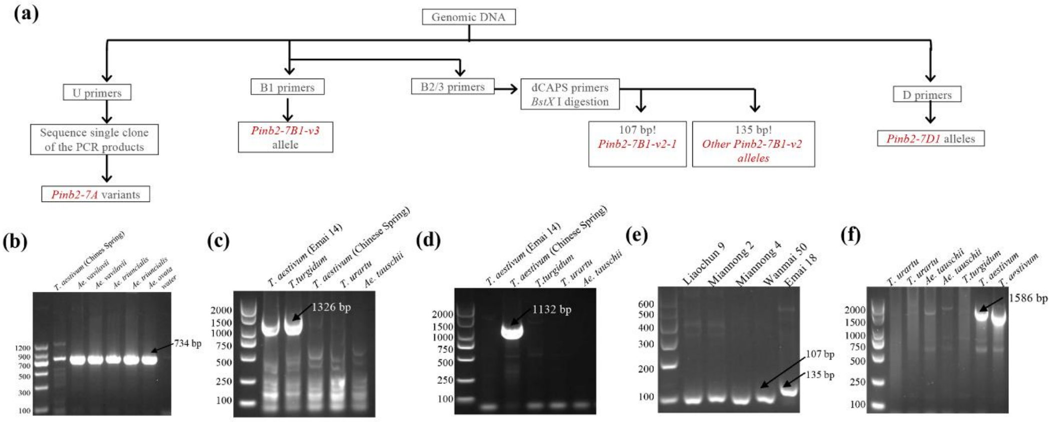


Figure 4. Genotyping of *Pinb2* genes using the new PCR method. (**a**) Diagram describing the workflow of set of PCR primers to genotype *Pinb2-7A*, *Pinb2-7B* and *Pinb2-7D* loci. (**b**) Primer pair U non-specifically amplified *Pinb2* genes encoded by *Pinb2-7A*, *Pinb2-7B* and *Pinb2-7D* loci from several *Triticeae* species. Sizes of DNA markers used are labeled. Specific PCR products are pointed out by black arrowheads. (**c**) PCR primer pair B1 specifically detected the *Pinb2-7B1-v3* allele in bread and durum wheats (*T. aestivum* and *T. turgidum* spp. *durum*). (**d**) Nested PCR primer sets B2/B3 specifically detected the *Pinb2-7B1-v2* allele in bread wheat (*T. aestivum*) Chinese Spring. (**e**) The dCAPS primers-derived PCR amplification and subsequent *BstX* I digestion distinguished the *Pinb2-7B1-v2-1* allele (107 bp) from the other *Pinb2-7B1-v2* alleles (135 bp). (**f**) PCR primer pair D specifically amplified the *Pinb2-7D1* gene in bread wheat.

To validate the primers' specificity, several accessions of bread wheat (*T. aestivum*), durum wheat (*T. turgidum* spp. *durum*), *Ae. tauschii* and *T. urartu* were used for PCR. Indeed, primer pair B1 amplified the *Pinb2-7B1* genes from bread wheat Emai 14 and durum wheat Ofanto but did not detect the *Pinb2-7B1-v2* allele from CS (Figure 4b,c). The sequencing results of purified PCR products showed that the *Pinb-2v* on chromosome 7B of Emai14 and *T. durum* are *Pinb-B2v3b* and *Pinb-B2v3a*, respectively. The dCAPS primer pair followed by enzyme digestion specifically distinguished the *Pinb2-7B1-v2-1* allele (with a 107-bp product) from other *Pinb2-7B1* alleles (with a 135-bp product, Figure 4e). Additionally, primer pair D only detected the *Pinb2-7D1* genes in bread wheat as designed (Figure 4f), while primer pair U had amplification for all the accessions used (Figure 4b). Overall, these PCR validation results demonstrated the feasibility of our *Pinb2* genotyping pipeline, as illustrated in Figure 4a.

We then sought to show the usefulness of our *Pinb2* genotyping primers with two examples. First, we explored the diversity of *Pinb2* genes using 70 selected bread wheat accessions and several accessions of *Ae. tauschii*, *Aegilops vavilovii*, *Aegilops triuncialis*, *Aegilops ovata* and *T. urartu*. A new *Pinb2-7D1* allele was amplified from the bread wheat varieties Zhengmai 101, Wanke 06229, Jimai 107, and Laoqimai, designated as *Pinb2-7D1-v1-6*. Four new alleles of *Pinb2-7D*1 were obtained from *Ae. cylindrical*, *Ae. vavilovii*, *Ae. triuncialis* and *Ae. geniculate*, designated as *Pinb2-7D1-v1-8*, *Pinb2-7D1-v1-9*, *Pinb2-7D1-v1-10* and *Pinb2-7D1-v1-11*, respectively (Figure 4b). We sequenced the five new *Pinb2* alleles, *Pinb2-7D1-v1-6*, *Pinb2-7D1-v1-8*, *Pinb2-7D1-v1-9*, *Pinb2-7D1-v1-10*, and *Pinb2-7D1-v1-11*. Sequence alignment and phylogenetic analysis of the new *Pinb2* alleles with the previously identified *Pinb2* sequences showed that the *Pinb2* sequences are clustered according to the A, B and D genomes (Figure 5 and Figure S6).

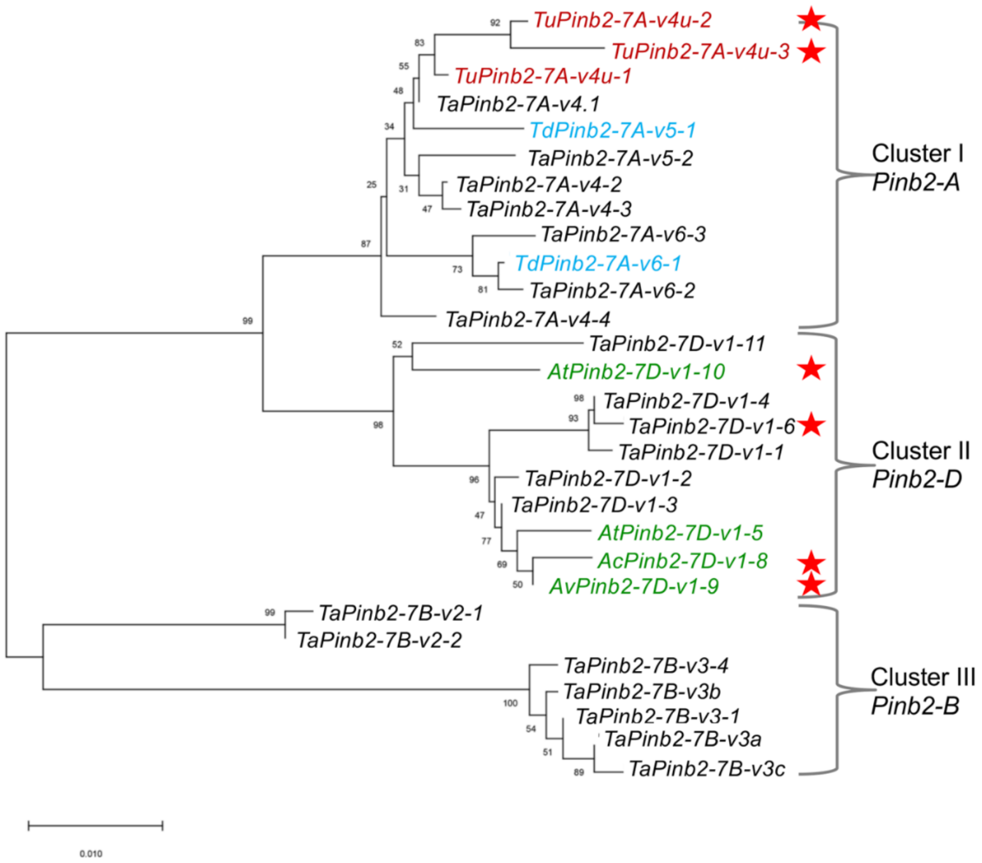


Figure 5. Phylogenetic tree of *Pinb2* variants highlights that these sequences are clustered according to the loci where they are located. Phylogeny of *Pinb2* alleles was constructed using the neighbor-joining method with 1000-time bootstrap and the substitution model of Kimura 2. Different *Pinb2* alleles from different *Triticeae* species in this study are labeled in different colors. The *Pinb2* alleles from durum wheats were labeled in blue color, bread wheat in black, *T. urartu* in red, *Ae. tauschii* in green. The new alleles identified in this study are indicated using red stars.

2.4. Association of Pina, Pinb and Pinb2 with Wheat Kernel Hardness

As another application of our *Pinb2* genotyping primers, we genotyped *Pinb2-7B1* and *Pinb2-7D1* loci for 70 Chinese bread wheat varieties, of which the kernel hardness phenotypes and *Pina*/*Pinb* genotypes were reported previously [4]. The 70 varieties were selected based on their highly correlated kernel hardness phenotypes between two field seasons (2016–2017 and 2017–2018) and their uniform *Pina-D1a* genotype, so as to simplify the genotype–phenotype association analysis. The genotypes of *Pinb2-7B1* and *Pinb2-7D1* for these 70 varieties are given in Table S8, with the *Pin* and *Pinb2* frequencies shown in Table 1.

Table 1. The distribution and frequency of *Pina*, *Pinb* and *Pinb2* for the 70 Chinese wheat varieties.

Genotypes		*Pinb2-7D1*			*Pinb2-7B1*			
		2v1-4	*2v1-3*	*2v1-6*	*2v2-1*	*2v3a*	*2v3b*	*2v3c*
Pina-D1a/	Number	19	2	1	6	5	11	0
Pinb-D1a	Freq. (%)	86.4	9.1	4.5	27.3	22.7	50	0
Pina-D1a/	Number	24	8	2	7	6	17	4
Pinb-D1b	Freq. (%)	70.6	23.5	5.9	20.6	17.6	50	11.8
Pina-D1a/	Number	9	4	1	6	0	7	1
Pinb-D1p	Freq. (%)	64.3	28.6	7.1	42.9	0	50	7.1
Total	Number	52	14	4	19	11	35	5
	Freq. (%)	74.3	20	5.7	27.1	15.7	50	7.1

For the *Pinb2-7B1* locus, four alleles (*Pinb2-7B1-v2-1*, *Pinb2-7B1-v3a*, *Pinb2-7B1-v3b* and *Pinb2-7B1-v3c*) were detected and *Pinb2-7B1-v3b* was the most prevalent allele (50%) followed by *Pinb2-7B1-v2-1*. For the *Pinb2-7D1* locus, three alleles (*Pinb2-7D1-v1-3*, *Pinb2-7D1-v1-4*, and *Pinb2-7D1-v1-6*) were identified, with *Pinb2-7D1-v1-4* being the most predominant allele. While all of the 70 varieties uniformly contained the *Pina-D1a* allele, these varieties have three *Pinb* alleles, *Pinb-D1a*, *Pinb-D1b*, and *Pinb-D1p*, of which the latter two are causal for the hard-kernel phenotype due to a point mutation and a lack of PINB protein, respectively (summarized by Li et al.) [4].

Among the 70 varieties, six had a medium-hard kernel (hardness index (HI) between 40 and 60) in one season but were detected as hard kernel (HI > 60) in the other season. These varieties were then classified as mixed kernel-hardness varieties (likely containing multiple *Pin-D1* haplotypes) [65–67] and excluded from further analysis. A *Chi*-square independence test showed that *Pinb-D1* was significantly associated with kernel hardness in both seasons ($P = 3.5 \times 10^{-11}$), and *Pinb2-7B1* was not correlated to kernel hardness (Table 2). *Chi*-square results didn't support a dependency between *Pinb2-7D1* and kernel hardness although a provisional significance was calculated ($P = 0.0503$). The results here highlight the usefulness of the new *Pinb2* genotyping method and its contribution to association studies of *Pinb2*. Whether *Pinb2-7D1* could play a minor role in kernel hardness, at least in some varieties under particular conditions, requires more large-scale, robust genetic analyses.

Table 2. *Chi*-square analysis of the association between *Pinb*, *Pinb2-7B1* or *Pinb2-7D1* loci with kernel hardness.

Alleles	2016–17			2017–18			*Chi*-Square*
	Soft	Medium	Hard	Soft	Medium	Hard	
Pinb-D1a	20	2	0	20	2	0	*Chi*-square = 54.8 *df* = 4 $P = 3.5 \times 10^{-11}$
Pinb-D1b	1	6	21	1	6	21	
Pinb-D1p	0	1	13	0	1	13	
Pinb2-7B1-2v2-1	6	3	9	6	3	9	*Chi*-square = 3.4 *df* = 6 *P* = 0.753
Pinb2-7B1-2v3a	5	1	5	5	1	5	
Pinb2-7B1-2v3b	10	4	16	10	4	16	
Pinb2-7B1-2v3c	0	1	4	0	1	4	
Pinb2-7D1-2v1-3	2	1	8	2	1	8	*Chi*-square = 9.5 *df* = 4 *P* = 0.0503
Pinb2-7D1-2v1-4	19	6	25	19	6	25	
Pinb2-7D1-2v1-6	0	2	1	0	2	1	

3. Discussion

Pinb2 is a group of important genes with several sequence features similar to the kernel hardness-determinant gene *Puroindolines,* though some previous results of association and expression do not support *Pinb2* as playing a major role in controlling wheat kernel hardness. Nevertheless, some genotype–phenotype analyses have indicated that *Pinb2* could be associated with kernel traits [26,34]. Recently, new insights into the functions of *Puroindolines* have emerged, including their abilities to interact with wheat gluten proteins and lipids, possibly through hydrophobic interactions and/or disulfide bonds besides the known TRD–polarlipid interaction [68–70]. PINB2 maintains the same cysteine-residue backbone as PIN (Figure 2 and Figure S4). Taken together, these pieces of information justify the need to explore the sequences and function of *Pinb2* in *Triticeae* species, as recently reported for PINs, which was previously mainly reliant on PCR-based genotyping using the limited sequence information of the *Pinb2* ORFs. To facilitate *Pinb2* analysis, we used the high-quality reference genomes of *T. aestivum* and other closely related *Triticeae* species to compare genomic synteny between *Pin* and *Pinb2*, phylogenetically position *Pinb2* among major groups of wheat seed proteins, and develop the new PCR primers for *Pinb2* genes/alleles. Similar types of studies, such as the large-scale identification of immunoresponsive allergens from wheat seed proteins and a genome-wide study of avenin-like proteins or MADS-box genes, were only made possible recently, thanks to the high-quality *Triticeae* genomes [48,51,71].

Using the high-quality wheat genome, we show that *Pinb2* loci are not colinear with the *Hardness* locus on chromosome 5DS, where *Pina* and *Pinb* reside. While there is a lack of synteny between the genomic segments, PINB2 proteins are, indeed, phylogenetically clustered with PINs as well as other ATIs and share the conserved cysteine residue backbone, helixes, and hydrophobic domain with PINs, suggesting an evolutionary relationship between these proteins. Sequence analysis highlights the lack of a functional TRD region and several important amino acids in PINB2, which are required for determining kernel hardness. The contrasting results between the synteny and phylogeny and sequence analyses allow us to speculate the possibility that *Pinb2* or *Puroindolines* might emerge in the progenitor of diploid *Triticeae* species, and a duplicated copy of *Pinb2* might reinsert into the *Ha* locus at chromosome 5DS, or vice versa, followed by an independent evolution for each loci as the divergence of diploid *Triticeae* species and polyploidization to form *Triticum aestivum*. This hypothesis would explain the syntenic alignment results and the emergence of unique TRD in Puroindolines. The evolutionary aspect of *Pinb2* and *Pin* went beyond the scope of our study and requires additional evolutionary and bioinformatics analyses using much broader genomics information, such as those from barley and rye.

Previous studies on the association of *Pinb2-7B* with kernel hardness drew somewhat different results and some results have indicated potential minor effects of certain *Pinb2-7B* alleles on kernel hardness [26,28,32–34]. *Pinb2* has also been reported to be associated with other kernel traits in wheat [26,32]. In the present study, we designed the new *Pinb2* PCR markers using the high-quality genome assemblies of bread wheat and several *Triticeae* species and proved the PCR markers to be useful. However, we acknowledged that the *Pinb2* PCR primers might not be perfect. For example, there are a few polymorphisms in the reverse U primer (Figure S1), although it did not affect the annealing of the primers and successful amplification using the wheat materials (Figure 4). Future work will be needed to improve the *Pinb2* genotyping primers toward a higher sensitivity and a broader adaptability for more wheat varieties and *Triticeae* species. With the new *Pinb2* genotyping primers described here, we expanded the genotyping ability of certain *Pinb2-7B* alleles to both three homoeologous *Pinb2* loci. The result of the *Chi*-square test didn't support the dependency of *Pinb2-7D* or *Pinb2-7B* on kernel hardness, although the association results of *Pinb2-7D* approached significance ($P = 0.0503$). As the major purpose of the present study is to conduct a comprehensive analysis of *Pinb2* genes using the high-quality wheat genome and to proof-of-concept the new *Pinb2* genotyping primers, rather than to draw definite solid conclusions for the association between *Pinb2* and kernel hardness, our interpretation comes with a caveat based on the possibility that the number of varieties may not be sufficient or the varieties used here may not represent broad enough genetic diversity to draw definite solid conclusions for genotype–phenotype association analysis. The marginal significance between *Pinb2-7D* and kernel hardness highlights the necessity of performing further genotype–phenotype association studies for the Pinb2 genes and to expand such appropriate studies to a larger and more representative collection of wheat accessions using the *Pinb2* genotyping primers reported here. The wheat accessions used previously are different between studies, also making it difficult to directly compare the results. Importantly, PINB2 proteins possibly hold the structure and/or sequence basis for interacting with gluten proteins or lipids [68–70,72].

More recently, studies provide evidence of PIN–gluten protein aggregation, likely through hydrophobic interactions, particularly the evidence for PIN–gliadin interaction affecting gliadins' aggregative properties [69,70]. Considering the variations in expression levels between *Pinb2* homoeologous copies observed in the RNA-seq data, it may be possible that certain PINB2 proteins with high expression levels could exert similar properties as PINs, interacting with gluten proteins or lipids through hydrophobic interactions. Thus, our results emphasize future research directions to address the issue whether PINB2 proteins could interact with gluten proteins and/or lipids as PINs, and whether such interactions would exert some effects on kernel hardness or other kernel traits.

4. Materials and Methods

4.1. Plant Material

A collection of seventy bread wheat varieties from the wheat-producing regions of the Yellow and Huai Valleys and Yangtze Valley of China, provided by the Institute of Crop Breeding and Cultivation of the Hubei Agricultural Academy of Science, and the Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, was used for genotyping *Pinb2* genes and for interrogating their association with kernel hardness using the new genotyping primers developed in the present study (Table S8). The kernel hardness phenotypes of these varieties, collected in the 2016–2017 and 2017–2018 field seasons, have been reported previously [4]. In both seasons, they were planted in the experimental field of Huazhong University of Science and Technology with a randomized complete block design consisting of five 200 cm long rows per accession.

Several *Triticum* and *Aegilops* species were used to validate the PCR specificity and to explore the genetic diversity of *Pinb2* genes; they include *Aegilops tauschii* (donor of the bread wheat A genome, AA), *Ae. vaviolovii*, *Ae. triuncialis*, *Ae. ovata*, *T. urartu* (donor of the bread wheat D genome, DD; accessions G1937 and G1906) and tetraploid *Triticum turgidum* spp. *durum* variety Ofanto (AABB). G1937 and G1906 were obtained from USDA-GRIN.

4.2. Phenotyping of Kernel Hardness

The kernel hardness index (HI) was determined with the Single Kernel Characterization System (SKCS) 4100 (Perten Instruments North America Inc., Springfield, IL, USA), following the American Association for Cereal Chemists (AACC) approved method as previously described [73]. Briefly, 200 kernels harvested from each replicate field plot were used for measuring HI. The wheat varieties were planted and measured in the 2016–2017 and 2017–2018 seasons, with two replicated plots for the 2016–2017 season and three replicated plots for the 2017–2018 season. These varieties were classified into soft-, medium-, and hard-kernels based on the HI value: varieties with HI less than 40 were the soft-kernel type, while those with HI greater than 60 were the hard-kernel type, with those with a HI value between 40 to 60 were the medium-kernel type.

4.3. Identification and Genotyping of Puroindoline-D1 and Puroindoline b-2 Variants

The genotyping of *Pin a* and *b* was described previously [4]. Six *Pinb2* variants reported previously, *Pinb-2v1*, *Pinb-2v2*, *Pinb-2v3*, *Pinb-2v4*, *Pinb-2v5*, and *Pinb-2v6*, were used to identify the gene models encoding *Pinb2* genes. Then, the coding sequences of *Pinb2* gene models and predicted protein sequences were obtained from the Gramene and *Triticeae* Multi-omics Center Website, the latter of which has been established by Shandong Agricultural University (http://202.194.139.32/blast/viroblast.php). The coding and flanking sequences of *Pinb2* genes were aligned and used to design primers specific to particular loci or variants. Specific primers were designed successfully in the light of variant- or loci-specific regions based on alignment (Figure S1, Table S7). Particularly, the conserved regions of the *Pinb2* variants allowed us to design the universal primer pair U. A pair of dCAPS primers was newly designed to identify allelic variation in *Pinb-B2v2-1* on chromosome 7B. Based on a special, single-nucleotide polymorphism of *Pinb-B2v2-1*, dCAPS Finder v2.0 was used to design dCAPS primer pair C for its use together with restriction enzyme *Bst*X I (Fermentas, Waltham, MA, USA) [74]. Using the new primer pairs described here, a pipeline for detecting known *Pinb2* genes and identifying new *Pinb2* variants at all three homoeologous loci will be possible (details in Results, Figure 4).

Genomic DNA samples from three wheat seedlings of each variety were extracted with the Cetyltrimethylammonium Ammonium Bromide (CTAB) method and used for *Pina*/*Pinb* and *Pinb2* genotyping [73]. PCR amplification was done in a 25 µL reaction volume containing 100 ng of genomic DNA, 0.2 µM primers, and 12.5 µL 2×Es Taq MasterMix (CWBIO, Beijing, China) using the the PCR program as follows: 95 °C for 3 min, 34 cycles of 95 °C for 30 s, 55–62 °C 30 s and 72 °C for 30 s, with a final extension of 5 min in a Bio-Rad-T100 thermal cycler. The PCR products were separated by 1.5% (*w*/*v*) agarose-gel electrophoresis and visualized under UV light after ethidium bromide staining. A sequencing service for the purified PCR products (provided by AuGCT company, Beijing, China) was used to validate the *Pina*/*Pinb* and *Pinb-2v* genotypes and to obtain the sequences of the newly identified alleles of *Pinb2* (Table 3).

Table 3. The information of new *Pinb2* alleles identified in this study.

New Designation	Previous *Pinb-2v* Designation	Species	Accession	NCBI Accession
TaPinb2-D1-v1-6	*Pinb-D2v1-6*	*Triticum aestivum*	Zhengmai 101, Wanke 06229, Jimai 107, Laoqimai	MN839440
AcPinb2-D1-v1-8	*Pinb-2v1-8*	*Aegilops cylindrical*	na	MN708354
AvPinb2-D1-v1-9	*Pinb-2v1-9*	*Aegilops vavilovii*	na	MN708355
AtPinb2-D1-v1-10	*Pinb-2v1-10*	*Aegilops triuncialis*	na	MN708356
AgPinb2-D1-v1-11	*Pinb-2v1-11*	*Aegilops geniculate*	na	MN708357
TuPinb2-A1-v4u-2	*Pinb-2v4u-2*	*Triticum urartu*	G1937 (PI 428230)	MN893165
TuPinb2-A1-v4u-3	*Pinb-2v4u-3*	*Triticum urartu*	G1906 (PI 428228)	MN893166

"na" = not applicable.

4.4. Synteny Analysis of the Genomic Segments Containing Pinb2 Genes

The gene order and annotation from several high-quality *Triticeae* species were used for syntenic analysis, including *T. aestivum* IWSGC RefSeq v1.0, the *T. urartu* genome, *Ae. tauschii* genome, and wild emmer wheat (*T. turgidum* spp. *dicoccoides*) genome [36–39]. High-confidence protein-coding genes flanking *Pina*/*Pinb* genes at 5DS, and those flanking *Pinb2* at 7AL, 7BL and 7DL (within ~2–3 Mb), were retrieved from the genomes. The orthologous gene pairs were determined by using reciprocal BLAST (10e-5, the score is greater than 70%, and the matching base is longer than 100 bp) to compare genes between any two of the genomes. Annotations of the genes and their abbreviated names shown in Table S3.

4.5. Quantitative Real-Time Reverse Transcription PCR (qRT-PCR)

Quantitative PCR primers were designed to target the gene-specific coding regions for *Pina*, *Pinb* and *Pinb2*, respectively, according to the sequence alignment by DNAMAN v6.0. *Actin* (TraesCS4B01G050600.1) was used as the internal reference gene for qPCR. Due to the high sequence identity within the *Pinb2* coding region, qRT-PCR primers for quantifying *Pinb2* were only able to detect the overall expression levels of all expressed copies at the three loci (Table S6). The total RNA was extracted from different tissues of the wheat cultivar Chinse Spring [75]. The first-strand cDNA was generated from 1 µg RNA using a FastKing RT kit and gDNase (Tiangen, China) and qRT-PCR was conducted using AceQ qPCR SYBR Green Master Mix (Vazyme, Nanjing, China). The data were evaluated between the three replicates by using the relative quantification method ($2^{-\Delta\Delta Ct}$).

4.6. Analyses of the Phylogeny and Protein Features for Wheat Seed Proteins

The predicted protein sequences from the bread wheat reference genome were used for phylogenetic analysis together with the predicted protein sequences of *Puroindoline* homologs in triticale and barley (namely, *Sina* and *Sinb* from hexaploid triticale, and *Hina*, *Hinb-1* and *Hinb-2* from barley, *Hordeum vulgare*) (Table S4). Domain analysis using profile hidden Markov Models (HMMER) was performed for the predicted protein sequences of *Pinb2* and other seed proteins to identify several protein domain signatures, for instance the Tryp_alpha_amyl domain (PF00234), HMW-GS domain (PF03157) and gliadin domain (PF13016) [76]. Annotations of wheat seed proteins were reported previously by Juhasz et al. [43] and avenin-like proteins (ALP) were annotated by Zhang et al. [51]. The presence of conserved cysteine residual patterns was used as a feature for the subfamily assignment of the prolamin superfamily [47,48]. The phylogenetic analysis followed a detailed protocol [77]. The protein sequences were aligned using MUltiple Sequence Comparison by Log-Expectation (MUSCLE), guided by an unweighted pair-group method with arithmetic means (UPGMA) tree. The phylogenetic tree of the wheat seed proteins was constructed using the maximum likelihood method provided in the MEGA7 software with a 500-time bootstrap using a Jones–Taylor–Thornton (JTT) model [78].

4.7. Statistical Analysis

The association of *Pinb* and *Pinb2* loci with kernel hardness phenotypes was determined by a *Chi*-square test of independence.

5. Conclusions

In conclusion, *Pinb2* genes consist of five copies rather than three homoeologous genes due to the additional tandem duplicated copies at *Pinb2-7A*. Based on the synteny and phylogeny analyses, *Pinb2* genes likely preserve the sequence features for interacting with gluten proteins through hydrophobic connections but lack the basis for determining kernel hardness, such as the TRD domain. These results are in line with the association analysis results: *Pinb2* genes do not exert major impacts on kernel hardness as *Pina*/*Pinb*. Leveraging the high-quality reference genome of bread wheat, we develop new *Pinb2* genotyping primers and demonstrate their application in identifying new *Pinb2* alleles and in facilitating association studies. The present study exemplifies an application of the high-quality *Triticeae* genomic resources, and the results implicate the areas for further study to unveil *Pinb2*'s function and its potential use in genetic engineering.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/21/4/1304/s1.

Author Contributions: G.H., G.Y., X.L. and Y.L. conceived and designed the experiments; X.L., F.S. and X.Y. performed the experiments; X.L., F.S., X.Y. and Y.L. analyzed the data; G.H. and G.Y. contributed reagents/materials/analysis tools. X.L., Y.L. drafted the manuscript. G.Y. and G.H. finalized the manuscript. funding acquisition, G.H. and G.Y. All authors have read and agreed to the published version of the manuscript.

Funding: The work was funded by National Genetically Modified New Varieties of Major Projects of China (2016ZX08010004-004), the National Natural Science Foundation of China (No. 31771418, 31570261) and Key Project of Hubei Province (2017AHB041).

Acknowledgments: The authors thank the anonymous editors and reviewers whose constructive comments were helpful in improving the quality of this work.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ATI	α-amylase inhibitor
ALP	avenin-like protein
BAC	bacterial artificial chromosome
BGGP	β-1-3- galactosyl- O-glycosyl-glycoprotein
CS	Chinese Spring
dCAPS	derived cleaved amplified polymorphic sequence
GSP	grain softness protein
Ha	Hardness
HD	hydropbic domain
HMW-GS	High-Molecular-Weight Glutenin Subunits
IWGSC	International Wheat Genome Sequencing Consortium
LTP	Lipid transfer proteins
ORF	open reading frame
PIN	puroindoline
TPM	Transcripts Per Million
TRD	tryptophan-rich domain
UPGMA	unweighted pair group method using arithmetic average

References

1. Martin, J.M.; Frohberg, R.C.; Morris, C.F.; Talbert, L.E.; Giroux, M.J. Milling and bread baking traits associated with puroindoline sequence type in hard red spring wheat. *Crop Sci.* **2001**, *41*, 228–234. [CrossRef]
2. Martin, J.M.; Meyer, F.D.; Morris, C.F.; Giroux, M.J. Pilot scale milling characteristics of transgenic isolines of a hard wheat over-expressing *Puroindolines*. *Crop Sci.* **2007**, *47*, 497–506. [CrossRef]
3. Li, Y.; Mao, X.; Wang, Q.; Zhang, J.R.; Li, X.Y.; Ma, F.Y.; Sun, F.S.; Chang, J.L.; Chen, M.J.; Wang, Y.S.; et al. Overexpression of *Puroindoline a* gene in transgenic durum wheat (*Triticum turgidum* ssp. *durum*) leads to a medium–hard kernel texture. *Mol. Breed.* **2014**, *33*, 545–554.
4. Li, X.; Li, Y.; Zhang, M.; Yu, X.; Hu, R.; Chang, J.; Yang, G.; Wang, Y.; He, G. Diversity of *Puroindoline* genes and their association with kernel hardness in Chinese wheat cultivars and landraces. *Mol. Breed.* **2019**, *39*, 61. [CrossRef]
5. Morris, C.F. *Puroindolines*: The molecular genetic basis of wheat grain hardness. *Plant Mol. Biol.* **2002**, *48*, 633–647. [CrossRef]
6. Bhave, M.; Morris, C.F. Molecular genetics of *Puroindolines* and related genes: Allelic diversity in wheat and other grasses. *Plant Mol. Biol.* **2008**, *66*, 205–219. [CrossRef]
7. Bhave, M.; Morris, C.F. Molecular genetics of *Puroindolines* and related genes: Regulation of expression, membrane binding properties and applications. *Plant Mol. Biol.* **2008**, *66*, 221–231. [CrossRef]
8. Beecher, B.; Bettge, A.; Smidansky, E.; Giroux, M. Expression of wild-type pinB sequence in transgenic wheat complements a hard phenotype. *Theor. Appl. Genet.* **2002**, *105*, 870–877. [CrossRef]
9. Giroux, M.J.; Morris, C.F. A glycine to serine change in puroindoline b is associated with wheat grain hardness and low levels of starch-surface friabilin. *Theor. Appl. Genet.* **1997**, *95*, 857–864. [CrossRef]
10. Hogg, A.C.; Sripo, T.; Beecher, B.; Martin, J.M.; Giroux, M.J. Wheat puroindolines interact to form friabilin and control wheat grain hardness. *Theor. Appl. Genet.* **2004**, *108*, 1089–1097. [CrossRef]
11. Krishnamurthy, K.; Giroux, M.J. Expression of wheat *puroindoline* genes in transgenic rice enhances grain softness. *Nat. Biotechnol.* **2001**, *19*, 162–166. [CrossRef] [PubMed]
12. Martin, J.M.; Meyer, F.D.; Smidansky, E.D.; Wanjugi, H.; Blechl, A.E.; Giroux, M.J. Complementation of the *pina* (null) allele with the wild type *Pina* sequence restores a soft phenotype in transgenic wheat. *Theor. Appl. Genet.* **2006**, *113*, 1563–1570. [CrossRef] [PubMed]
13. Zhang, J.; Martin, J.M.; Beecher, B.; Morris, C.F.; Hannah, L.C.; Giroux, M.J. Seed-specific expression of the wheat *puroindoline* genes improves maize wet milling yields. *Plant Biotechnol. J.* **2009**, *7*, 733–743. [CrossRef] [PubMed]
14. Wang, Q.; Li, Y.; Sun, F.; Li, X.; Wang, P.; Yang, G.; He, G. Expression of *Puroindoline a* in durum wheat affects milling and pasting properties. *Front. Plant Sci.* **2019**, *10*, 482. [CrossRef] [PubMed]

15. Wang, Q.; Li, Y.; Sun, F.; Li, X.; Wang, P.; Chang, J.; Wang, Y.; Yang, G.; He, G. Co-expression of high-molecular-weight glutenin subunit *1Ax1* and *Puroindoline a* (*Pina*) genes in transgenic durum wheat (*Triticum turgidum* ssp. *durum*) improves milling and pasting quality. *BMC Plant Biol.* **2019**, *19*, 126.
16. Rai, A.; Mahendru-Singh, A.; Raghunandan, K.; Kumar, T.; Sharma, P.; Ahlawat, A.; Singh, S.; Ganiewala, D.; Shukla, R.; Sivasamy, M. Marker-assisted transfer of *PinaD1a* gene to develop soft grain wheat cultivars. *3 Biotech.* **2019**, *9*, 183. [CrossRef]
17. Ma, X.; Xue, H.; Sun, J.; Sajjad, M.; Wang, J.; Yang, W.; Li, X.; Zhang, A.; Liu, D. Transformation of *Pinb-D1x* to soft wheat produces hard wheat kernel texture. *J. Cereal Sci.* **2020**, *91*, 102889. [CrossRef]
18. Ali, I.; Sardar, Z.; Rasheed, A.; Mahmood, T. Molecular characterization of the *puroindoline-a* and *b* alleles in synthetic hexaploid wheats and in silico functional and structural insights into *Pina-D1*. *J. Theor. Biol.* **2015**, *376*, 1–7. [CrossRef]
19. Chen, F.; He, Z.H.; Xia, X.C.; Xia, L.Q.; Zhang, X.Y.; Lillemo, M.; Morris, C.F. Molecular and biochemical characterization of *puroindoline a* and *b* alleles in Chinese landraces and historical cultivars. *Theor. Appl. Genet.* **2006**, *112*, 400–409. [CrossRef]
20. Kumar, R.; Arora, S.; Singh, K.; Garg, M. *Puroindoline* allelic diversity in Indian wheat germplasm and identification of new allelic variants. *Breed. Sci.* **2015**, *65*, 319–326. [CrossRef]
21. Ma, X.; Sajjad, M.; Wang, J.; Yang, W.; Sun, J.; Li, X.; Zhang, A.; Liu, D. Diversity, distribution of *Puroindoline* genes and their effect on kernel hardness in a diverse panel of Chinese wheat germplasm. *BMC Plant Biol.* **2017**, *17*, 158. [CrossRef]
22. Wang, J.; Sun, J.Z.; Liu, D.C.; Yang, W.L.; Wang, D.W.; Tong, Y.P.; Zhang, A.M. Analysis of *Pina* and *Pinb* alleles in the microcore collections of Chinese wheat germplasm by Ecotilling and identification of a novel Pinb allele. *J. Cereal Sci.* **2008**, *48*, 836–842. [CrossRef]
23. Ayala, M.; Guzmán, C.; Alvarez, J.; Peña, R. Characterization of genetic diversity of *puroindoline* genes in Mexican wheat landraces. *Euphytica* **2012**, *190*, 53–63. [CrossRef]
24. Klimushina, M.; Divashuk, M.; Mokhammed, T.; Semenov, O.; Karlov, G. Analysis of allelic state of genes responsible for baking properties in allocytoplasmic wheat hybrids. *Russ. J. Genet.* **2013**, *49*, 530–538. [CrossRef]
25. Morris, C.; Kiszona, A.; Peden, G. Registration of extra-hard kernel near-isogenic hexaploid wheat genetic stocks lacking *puroindoline* genes. *J. Plant Regist.* **2020**, *2*. [CrossRef]
26. Chen, F.; Zhang, F.Y.; Cheng, X.Y.; Craig, M.; Xu, H.X.; Dong, Z.D.; Zhan, K.H.; Cui, D.Q. Association of *Puroindoline b-B2* variants with grain traits, yield components and flag leaf size in bread wheat (*Triticum aestivum* L.) varieties of the Yellow and Huai Valleys of China. *J. Cereal Sci.* **2010**, *52*, 247–253. [CrossRef]
27. Wilkinson, M.; Wan, Y.; Tosi, P.; Leverington, M.; Snape, J.; Mirchell, A.C.R.; Shewry, P.R. Identification and genetic mapping of variant forms of *puroindoline b* expressed in developing wheat grain. *J. Cereal Sci.* **2008**, *48*, 722–728. [CrossRef]
28. Chen, F.; Beecher, B.S.; Morris, C.F. Physical mapping and a new variant of *Puroindoline b-2* genes in wheat. *Theor. Appl. Genet.* **2010**, *120*, 745–751. [CrossRef]
29. Chen, F.; Xu, H.X.; Zhang, F.Y.; Xia, X.C.; He, Z.H.; Wang, D.W.; Dong, Z.D.; Zhan, K.H.; Cheng, X.Y.; Cui, D.Q. Physical mapping of *puroindoline b-2* genes and molecular characterization of a novel variant in durum wheat (*Triticum turgidum* L.). *Mol. Breed.* **2011**, *28*, 153–161. [CrossRef]
30. Geng, H.; Beecher, B.S.; He, Z.H.; Morris, C.F. Physical Mapping of *Puroindoline b-2* Genes in Wheat using 'Chinese Spring' Chromosome Group 7 Deletion Lines. *Crop Sci.* **2012**, *52*, 2674–2678. [CrossRef]
31. Ramalingam, A.; Palombo, E.A.; Bhave, M. The *Pinb-2* genes in wheat comprise a multigene family with great sequence diversity and important variants. *J. Cereal Sci.* **2012**, *56*, 171–180. [CrossRef]
32. Mohler, V.; Schmolke, M.; Paladey, E.; Seling, S.; Hartl, L. Association analysis of *Puroindoline-D1* and *Puroindoline b-2* loci with 13 quality traits in European winter wheat (*Triticum aestivum* L.). *J. Cereal Sci.* **2012**, *56*, 623–628. [CrossRef]
33. Geng, H.; Beecher, B.S.; He, Z.; Kiszonas, A.M.; Morris, C.F. Prevalence of *Puroindoline D1* and *Puroindoline b-2* variants in U.S. Pacific Northwest wheat breeding germplasm pools, and their association with kernel texture. *Theor. Appl. Genet.* **2012**, *124*, 1259–1269. [CrossRef]
34. Chen, F.; Zhang, F.; Li, H.; Morris, C.F.; Cao, Y.; Shang, X.; Cui, D. Allelic variation and distribution independence of *Puroindoline b-B2* variants and their association with grain texture in wheat. *Mol. Breed.* **2013**, *32*, 399–409. [CrossRef]

35. Giroux, M.J.; Kim, K.H.; Hogg, A.C.; Martin, J.M.; Beecher, B. The *Puroindoline b-2* variants are expressed at low levels relative to the *Puroindoline D1* genes in wheat seeds. *Crop Sci.* **2013**, *53*, 833–841. [CrossRef]
36. International Wheat Genome Sequencing Consortium (IWGSC). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **2018**, *361*, eaar7191. [CrossRef]
37. Avni, R.; Nave, M.; Barad, O.; Baruch, K.; Twardziok, S.O.; Gundlach, H.; Hale, I.; Mascher, M.; Spannagl, M.; Wiebe, K.; et al. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **2017**, *357*, 93–97. [CrossRef]
38. Ling, H.; Ma, B.; Shi, X.; Liu, H.; Dong, L.; Sun, H.; Cao, Y.; Gao, Q.; Zheng, S.; Li, Y.; et al. Genome sequence of the progenitor of wheat A subgenome *Triticum Urartu*. *Nature* **2018**, *557*, 424–428. [CrossRef]
39. Luo, M.C.; Gu, Y.; Puiu, D.; Wang, H.; Twardziok, S.O.; Deal, K.R.; Huo, N.; Zhu, T.; Wang, L.; Wang, Y.; et al. Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature* **2017**, *551*, 498–502. [CrossRef]
40. Zhao, G.; Zou, C.; Li, K.; Wang, K.; Li, T.; Gao, L.; Zhang, X.; Wang, H.; Yang, Z.; Liu, X.; et al. The *Aegilops tauschii* genome reveals multiple impacts of transposons. *Nat. Plants* **2017**, *3*, 946–955. [CrossRef]
41. Dong, J.; Feng, Y.; Kumar, D.; Zhang, W.; Zhu, T.; Luo, M.C.; Messing, J. Analysis of tandem gene copies in maize chromosomal regions reconstructed from long sequence reads. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 7949–7956. [CrossRef]
42. Li, C.; Xiang, X.; Huang, Y.; Zhou, Y.; An, D.; Dong, J.; Zhao, C.; Liu, H.; Li, Y.; Wang, Q.; et al. Long-read sequencing reveals genomic structural variations that underlie creation of quality protein maize. *Nat. Commun.* **2020**, *11*, 17. [CrossRef]
43. Zhang, J.; Zhang, X.; Tang, H.; Zhang, Q.; Hua, X.; Ma, X.; Zhu, F.; Jones, T.; Zhu, X.; Bowers, J.; et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* **2018**, *50*, 1565–1573. [CrossRef]
44. Shi, J.; Ma, X.; Zhang, J.; Zhou, Y.; Liu, M.; Huang, L.; Sun, S.; Zhang, X.; Gao, X.; Zhan, W.; et al. Chromosome conformation capture resolved near complete genome assembly of broomcorn millet. *Nat. Commun.* **2019**, *10*, 464. [CrossRef]
45. Chantret, N.; Salse, J.; Sabot, F.; Rahman, S.; Bellec, A.; Laubin, B.; Dubois, I.; Dossat, C.; Sourdile, P.; Philippe, J.; et al. Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *Plant Cell* **2005**, *17*, 1033–1045. [CrossRef]
46. Li, W.; Huang, L.; Gill, B.S. Recurrent deletions of *Puroindoline* genes at the grain hardness locus in four independent lineages of polyploid wheat. *Plant Physiol.* **2008**, *146*, 200–212. [CrossRef]
47. Charles, M.; Tang, H.; Belcram, H.; Paterson, A.; Gornick, P.; Chalhoub, B. Sixty million years in evolution of soft grain trait in grasses: Emergence of the softness locus in the common ancestor of pooideae and ehrhartoideae, after their divergence from panicoideae. *Mol. Biol. Evol.* **2009**, *26*, 1651–1661. [CrossRef]
48. Juhasz, A.; Belova, T.; Florides, C.G.; Maulis, C.; Fischer, I.; Gell, G.; Birinyi, Z.; Ong, J.; Keeble-Gagnere, G.; Maharajan, A.; et al. Genome mapping of seed-borne allergens and immunoresponsive proteins in wheat. *Sci. Adv.* **2018**, *4*, eaar8602. [CrossRef]
49. Ma, F.; Li, M.; Yu, L.; Li, Y.; Liu, Y.; Li, T.; Liu, W.; Wang, H.; Zheng, Q.; Li, K.; et al. Transformation of common wheat (*Triticum aestivum* L.) with *avenin-like b* gene improves flour mixing properties. *Mol. Breed.* **2013**, *32*, 853–865. [CrossRef]
50. Ma, F.; Li, M.; Li, T.; Liu, W.; Liu, Y.; Li, Y.; Hu, W.; Zheng, Q.; Wang, Y.; He, G. Overexpression of avenin-like b proteins in bread wheat (*Triticum aestivum* L.) improves dough mixing properties by their incorporation into glutenin polymers. *PLoS ONE* **2013**, *8*, e66758. [CrossRef]
51. Zhang, Y.; Hu, X.; Islam, S.; She, M.; Peng, Y.; Yu, Z.; Wylie, S.; Juhasz, A.; Dowla, M.; Yang, R.; et al. New insights into the evolution of wheat avenin-like proteins in wild emmer wheat (*Triticum dicoccoides*). *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 13312–13317. [CrossRef]
52. Wang, Y.; Li, M.; Guan, Y.; Li, L.; Sun, F.; Han, J.; Chang, J.; Chen, M.; Yang, G.; He, G. Effects of an additional cysteine residue of avenin-like b protein by site-directed mutagenesis on dough properties in wheat (*Triticum aestivum* L.). *J. Agric. Food Chem.* **2019**, *67*, 8557–8572. [CrossRef]
53. Beecher, B.; Bowman, J.; Martin, J.M.; Bettge, A.D.; Morris, C.F.; Blake, T.K.; Giroux, M.J. Hordoindolines are associated with a major endosperm-texture QTL in Barley (*Hordeum vulgare*). *Genome* **2002**, *45*, 584–591. [CrossRef]

54. Feiz, L.; Beecher, B.S.; Martin, J.M.; Giroux, M.J. In planta mutagenesis determines the functional regions of the wheat puroindoline proteins. *Genetics* **2009**, *183*, 853–860. [CrossRef]
55. Feiz, L.; Martin, J.M.; Giroux, M.J. Creation and functional analysis of new *Puroindoline* alleles in *Triticum aestivum*. *Theor. Appl. Genet.* **2009**, *118*, 247–257. [CrossRef]
56. Takahashi, A.; Ikeda, T.M.; Takayama, T.; Yanagisawa, T. A barley Hordoindoline mutation resulted in an increase in grain hardness. *Theor. Appl. Genet.* **2010**, *120*, 519–526. [CrossRef]
57. Gasparis, S.; Orczyk, W.; Nadolska-Orczyk, A. *Sina* and *Sinb* genes in triticale do not determine grain hardness contrary to their orthologs *Pina* and *Pinb* in wheat. *BMC Plant Biol.* **2013**, *13*, 190. [CrossRef]
58. Shabrangy, A.; Roustan, V.; Reipert, S.; Weidinger, M.; Roustan, P.J.; Stoger, E.; Weckwerth, W.; Ibl, V. Using RT-qPCR, proteomics, and microscopy to unravel the spatio-temporal expression and subcellular localization of hordoindolines across development in barley endosperm. *Front. Plant Sci.* **2018**, *9*, 775. [CrossRef]
59. Alfred, R.L.; Palombo, E.A.; Panozzo, J.F.; Bhave, M. The cooperative interaction of puroindolines in wheat grain texture may involve the hydrophobic domain. *J. Cereal Sci.* **2014**, *60*, 323–330. [CrossRef]
60. Geneix, N.; Dalgalarrondo, M.; Bakan, B.; Rolland-Sabate, A.; Elmorjani, K.; Marion, D. A single amino acid substitution in puroindoline b impacts its self-assembly and the formation of heteromeric assemblies with puroindoline a. *J. Cereal Sci.* **2015**, *64*, 116–125. [CrossRef]
61. Borrill, P.; Ramirez-Gonzalez, R.H.; Uauy, C. expVIP: A customizable RNA-seq data analysis and visualization platform. *Plant Physiol.* **2016**, *170*, 2172–2186. [CrossRef]
62. Ramirez-Gonzalez, R.H.; Borrill, P.; Lang, D.; Harrington, S.A.; Brinton, J.; Venturini, L.; Davey, M.; Jacobs, J.; van Ex, F.; Pasha, A.; et al. The transcriptional landscape of polyploid wheat. *Science* **2018**, *361*, eaar6089. [CrossRef]
63. Yu, X.; Wang, T.; Zhu, M.; Zhang, L.; Zhang, F.; Jing, E.; Ren, Y.; Wang, Z.; Xin, Z.; Lin, T. Transcriptome and physiological analyses for revealing genes involved in wheat response to endoplasmic reticulum stress. *BMC Plant Biol.* **2019**, *19*, 193. [CrossRef]
64. Geng, H.; Beecher, B.; Pumphrey, M.; He, Z.; Morris, C.F. Segregation analysis indicates that *Puroindoline b-2* variants 2 and 3 are allelic in *Triticum aestivum* and that a revision to *Puroindoline b-2* gene symbolization is indicated. *J. Cereal Sci.* **2013**, *57*, 61–66. [CrossRef]
65. Chen, F.; He, Z.H.; Xia, X.C.; Lillemo, M.; Morris, C.F. A new *puroindoline b* mutation presented in Chinese winter wheat cultivar Jingdong 11. *J. Cereal Sci.* **2005**, *42*, 267–269. [CrossRef]
66. Lillemo, M.; Chen, F.; Xia, X.C.; William, M.; Peña, R.J.; Trethowan, R.; He, Z.H. *Puroindoline* grain hardness alleles in CIMMYT bread wheat germplasm. *J. Cereal Sci.* **2006**, *44*, 86–92. [CrossRef]
67. Morris, C.F.; Lillemo, M.; Simeone, M.C.; Giroux, M.J.; Babb, S.L.; Kidwell, K.K. Prevalence of puroindoline grain hardness genotypes among historically significant North American spring and winter wheats. *Crop Sci.* **2001**, *41*, 218–228. [CrossRef]
68. Sanders, M.R.; Clifton, L.A.; Frazier, R.A.; Green, R.J. Role of lipid composition on the interaction between a tryptophan-rich protein and model bacterial membranes. *Langmuir* **2016**, *32*, 2050–2057. [CrossRef]
69. Quayson, E.T.; Marti, A.; Morris, C.F.; Marengo, M.; Bonomi, F.; Seetharaman, K.; Iametti, S. Structural consequences of the interaction of puroindolines with gluten proteins. *Food Chem.* **2018**, *253*, 255–261. [CrossRef]
70. Geneix, N.; Dlgalarrondo, M.; Tassy, C.; Nadaud, I.; Barret, P.; Bakan, B.; Elmorjani, K.; Marion, D. Relationships between puroindoline-prolamin interactions and wheat grain hardness. *bioRxv* **2019**. [CrossRef]
71. Schilling, S.; Kennedy, A.; Pan, S.; Jermiin, L.S.; Melzer, R. Genome-wide analysis of MIKC-type *MADS-box* genes in wheat: Pervasive duplications, functional conservation and putative neofunctionalization. *New Phytol.* **2020**, *225*, 511–529. [CrossRef]
72. Qin, H.; Ma, D.; Huang, X.; Zhang, J.; Sun, W.; Hou, G.; Wang, C.; Guo, T. Accumulation of glycolipids in wheat grain and their role in hardness during grain development. *Crop J.* **2019**, *7*, 19–29. [CrossRef]
73. Li, Y.; Wang, Q.; Li, Y.; Xiao, X.; Sun, S.; Wang, C.; Hu, W.; Feng, J.; Chang, L.; Chen, J.; et al. Coexpression of the high molecular weight glutenin subunit *1Ax1* and *puroindoline* improves dough mixing properties in durum wheat (*Triticum turgidum* L. ssp. *durum*). *PLoS ONE* **2012**, *7*, e50057. [CrossRef]
74. Neff, M.M.; Turk, E.; Kalishman, M. Web-based primer design for single nucleotide polymorphism analysis. *Trends Genet.* **2002**, *18*, 613–615. [CrossRef]
75. Sun, F.; Liu, X.; Wei, Q.; Liu, J.; Yang, T.; Jia, L.; Wang, Y.; Yang, G.; He, G. Functional characterization of *TaFUSCA3*, a B3-superfamily transcription factor gene in the wheat. *Front. Plant Sci.* **2017**, *8*, 1133. [CrossRef]

76. Potter, S.C.; Luciani, A.; Eddy, Y.; Park, S.R.; Lopez, R.; Finn, R.D. HMMER web server: 2018 update. *Nucleic Acids Res.* **2018**, Web Server Issue 46. W200–W204. [CrossRef]
77. Hall, B.G. Building phylogenetic trees from molecular data with MEGA. *Mol. Biol. Evol.* **2013**, *30*, 1229–1235. [CrossRef]
78. Kumar, S.; Stechelr, G.; Tamura, K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **2016**, *33*, 1870–1874. [CrossRef]

International Journal of
Molecular Sciences

MDPI

Article

Genome-Wide Characterization and Expression Profiling of GASA Genes during Different Stages of Seed Development in Grapevine (*Vitis vinifera* L.) Predict Their Involvement in Seed Development

Bilal Ahmad [1,2,†], Jin Yao [1,2,†], Songlin Zhang [1,2], Xingmei Li [1,2], Xiuming Zhang [1,2], Vivek Yadav [2] and Xiping Wang [1,2,*]

[1] State Key Laboratory of Crop Stress Biology in Arid Areas, College of Horticulture, Northwest A&F University, Yangling 712100, China; bajwa1999@nwafu.edu.cn (B.A.); jin.yao@nwafu.edu.cn (J.Y.); zhangsonglin@nwafu.edu.cn (S.Z.); 18838933960@163.com (X.L.); zhangxiuming00@126.com (X.Z.)

[2] Key Laboratory of Horticultural Plant Biology and Germplasm Innovation in Northwest China, Northwest A&F University, Ministry of Agriculture, Yangling 712100, China; vivekyadav@nwafu.edu.cn

* Correspondence: wangxiping@nwsuaf.edu.cn; Tel.: +86-29-8708-2129

† These authors contributed equally to this work.

Received: 30 December 2019; Accepted: 1 February 2020; Published: 6 February 2020

Abstract: Members of the plant-specific GASA (gibberellic acid-stimulated Arabidopsis) gene family have multiple potential roles in plant growth and development, particularly in flower induction and seed development. However, limited information is available about the functions of these genes in fruit plants, particularly in grapes. We identified 14 GASA genes in grapevine (*Vitis vinifera* L.) and performed comprehensive bioinformatics and expression analyses. In the bioinformatics analysis, the locations of genes on chromosomes, physiochemical properties of proteins, protein structure, and subcellular positions were described. We evaluated GASA proteins in terms of domain structure, exon-intron distribution, motif arrangements, promoter analysis, phylogenetic, and evolutionary history. According to the results, the GASA domain is conserved in all proteins and the proteins are divided into three well-conserved subgroups. Synteny analysis proposed that segmental and tandem duplication have played a role in the expansion of the GASA gene family in grapes, and duplicated gene pairs have negative selection pressure. Most of the proteins were predicted to be in the extracellular region, chloroplasts, and the vacuole. In silico promoter analysis suggested that the GASA genes may influence different hormone signaling pathways and stress-related mechanisms. Additionally, we performed a comparison of the expression between seedless (Thompson seedless) and seeded (Red globe) cultivars in different plant parts, including the ovule during different stages of development. Furthermore, some genes were differentially expressed in different tissues, signifying their role in grapevine growth and development. Several genes (*VvGASA2* and *7*) showed different expression levels in later phases of seed development in Red globe and Thompson seedless, suggesting their involvement in seed development. Our study presents the first genome-wide identification and expression profiling of grapevine GASA genes and provides the basis for functional characterization of GASA genes in grapes. We surmise that this information may provide new potential resources for the molecular breeding of grapes.

Keywords: bioinformatics; grapevine; ovule abortion; *VvGAST*; GASR; *Cis*-elements

1. Introduction

Snakins are plant antimicrobial peptides (AMPs) of the GASA (gibberellic acid-stimulated Arabidopsis) gene family. These peptides have varied functions in response to various biotic and abiotic stresses via hormonal crosstalk [1] Snakin/GASA/ GAST (GA-stimulated transcripts)/GASR (GA-stimulated regulator) is a cysteine-rich low molecular weight peptide and part of the gibberellin-regulated family [2]. Mostly, hormone-regulated gene families affect various physiological processes and plant development. Various developmental roles have been speculated for GASA genes such as lateral root initiation and development, leaf expansion, flower induction, fruit cell size regulation, seed development and germination in many monocot and dicot plants [3,4]. Apart from these, most of the GASA genes are involved in hormonal (gibberellic acid, abscisic acid, and naphthalene acetic acid) signaling pathways [5,6]. For example, in rice, *OsGSR1* (a member of GASA gene family) influences Brassinosteroid signaling by interacting with *DIM*/*DWF1* [7].

The GASA gene family is highly specific to plants; the name was assigned according to the first identified member GAST-1 from tomatoes [2]. GASA genes encode low molecular weight proteins (80–270 amino acids) and have three different domains: (1) a N-terminal signal peptide of 18–29 amino acids; (2) a highly variable region (7–31 amino acids) displaying a difference between family members both in amino acid composition and sequence length; and (3) a C-terminal GASA domain of 60 amino acids with 12 conserved cysteine residues that contribute to the biochemical stability of the molecule [3,8]. In Arabidopsis, AtGASA2/AtGASA23, AtGASA5, and AtGASA14 are involved in ABA signaling [9]. Some GASA members may have opposite functions, e.g., AtGASA5 inhibits flowering while AtGASA4 promotes flowering [3,10,11]. Furthermore, GASA family members also play a role in disease resistance; for example, in rubber plants, GASA genes (HbGASA) were up-regulated upon encounter with the fungal pathogen *Colletotrichum gloeosporioides*. The HbGASA gene induced the production of reactive oxygen, signifying its role in plant innate immunity [12]. In *Solanum tuberosum* subsp. tuberosum cv. Kennebec, the expression of snakin1, 2, and 3 was found to be affected by the inoculation of bacterial fungal pathogens [13].

In the recent past, several studies focused on the functional characterization of these low molecular weight peptide-proteins in different plant species such as *Arabidopsis*, tomato, rice, potato, maize, wheat, apple, soybean, rubber plant, gerbera, strawberry, French bean, beechnut, pepper, and petunia [5,12–18]. Different studies found that GASA genes have potential roles in flower induction, seed size, seed development, and fruit size regulation, e.g., in Arabidopsis, overexpression of GASA4 positively affected seed size, seed weight, and seed yield [11]. Likewise, another member of this gene family, TaGASR7, has been found to be associated with grain length in wheat [19,20]. Similarly, application of different growth regulators (i.e., GA3, 6-BA, and sugar) in apple showed that GASA genes may have a role in flower induction [17]. While working with rice, Muhammad et al. [21] found that GASA genes increased grain size and length. Little information is available about the functions of GASA genes in fruit plants; however, for grapevine GASA genes; this will be the first study to provide such information.

Grapevine (*Vitis vinifera* L.) is among the top fruits crops grown all over the world, with an annual production of 7.9 million ha globally [22]. The cultivation is important due to its multipurpose use including table grapes, juice, raisins, and wine production. Grapes are supposed as an ideal model plant to study and understand the berry development phenomenon in perennial fruit crops. Recently, the demand for seedless grapes is also increasing [23,24]. In recent years, scientists have reported genes such as *VvYABBY4* and *VvHB58* as having a role in grapevine fruit and seed development [25,26]. However, the key genes mediating this process still needs to be explored, such as, for example, GASA genes. The important role of GASA genes as a regulator of different stages of plant growth, especially in flower induction, seed size, and seed weight in other fruits, justifies a detailed bioinformatics and expression profiling of this gene family in grapes. In this experiment, we performed a detailed bioinformatics study of the GASA gene family in grapes, including chromosomal locations and gene structure, sequence homology, evolutionary history, synteny analysis, cis-acting element analysis, in silico analysis of protein structure, and subcellular localization. We also investigated the expression of GASA genes during different phases of seed development as well as in different tissues in seedless and seeded grape cultivars. The findings of this experiment will provide foundations for further detailed studies of GASA genes in grapes as well as in other fruit plants.

2. Results

2.1. Genome-Wide Identification and Protein Features of GASA Genes in Grapevine

A total of 14 putative GASA genes were identified in the grapevine genome. These genes were named according to their locations on the chromosomes. Complete information about grapevine GASA genes is presented in Table 1.

Table 1. Detailed information of grapevine GASA (gibberellic acid-stimulated Arabidopsis) genes.

Gene Locus ID	Gene ID	Accession No.	Chromosome No.	Start Site	End Site	CDS (bp)	ORF (aa)
GSVIVT01020178001	*VvGASA1*	CBI32100	1+	9381743	9382568	327	108
GSVIVT01037887001	*VvGASA2*	CBI22497	3-	6715491	6716068	267	88
GSVIVT01000168001	*VvGASA3*	CBI33733	7+	15821751	15822734	204	67
GSVIVT01033563001	*VvGASA4*	CBI30071	8-	19734994	19736181	321	106
GSVIVT01032528001	*VvGASA5*	CBI34969	14+	28107504	28112390	897	298
GSVIVT01011412001	*VvGASA6*	CBI22214	14-	29446051	29447932	321	106
GSVIVT01008003001	*VvGASA7*	CBI15224	17+	6769396	6770752	315	104
GSVIVT01007817001	*VvGASA8*	CBI15083	17+	8741561	8742409	336	111
GSVIVT01009384001	*VvGASA9*	CBI19434	18+	7913344	7915123	267	88
GSVIVT01009902001	*VvGASA10*	CBI19861	18-	12275798	12276616	297	98
GSVIVT01034477001	*VvGASA11*	CBI18167	18+	20718815	20720279	339	112
GSVIVT01034476001	*VvGASA12*	CBI18166	18-	20720304	20720676	225	74
GSVIVT01003387001	*VvGASA13*	CBI25689	Un-	9775242	9775609	195	64
GSVIVT01003388001	*VvGASA14*	CBI25690	Un-	9791751	9792551	321	106

CDS: coding sequence; Chr: chromosome; ORF: open reading frame; Un: unknown chromosome.

The protein sequence of VvGASA genes varied from 64 (*VvGASA13*) to 298 (*VvGASA5*) amino acids with a MW (molecular weight) of 7.28 to 31.96 kDa. The average length of grapevine GASA proteins was 109 aa, while the average MW was 12.06 kDa. Apart from these, the isoelectric point (PI) ranged from 8.50 (*VvGASA13*) to 9.64 (*VvGASA5*), while for most of the proteins (72%), instability index values were more than 40. According to the Grand average of hydropathicity (GRAVY), the GASA proteins are hydrophilic except for *VvGASA8* and *VvGASA14*. As far as the amino acid content of proteins was concerned, cysteine, lysine, and leucine were predominant amino residues, whereas the aliphatic index ranged from 33.43 to 81.79. Detailed information about protein characteristics can be seen in Table 2.

Prediction of the subcellular positions of proteins can give important hints about their roles. From in silico analysis, the subcellular locations and structures of the proteins were determined. Most of the grapevine GASA genes were found in the apoplast (cell wall), vacuole, chloroplast, and cytoplasm (Table 2). All proteins of GASA genes have flexible structure due to the presence of coils. Members of the group 1 GASA gene family have more coils compared to other groups as shown in Figure 1. All proteins have at least two large α helices while β sheets are not common. Apart from two large α helices, group 1 proteins and GASA12 also have two small α helices.

Table 2. Amino acid composition and physiochemical characteristics of GASA proteins.

Gene	MW	PI	Major Amino Acid%	Instability Index	Aliphatic Index	GRAVY	Localization Predicted
GASA1	11.96	8.61	C(11.9), L(8.3), R(7.3)	36.64	67.06	−0.172	extr., vacu.
GASA2	9.71	9.02	C(14.8), K(12.5), L(11.4)	38.49	58.75	−0.281	chlo, nucl., extr
GASA3	7.28	8.87	C(17.9), K(13.4), G(10.4)	41.98	33.43	−0.515	chlo., nucl., cyto., extr.
GASA4	11.85	9.20	C(11.2), P(11.2), T(9.3)	51.40	45.61	−0.421	extr., chlo., nucl.
GASA5	31.96	9.64	P(23.7), V(9.7), K(8.4)	67.94	78.53	−0.241	cyto., ER
GASA6	11.79	9.30	C(11.3), K(11.3), G(9.4)	35.75	53.40	−0.289	extr., vacu., chlo.
GASA7	11.62	9.22	C(11.4), K(11.4), P (8.6)	38.76	57.62	−0.233	extr., vacu.
GASA8	12.27	8.66	C(10.7), L(9.8), G(8.9)	45.33	80.80	0.046	extr., chlo., vacu.
GASA9	9.71	8.96	C(13.5), K(13.5), S(10.1)	44.76	54.83	−0.206	extr., chlo., vacu.
GASA10	10.35	8.50	C(12.1), A(10.1), S(10.1)	47.05	63.23	−0.143	extr., vacu.
GASA11	12.62	9.52	S(11.5), C(10.6), K(10.6)	50.04	68.23	−0.344	extr.
GASA12	8.36	9.00	C(16), K(10.7), A(10.7)	42.75	49.47	−0.417	mito., chlo., cyto.
GASA13	7.42	8.50	C(17.2), K(10.9), Y(9.4)	42.31	41.09	−0.492	nucl., cyto., mito.
GASA14	11.96	8.80	C(12.3), L(12.3), K(11.3)	45.71	81.79	0.103	chlo., nucl., extr.

MW: molecular weight (kDa), pI: isoelectric point, GRAVY: grand average of hydropathicity,A: Ala, R: Arg, C: Cys, G: Gly, L: Leu, K: Lys, P: Pro, S: Ser, T: Thr, Y: Tyr, Extra: extracellular, Vacu: vacuoles, Chlo: chloroplast, Cyto: cytoplasm, Mito: mitochondria, Nucl: nucleus, Plas: plastids, and ER endoplasmic reticulum.

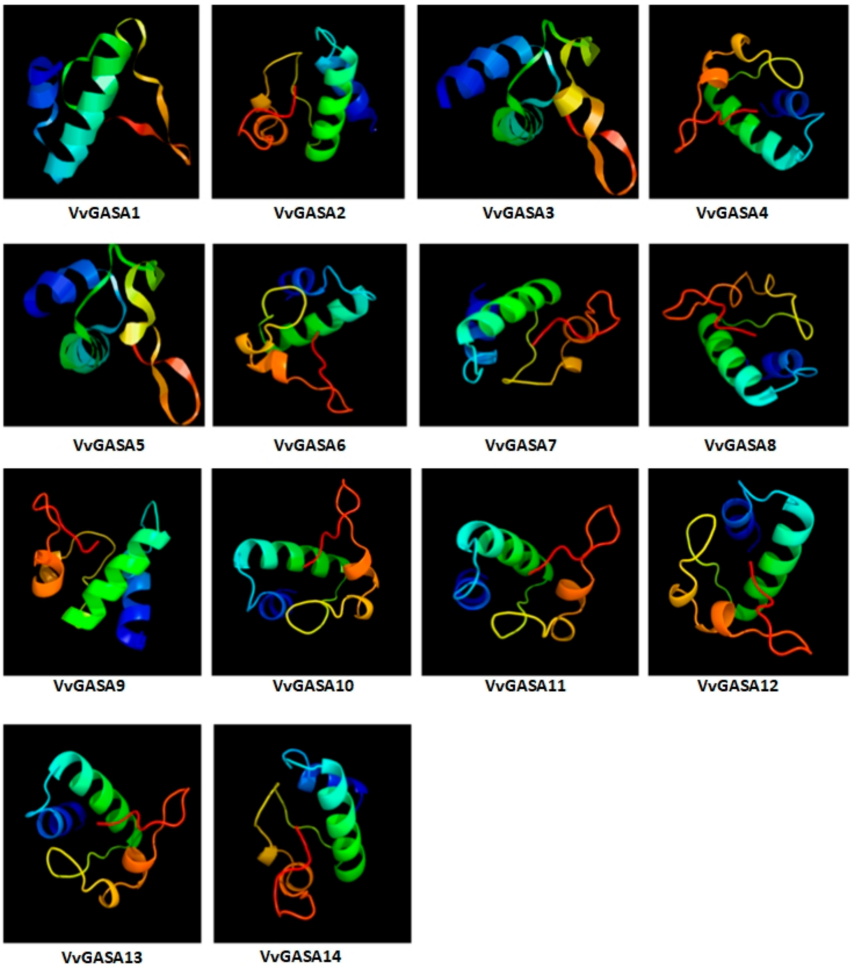


Figure 1. Predicted 3-D structures of GASA proteins.

2.2. Phylogenetic Analysis of GASA Genes from Grape, Apple and Arabidopsis

To describe the phylogeny and to assist in the classification of the GASA gene family in grapevine, a phylogenetic tree was constructed among grapes, apple, and *Arabidopsis*-aligned GASA protein sequences. The analysis included 55 GASA genes comprising 14, 26, and 15 from grapes, apple, and *Arabidopsis*, respectively. As shown in Figure 2, the genes were divided into three groups named G1, G2, and G3. In the distribution of VvGASA genes into different groups, the previous trend was noted, i.e., G2 contained the least (three) VvGASA genes, while G3 contained the most (six) genes. Grapes and *Arabidopsis* have the same number of genes in G1 and G3, while G2 has a difference of one gene. The predicted length of grapevine proteins in the G1, G2, and G3 groups ranges from 64–106, 67–88, and 74–298 amino acids, respectively.

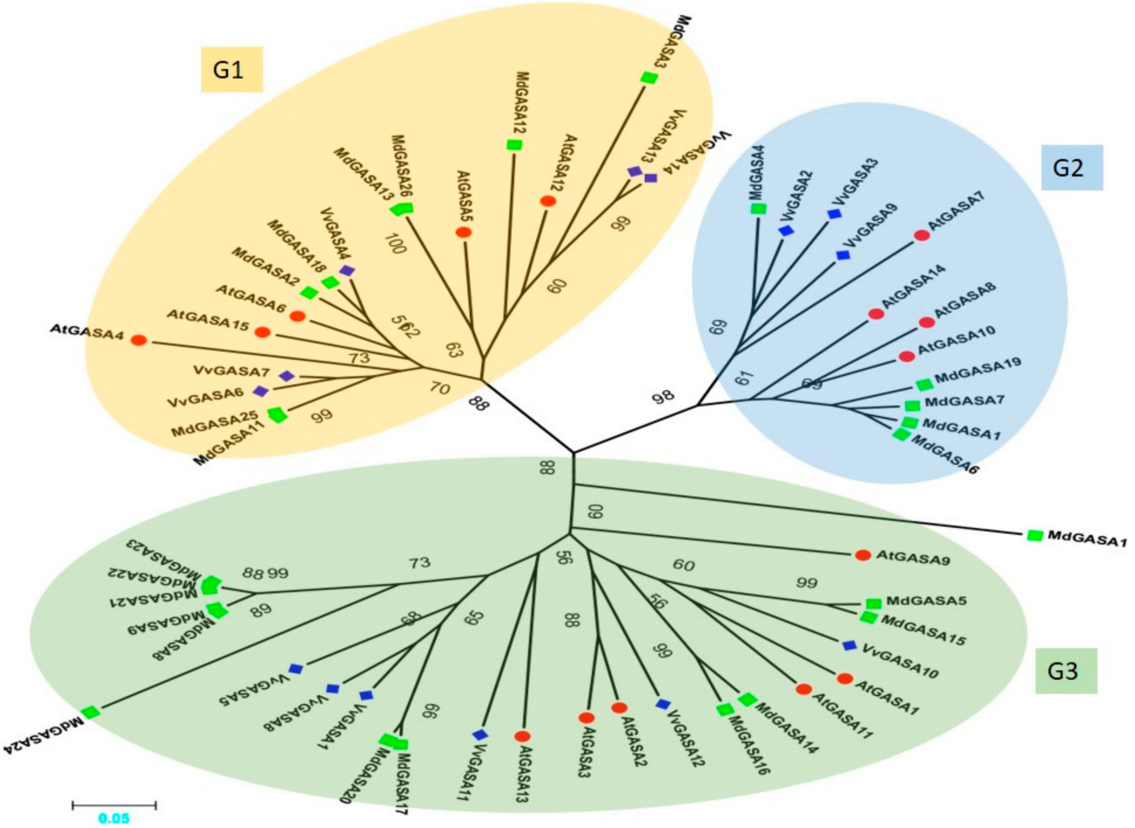


Figure 2. Phylogenetic tree of GASA genes of *Vitis vinifera*, *Malus domestica*, and *Arabidopsis thaliana*. Blue-colored diamonds represent grapevine protein, green-colored squares represent apple proteins, and red-colored circles represent *Arabidopsis* proteins. Different colored oval shapes indicate different groups. Numbers near the tree branches indicate bootstrap values (BS) and BS values less than 50 are not shown.

2.3. Analysis of Conserved Motifs, Domain Architecture, and Gene Structure

To further explore the phylogeny of grapevine GASA genes, an unrooted tree was constructed between VvGASA genes (Figure 3A). In concordance with the phylogenetic tree including the *Arabidopsis*, grapes, and apple GASA genes, this analysis also supported the classification of GASA genes into three groups. The exon–intron structural analysis of VvGASA genes was performed by the Gene Structure Display Server program to gain some perceptible information about the paralogous genes. Each member of G2 contained two exons (Figure 3C), showing that this group is structurally more conserved as compared to the others. In G1, three members (*GASA4*, *6*, and *7*) have four exons, while the remaining two have two exons per gene. Meanwhile, in G3, three members out of six have the same (four) exon number, and others have a different number of exons, showing that G3 is less conserved. The genes with similar exon numbers, positions, and lengths were closely related paralogous gene pairs. Moreover, the similarity index of protein sequences varied in each clade, with 59.4%, 57.4%, and 20.23% in G1, G2, and G3 groups, respectively. Furthermore, the similarity index of all grapevine GASA gene was 16.87%. This suggests that not only the exon number but also the protein similarity index is conserved with in the same clade. The motif distribution pattern was determined using an online server (MEME). GASA protein sequences have variations in motif length and number but have similar motif distribution patterns within the same group (Figure 3B). The highly conserved motifs 1 and 2 (gibberellin regulated protein (GASA domain); IPR003854) were detected in all fourteen genes of the GASA family, whereas, except for two members (GASA8 and 13), all members have motif 3.

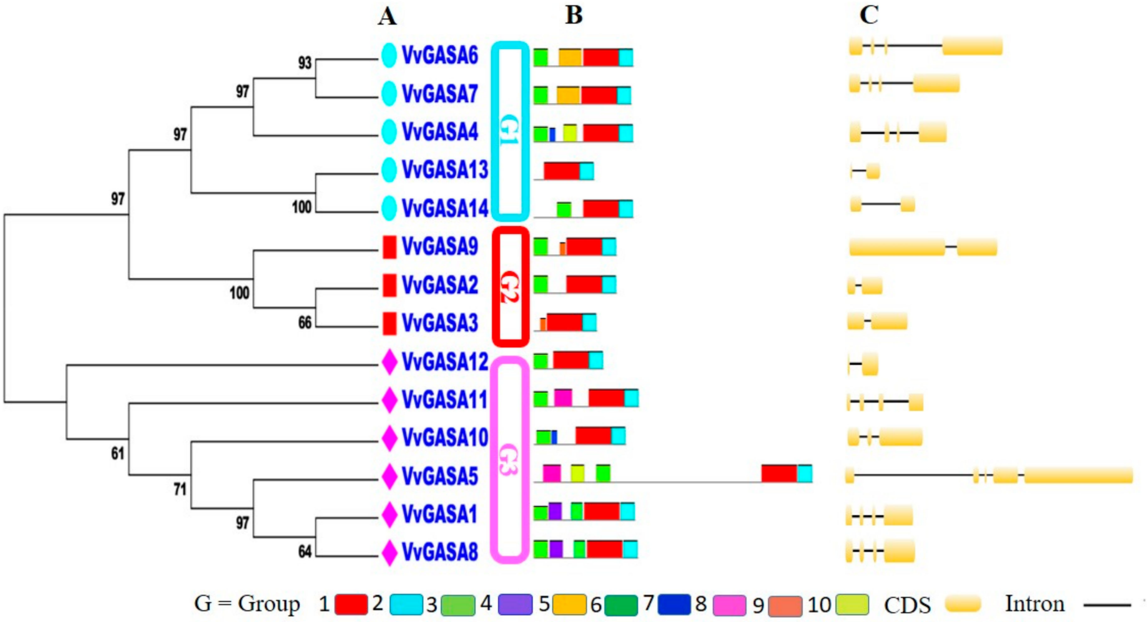


Figure 3. Analysis of grapevine GASA genes. (**A**) Phylogenetic tree of grapevine GASA genes. Different boxes are colored to indicating different groups. Numbers near the tree branches indicate bootstrap values. (**B**) Motif analysis. The different colors of boxes denote different motif numbers. The length of box indicates motif length. (**C**) Exon-intron distribution. CDS denotes exons.

Moreover, the predicted number of motifs in the G1, G2, and G3 groups ranges from 2–5, 3–4, and 3–5 motifs, respectively. Therefore, these results strongly support G2 as being more conserved with respect to motif number as compared to other groups. According to the study, for members of the same clade of phylogenetic tree especially, paralogous gene pairs (*VvGASA6*/*VvGASA7* and *VvGASA*1/*VvGASA8*) shared an almost similar motif distribution either with respect to gene length or motif number. Some of the identified motifs were specific to only paralogous gene pairs; for example, motif 5 presented only in *GASA6* and *GASA7* while motif 6 was present only in *GASA1* and *GASA8*. The protein motifs that are limited to only one *VvGASA* group may have some special functions. These results further verify our classification and justify the credibility of phylogeny and exon-intron analysis for classification. To further explore the phylogenetic relationships among grapevine GASA genes, the presence of the GASA domain was examined in all genes. For this, full length VvGASA proteins were aligned. The conserved domain in *VvGASA* sequences were confirmed with SMART and multiple sequence alignment. All of the putative VvGASA proteins shared a conserved GASA domain on the C-terminal comprising about sixty amino acids with twelve cysteine residues (Figure S1).

2.4. Grapevine Genes Duplication and Evolutionary Analysis

According to our results, 14 *VvGASA* genes were randomly distributed on eight out of the 20 chromosomes (Figure 4). Chromosome 18 has the larger proportion of GASA genes (4; 35%). According to the criteria mentioned in Materials and Methods, four genes are tandemly duplicated (Table 3) and clustered by two duplication events on Chromosome 18 (GASA11 and 12) and an uncharacterized chromosome (GASA13 and 14). These duplicated genes belong to groups 1 and 3, and no tandem duplication was observed in group 2. Apart from tandem duplication, four pairs (VvGASA3/9, VvGASA7/6, VvGASA8/5, and VvGASA9/2) (Figure 4, Table S2) of segmental duplication were also observed between seven genes.

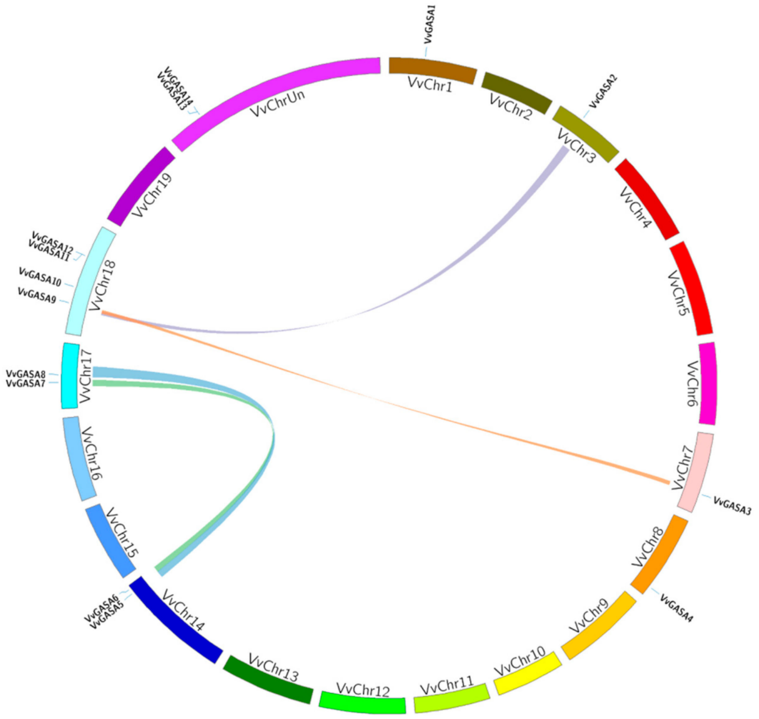


Figure 4. Chromosomal distribution and synteny analysis of grapevine GASA genes. Syntenic regions and chromosomal regions are depicted in different colors (Chr: chromosomes).

Table 3. Duplications of GASA genes in grapes.

Gene1	Gene2	Ka	Ks	Ka/Ks	Selection Pressure	Gene Duplications
GASA7	*GASA6*	0.156	1.5871	0.0928	Purifying selection	Segmental
GASA8	*GASA5*	0.4568	1	0.4568	Purifying selection	Segmental
GASA9	*GASA2*	0.1985	0.929	0.213	Purifying selection	Segmental
GASA3	*GASA9*	0.1781	1.1731	0.151	Purifying selection	Segmental
GASA12	*GASA11*	0.196	0.333	0.585	Purifying selection	Tandem
GASA14	*GASA13*	0.040	0.074	0.540	Purifying selection	Tandem

This indicates that tandem and segmental duplication both have contributed in the expansion of the GASA family in grapevine. Interestingly, *VvGASA9* paired with two genes *GASA3* and *9*. All pairs of duplicated genes (segmental or tandem) "belonged to the same group suggesting common ancestor". In conclusion, 78.5% of GASA genes (11 out of 14) underwent duplication events, which may provide clues for the expansion and functional potential of the GASA gene family. The ratio between the non-synonymous (Ka) and synonymous (Ks) can be used to describe the history of the evolutionary process [27]. The ratio between Ka and Ks was calculated for duplicated gene pairs. All duplicated gene pairs (tandem and segmental duplication) have Ka/Ks values less than 1, which suggests purifying selection, whereas the average of Ka/Ks values were 0.565 and 0.224 in tandemly and segmentally duplicated gene pairs, respectively. According to these results, segmentally duplicated genes are more conserved compared to tandemly duplicated genes.

2.5. GASA Genes Expression Profiling During Seed Development

To identify whether some GASA genes have a role in ovule abortion or seed development, we performed real-time, quantitative RT-PCR of GASA genes during different phases of seed development in seeded and seedless cultivars.

As shown in Figure 5, GASA2, GASA4, and GASA11 were highly expressed during all stages of seed development in seeded cultivars compared to seedless cultivars, whereas GASA6, GASA7, and GASA8 showed expression in the later three seed developmental stages (34, 40, and 50 DAF (days after full bloom)) in the Red globe. However, GASA5 was significantly highly expressed in seedless cultivar compared to seeded grape cultivar. These results suggest that the above-mentioned differentially expressed genes may have a role in ovule abortion or seed development.

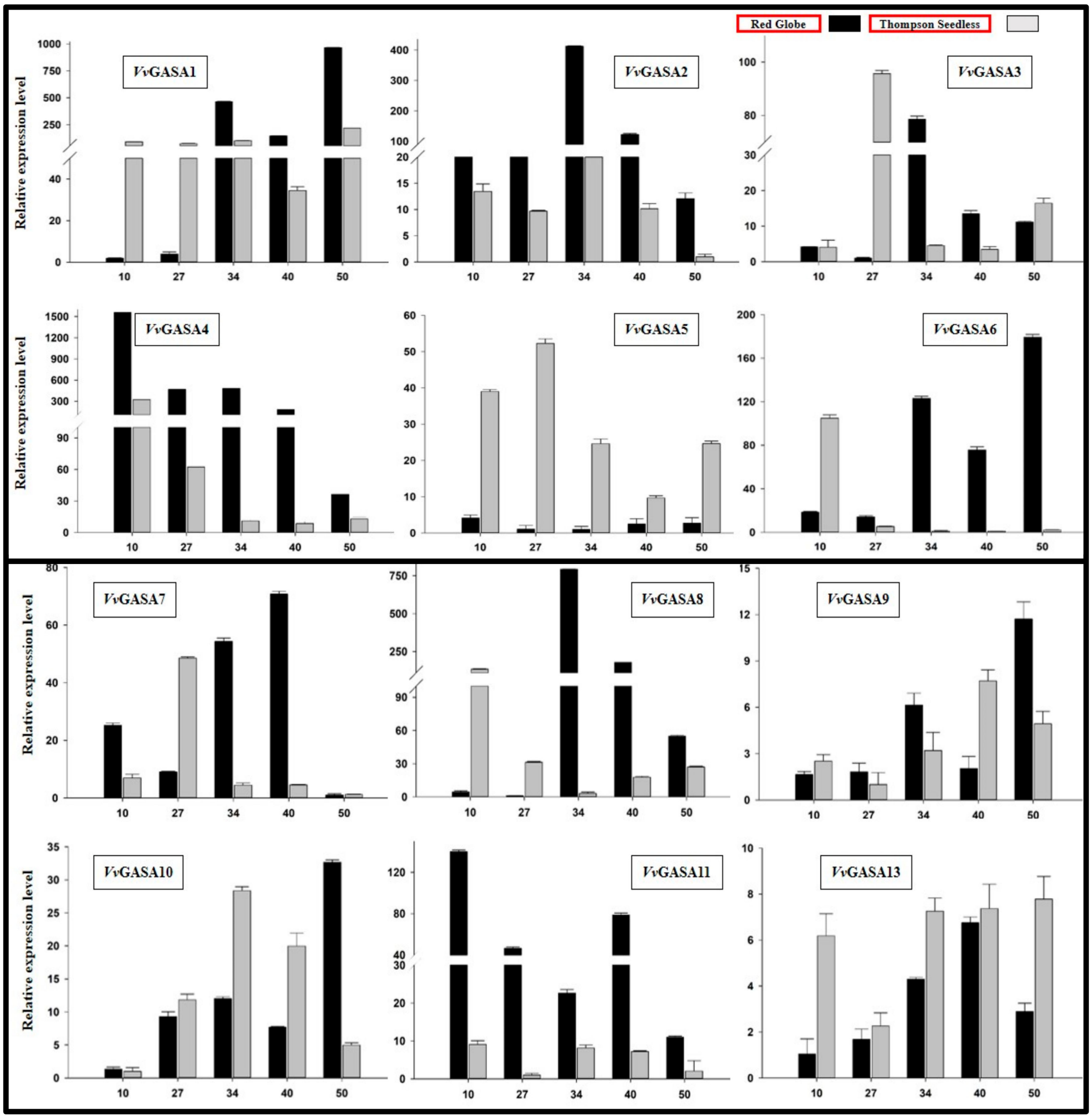


Figure 5. Real-time PCR analysis of grapevine GASA genes at different stages of seed development in seedless and seeded cultivar. Numbers on *x*-axis denote number of days after full bloom (DAF).

2.6. Tissue Specific Expression Profiling of Grapevine GASA Candidates

The spatio-temporal expression analysis of genes can provide information about gene function [28]. We performed real-time RT-PCR for expression profiling of the grapevine GASA genes in the leaf, tendril, stem, flower, and fruit of the Thompson seedless and Red globe (Figure 6).

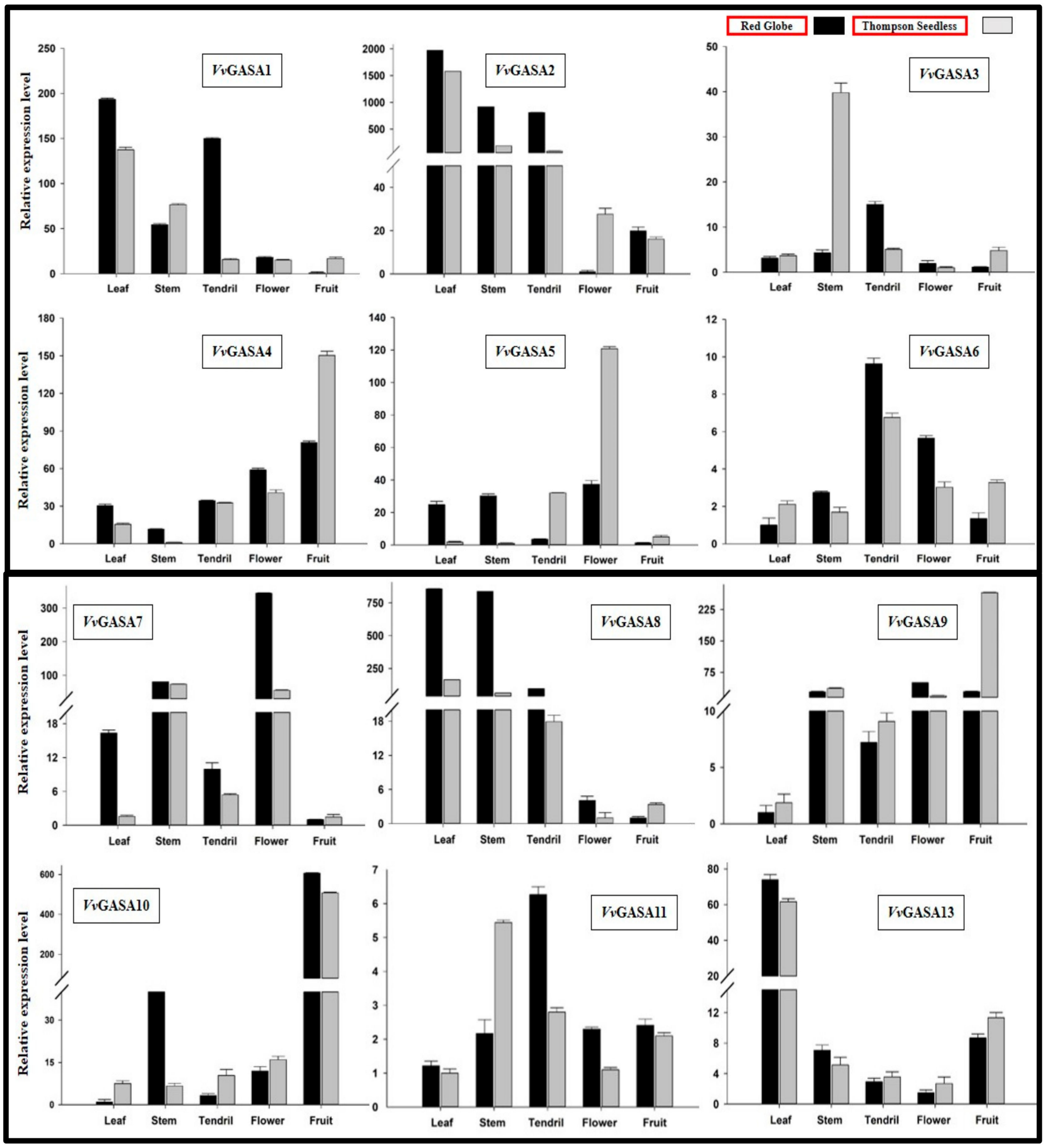


Figure 6. Real-time PCR analysis of different plant parts.

We noted that some of the genes (VvGASA2 and VvGASA11) are expressed relatively ubiquitously. However, most of the GASA genes showed different levels of expression in all tissues both in seeded and seedless cultivars. For example, VvGASA7 and VvGASA8 had greater expression in all tissues (except fruit) of Red Globe, whereas in tissues of Thompson seedless, a moderate level of expression of these genes was noted, suggesting their role in seed development. In general, most of the genes were highly expressed in vegetative plant parts (leaf, stem, and tendril) compared to reproductive organs (flower and fruit), suggesting their role in plant development.

2.7. Promoter Analysis of GASA Genes

To further explore the regulatory mechanisms of grapevine GASA genes, in silico promoter analysis was performed (Figure 7). Several plant hormone (P box, ERE, CGTCA, ABRE, AuxRR-core, TGA-element, GARE, TCA element, and SARE)-related cis-elements were identified in the promoter region of VvGASA genes. However, there were more cis-elements related to ethylene, gibberellic acid and salicylic acid. Cis-elements related to different types of stresses (LTR, STRE, and TCA-motif) and disease resistance (W-box, WUN, WRKY, and TC-rich repeats) were identified in most of the genes. Apart from these, cis-elements involved in endosperm expression (AAGAA-motif) and meristem activation (CCGTCC-box) were identified in the promoters of 10 genes. Moreover, cis-elements having role in anaerobic respiration (ARE) and light response (BOX4, GATA, G BOX, and I BOX) were found in the promoters of all GASA genes.

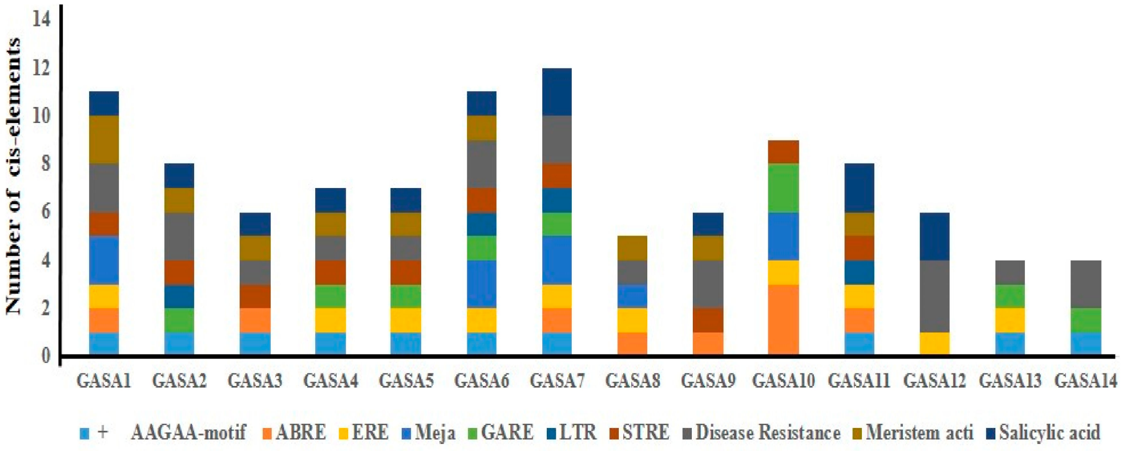


Figure 7. *Cis*-element prediction in the VvGASA promoters.

3. Discussion

GASA gene family members have different critical roles in plant growth and development by influencing plant hormone levels and signal transduction pathways [29]. The members of the same GASA gene family may have the same or reverse functions during vegetative and reproductive growth. Various reports have mentioned negative correlations among GASA gene family members with regard to their function, e.g., overexpression of *AtGASA5* inhibited stem elongation and delay in flowering time, but overexpression of *AtGASA6* promoted early flowering [11]. Moreover, the co-suppression of *GASA4* and *GASA6* in *Arabidopsis* causes a delay in flowering time. Due to these complexities in the functional mechanisms, little information about the exact or more precise functions of GASA genes are available [3,11,30,31]. *PpyGAST* genes influenced bud dormancy by participating in GA biosynthesis and ABA signaling pathways [32]. In strawberry, synergistic action of GAST1 and GAST2 affected fruit cell size, suggesting their role in final fruit size determination [33]. However, according to our information, this study represents the first comprehensive genome-wide identification and expression profiling of GASA genes in grapevine.

In this experiment, we identified 14 GASA genes in grapes, performed comprehensive bioinformatics and expression analysis in different plant structures as well as for different stages of ovule development in seeded and seedless cultivars. The identified genes were divided into three subgroups (G1, G2, and G3) based on their phylogenetic analysis with other species including *Arabidopsis* and apple. According to phylogenetic analysis of GASA genes, grape is more phylogenetically related with *Arabidopsis*. As far as the number of genes in subgroups are concerned, we observed the previous trend [17,18]: Group 3 contained the most (six) genes, whereas G2 contained the least (three) number of genes. According to this study, group 2 is more conserved with regard to exon-intron numbers or conserved motifs, suggesting that during the evolutionary process, other groups (G1 and 3) have either gained or lost exons leading to the difference in the number of exon-introns.

Gene duplication, which has played a major role in the evolution of gene families, can take place through four mechanisms: whole genome duplication, tandem duplication, segmental duplication, and transposition events [34]. However, segmental and tandem duplication have contributed more to the expansion and functional divergence of gene families [35]. According to our findings (Table 3), both segmental and tandem duplications contributed in the evolutionary process of grapevine GASA genes. The result of this study corroborates the previous findings that segmental duplications (7 out of 14 genes) has occurred more frequently compared to tandem duplication (4 out of 14 genes). We noticed an uneven distribution of genes on different chromosomes, as chromosome 18 contained 35% of genes. These findings suggest that duplication of GASA genes has occurred on chromosome 18 during the expansion of the grapevine GASA gene family. This finding is also supported by our observation that most of the (3 out of 4) genes on chromosome 18 underwent duplication and divided into two different groups (G2 and G3).

The Ka/Ks ratio can provide information about phylogenetic reconstruction, evolutionary process, and selection pressure. The evolution of new genes take place due to selection and mutation [36]. According to our results, the Ka/Ks values of all duplicated gene pairs were less than 1, suggesting negative (purifying) selection. Therefore, we can predict that grapevine GASA genes are less exposed to environmental changes. The analysis of the promoter region of a gene can provide clues related to its function and assist in functional characterization [37]. The presence of GA-responsive elements (P Box and GARE) in all members of group1, suggests their role in GA signaling pathways and seed development. Apart from hormone- and disease-related motifs, the presence of motifs involved in endosperm (AAGAA-motif) expression and meristem activation (CCGTCC-box) in the promoters of 10 genes suggests that GASA genes are involved in complex regulatory mechanisms affecting the expression of a gene.

Expression profiling of genes in different plant parts and organs can provide important clues for their functional characterization. In different plant species, GASA genes have shown their spatiotemporal specificity, probably due to their involvement in different hormonal signaling pathways [7,18]. In previous findings, exogenous application of GA increased expression of GsGASA1 in leaves but down-regulated its expression in roots of the soybean [6]. However, in Arabidopsis, GA up-regulated GAST1 expression in meristem tissues but showed negative results in roots and leaves. These results showed that GASA genes have tissue specific responses towards GA application [3]. In our findings, some genes showed tissue-specific expression such as GASA1 and 2 (highly expressed in leaves of both cultivars), whereas GASA9 and 10 showed high expression in the fruit and seed of both cultivars (Figure 6). For example, two paralogous genes, GASA6 and 7, were significantly highly expressed during all phases of ovule development in Red globe compared to Thompson seedless. This suggests that these two genes may have a role in seed development. Our observations are justified by previous findings, as TaGASR7 (Accession number AHM24216), the orthologous gene of VvGASA7, has been well-studied in wheat for its role in grain size length [20]. In *Arabidopsis*, AtGASA4 (AT5G15230), which is highly similar to VvGASA7, has its role in flower meristem development and positively regulates seed size and yield [11]. Apart from these, the orthologous gene of VvGASA7 in rice (OsGASR7; Os06g0266800) also affected grain seed length, suggesting its role in seed development [37]. Therefore, our findings suggest that some grapevine GASA genes may have a role in seed development. The functional characterization of *VvGAS7* will help scientists in exploring the mechanism of seed development in grapevine in future studies.

In contrast to VvGAS7, during seed development stages, VvGASA5 showed high expression in Thompson seedless and almost undetectable expression in Red globe, suggesting its role in ovule abortion. High expression of VvGASA2 during different stages of seed development in seeded cultivars also supports our hypothesis that GASA genes may have a role in seed development. Li et al. [38] reported that knock-down of OsGASR9 (which is homologous of VvGASA2) reduced plant height, seed size, and overall plant yield, whereas its overexpression also increased seed size, plant height, and yield. They reported that these findings are due to the involvement of GASA genes in GA pathways. Most of the GASA genes are involved in the hormonal signaling pathway, especially in GA and ABA, and influence many plant functions, e.g., bud dormancy, seed size, and yield. Although the potential roles of genes can be predicted based on the functions of their orthologous genes in other crops, functional analyses are needed to confirm their roles in specific crops. Finally, the functional characterization of grapevine GASA genes will help scientists in understanding the molecular mechanism of seed development in grapes.

4. Materials and Methods

4.1. Annotation and Identification of Putative Grapevine GASA Genes

We combined two complementary homology-based approaches to identify GASA genes in the grapevine genome. In the first approach, proteins annotated as GASA genes from *Arabidopsis* and apple [17] were used as queries to search open reading frame translations cataloged from a reference grapevine genome sequence (http://www.genoscope.cns.fr), Grapevine Genome CRIBI Biotech website (http://genomes.cribi.unipd.it/), and the National Centre for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/), using the Basic Local Alignment Search Tool [39]. In the second method, the Hidden Markov Model (HMM) profile of the GASA domain (PFAM 02704) was used to search grapevine GASA proteins in the 12X coverage assembly of the *V. vinifera* PN40024 genome [40]. The NCBI Conserved Domain Database (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) and Simple Modular Architecture Research Tool (SMART; http://smart.emblheidelberg.de/) were used to check the presence of the complete GASA domain in obtained protein sequences [41]. Finally, all non-redundant putative protein sequences with a conserved GASA domain were considered and used for further analysis. This approach led to the designation of 14 GASA-domain-encoding genes.

4.2. Physicochemical Properties and Phylogeny Analysis

All identified VvGASA gene protein sequences, coding sequences, genomic sequences, and related information about accession number, start–end position of the gene, the number of amino acids, and chromosome location were downloaded from Grape Genome Database and NCBI. Information about the physiochemical properties of GASA proteins was obtained from the online ExPASy program (http://web.expasy.org/protparam/) using protein sequences [42]. In silico analysis of subcellular location and tertiary structure of proteins was performed using online programs The WOLF PSORT II program (http://www.genscript.com/wolfpsort.html) [43] and PHYRE server v2.0 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index), respectively. Multiple sequence alignment of GASA proteins was completed using DNAMAN (Version 8, Lynnon Bio-soft, Canada) with default parameters. The phylogenetic trees including 55 proteins or with 14 GASA grapevine proteins were constructed with MEGA 5.0 software by using the following parameters: the neighbor-joining (NJ) method, 'W' approach for sequence alignments, 1000 bootstrap iterations, "p-distance", "Complete Deletion", and gap setting [44]. The phylogenetic tree among different plant species included 14, 26, and 15 protein sequences from grapes, apple, and *Arabidopsis*, respectively.

4.3. Exon–Intron, Gene Structures, Conserved Motif, and Promoter Analysis

Exon–intron study of the VvGASA genes was performed using aligned coding sequences and genomic sequences in the online Gene Structure Display Server 2.071 (http://gsds.cbi.pku.edu.cn/index.php). MEME 4.11.2 (http://meme-suite.org/tools/meme), an online program, was used to find up to ten conserved motifs [45]. For cis-acting elements analysis, the upstream sequence (1.5-kb) of each gene was examined through the PlantCARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/) online program.

4.4. Synteny Analysis and Calculation of Ka/Ks Ratio for Duplicated Genes

The tandem and segmental duplication of genes were calculated according to their physical position on individual chromosomes. Two or more genes present on the same chromosome within a 200 kb region were considered as tandemly duplicated [46]. For segmental duplication, data were retrieved from the Plant Genome Duplication Database (http://chibba.agtec.uga.edu/duplication/) [47] and a diagram was generated using the circos program, version 0.63 (http://circos.ca/). The Ka (non-synonymous substitution rate) and Ks (synonymous substitution rate) of duplicated genes were determined using an online tool (http://services.cbu.uib.no/tools/kaks). The ratio of Ka/Ks was used to estimate the selection pressure mode [48]. The Ka/Ks ratio can reveal three different situations: positive (Ka/Ks > 1), negative (Ka/Ks < 1), and neutral (Ka/Ks = 1) [49].

4.5. Plant Materials

In this experiment, plant samples were collected from two grape cultivars, Red globe (Seeded cultivar) and Thompson seedless (seedless cultivar), grown under natural field conditions in the grape orchard of Northwest A&F University, Yangling, China (34°200′ N 108°240′ E). All the samples, including young leaves, tendrils, stems, flowers, and fruits (42 DAF, days after full bloom), were collected from the healthy plants. Apart from these, seed samples were taken at different developmental stages of fruits at 10, 27, 34, 40, and 50 DAF. After collection, all plant parts were immediately frozen in liquid nitrogen and preserved at −80 °C for RNA extraction.

4.6. Total RNA Extraction and Expression Analysis by RT-PCR

The EZNA Plant RNA Kit (R6827-01, OMEGA Biotek, Norcross, GA, USA) and a Nano Drop Spectrophotometer (Thermo Fisher Scientific, Yokohama, Japan) were used to extract and quantify RNA, respectively. Prime Script RTase (Trans Gen Biotech, Beijing, China) was used to synthesize first-strand cDNA from extracted RNA; cDNA was diluted six times and stored at −40 °C for future study. Primer Premier 7.0 (Table S3) was used to design gene-specific primers. Quantitative RT-PCR was carried out for selected genes using SYBR Green (Trans Gen Biotech, Beijing, China) on an IQ5 real-time PCR machine (Bio-Rad, Hercules, CA, USA). The total reaction mixture was 20 µL consisting of 10 µL SYBR green, 7 µL sterile distilled water, 1 µL of cDNA template, 0.8 µL each primer (1.0 µM), and 0.4 µL of Rox reference dye1. The reaction was executed with the following parameters: 95 °C for 30 s, followed by 42 cycles of 95 °C for 10 s and 60 °C for 30 s. The transcript level was normalized by using the *VvActin* gene as an internal control. Each reaction was carried out with three technical and biological replicates. The comparative CT method, also known as the $2^{-\Delta\Delta CT}$method, was used to calculate relative expression levels where ΔΔCT = [(CT target gene − CT control gene) Sample A − (CT target gene − CT control gene) Sample B] [50]. Sigma Plot 12.5 was used to draw graphs.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/21/3/1088/s1.

Author Contributions: X.W., B.A., and S.Z. designed the experiments. B.A. and J.Y. conducted the experiments. V.Y. helped in RT-PCR. X.W. and X.Z. supplied materials/analyses. X.W. provided guidance throughout the study. X.L. helped in bioinformatics work. B.A. and X.W. wrote the manuscript. All authors approved the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China (U1603234), the Program for Innovative Research Team of Grape Germplasm Resources and Breeding (2013KCT-25). We thank Habib-ur-Rehman for editing this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Oliveira-Lima, M.; Benko-Iseppon, A.M.; Neto, J.R.; Rodriguez-Decuadro, S.; Kido, E.A.; Crovella, S.; Pandolfi, V. Snakin: Structure, roles and applications of a plant antimicrobial peptide. *Curr. Protein Pept. Sci.* **2017**, *18*, 368–374. [CrossRef]
2. Shi, L.; Olszewski, N.E. Gibberellin and abscisic acid regulate GAST1 expression at the level of transcription. *Plant Mol. Biol.* **1998**, *38*, 1053–1060. [CrossRef] [PubMed]
3. Aubert, D.; Chevillard, M.; Dorne, A.M.; Arlaud, G.; Herzog, M. Expression patterns of GASA genes in Arabidopsis thaliana: the GASA4 gene is up-regulated by gibberellins in meristematic regions. *Plant Mol. Biol.* **1998**, *36*, 871–883. [CrossRef] [PubMed]
4. Trapalis, M.; Li, S.F.; Parish, R.W. The Arabidopsis GASA10 gene encodes a cell wall protein strongly expressed in developing anthers and seeds. *Plant Sci.* **2017**, *260*, 71–79. [CrossRef] [PubMed]
5. Furukawa, T.; Sakaguchi, N.; Shimada, H. Two OsGASR genes, rice GAST homologue genes that are abundant in proliferating tissues, show different expression patterns in developing panicles. *Genes Genet. Syst.* **2006**, *81*, 171–180. [CrossRef] [PubMed]
6. Li, K.L.; Bai, X.; Li, Y.; Cai, H.; Ji, W.; Tang, L.L.; Wen, Y.D.; Zhu, Y.M. GsGASA1 mediated root growth inhibition in response to chronic cold stress is marked by the accumulation of DELLAs. *J. Plant Physiol.* **2011**, *168*, 2153–2560. [CrossRef] [PubMed]
7. Wang, L.; Wang, Z.; Xu, Y.Y.; Joo, S.H.; Kim, S.K.; Xue, Z.; Xu, Z.H.; Wang, Z.Y.; Chong, K. OsGSR1 is involved in crosstalk between gibberellins and brassinosteroids in rice. *Plant J.* **2009**, *57*, 498–510. [CrossRef] [PubMed]
8. Silverstein, K.A.; Moskal, W.A.; Wu, H.C.; Underwood, B.A.; Graham, M.A.; Town, C.D.; Vanden Bosch, K.A. Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants. *Plant J.* **2007**, *51*, 262–280. [CrossRef]
9. Zhang, S.C.; Wang, X.J. Expression pattern of GASA, downstream genes of DELLA, in Arabidopsis. *Chin. Sci. Bull.* **2008**, *53*, 3839–3846. [CrossRef]
10. Zhang, S.; Yang, C.; Peng, J.; Sun, S.; Wang, X. GASA5, a regulator of flowering time and stem growth in Arabidopsis Thaliana. *Plant Mol. Biol.* **2009**, *69*, 745–759. [CrossRef]
11. Roxrud, I.; Lid, S.E.; Fletcher, J.C.; Schmidt, E.D.; Opsahl-Sorteberg, H.G. GASA4, one of the 14-member Arabidopsis GASA family of small polypeptides, regulates flowering and seed development. *Plant Cell Physiol.* **2007**, *48*, 471–483. [CrossRef] [PubMed]
12. An, B.; Wang, Q.; Zhang, X.; Zhang, B.; Luo, H.; He, C. Comprehensive transcriptional and functional analyses of *HbGASA* genes reveal their roles in fungal pathogen resistance in *Hevea brasiliensis*. *Tree Genet. Genom.* **2018**, *14*, 41. [CrossRef]
13. Nahirñak, V.; Rivarola, M.; Gonzalez de Urreta, M.; Paniego, N.; Hopp, H.E.; Almasia, N.I.; Vazquez-Rovere, C. Genome-wide Analysis of the Snakin/GASA Gene Family in *Solanum tuberosum* cv. Kennebec. *Am. J. Potato Res.* **2016**, *93*, 172–188. [CrossRef]
14. Herzog, M.; Dorne, A.M.; Grellet, F. GASA, a gibberellin-regulated gene family from *Arabidopsis thaliana* related to the tomato GAST1 gene. *Plant Mol. Biol.* **1995**, *27*, 743–752. [CrossRef]
15. Ben-Nissan, G.; Lee, J.Y.; Borohov, A.; Weiss, D. GIP, a Petunia Hybrida GA induced cysteine-rich protein: A possible role in shoot elongation and transition to flowering. *Plant J.* **2004**, *37*, 229–238. [CrossRef] [PubMed]
16. Zhang, L.Y.; Geng, X.L.; Zhang, H.Y.; Zhou, C.L.; Zhao, A.J.; Wang, F.; Zhao, Y.; Tian, X.J.; Hu, Z.R.; Xin, M.M.; et al. Isolation and characterization of heat-responsive gene TaGASR1 from wheat (*Triticum aestivum* L.). *J. Plant Biol.* **2017**, *60*, 57–65. [CrossRef]
17. Fan, S.; Zhang, D.; Xing, L.; Qi, S.; Du, L.; Wu, H.; Shao, H.; Li, Y.; Ma, J.; Han, M. Comprehensive analysis of GASA family members in the *Malus domestica* genome: Identifcation, characterization, and their expressions in response to apple flower induction. *BMC Genom.* **2017**, *18*, 827. [CrossRef]
18. Ahmad, M.Z.; Sana, A.; Jamil, A.; Nasir, J.A.; Ahmed, S.; Hameed, M.U.; Abdullah. A genome-wide approach to the comprehensive analysis of GASA gene family in *Glycine max*. *Plant Mol. Biol.* **2019**, *100*, 607. [CrossRef]
19. Ling, H.Q.; Zhao, S.; Liu, D.; Wang, J.; Sun, H.; Zhang, C.; Fan, H.; Li, D.; Dong, L.; Tao, Y. Draft genome of the wheat A-genome progenitor *Triticumurartu*. *Nature* **2013**, *496*, 487–490. [CrossRef]

20. Dong, L.; Wang, F.; Liu, T.; Dong, Z.; Li, A.; Jing, R.; Mao, L.; Li, Y.; Liu, X.; Zhang, K.; et al. Natural variation of TaGASR7-1 A1affects grain length in common wheat under multiple cultivation conditions. *Mol. Breed.* **2014**, *34*, 937. [CrossRef]
21. Muhammad, I.; Li, W.Q.; Jinga, X.Q.; Zhoua, M.R.; Shalmania, A.; Ali, M.; Weia, X.Y.; Sharif, R.; Liua, W.T.; Chen, K.M. A systematic in silico prediction of Gibberellic acid stimulated GASA family members: a novel small peptide contributes to floral architecture and transcriptomic changes induced by external stimuli in rice. *J. Plant Physiol.* **2019**. [CrossRef] [PubMed]
22. FAO STAT 2018. Available online: http://www.fao.org/land-water/databases-and-software/crop-information/grape/en/ (accessed on 30 December 2019).
23. This, P.; Lacombe, T.; Thomas, M.R. Historical origins and genetic diversity of wine grapes. *Trends. Genet.* **2006**, *22*, 511–519. [CrossRef] [PubMed]
24. FAO; OIV. *FAO-OIV Focus Table and Dried Grapes*; Non-Alcoholic Products of the Vitivini Cultural Sector Intended for Human Consumption; FAO: Rome, Italy, 2016; Volume I7042, p. 64. Available online: www.oiv.int/public/medias/5268/fao-oiv-focus-2016 (accessed on 30 December 2019).
25. Zhang, S.; Wang, L.; Sun, X.; Li, Y.; Yao, J.; Nocker, S.v.; Wang, X. Genome-Wide Analysis of the YABBY Gene Family in Grapevine and Functional Characterization of VvYABBY4. *Front. Plant Sci.* **2019**, *10*, 1207. [CrossRef] [PubMed]
26. Li, Y.D.; Zhang, S.L.; Dong, R.Z.; Wang, L.; Yao, J.; Nocker, S.V.; Wang, X. The grapevine homeobox gene VvHB58 infuences seed and fruit development through multiple hormonal signaling pathways. *BMC Plant Biol.* **2019**, *19*, 523. [CrossRef]
27. Li, W.H.; Gojobori, T.; Nei, M. Pseudogenes as a paradigm of neutral evolution. *Nature* **1981**, *292*, 237–239. [CrossRef]
28. Bassett, D.E.; Eisen, M.B.; Boguski, M.S. Gene expression informatics–it's all in your mine. *Nat. Genet.* **1999**, *21*, 51–55. [CrossRef]
29. Zhang, S.; Wang, X. One new kind of phytohormonal signaling integrator: Up-and-coming GASA family genes. *Plant Signal. Behav.* **2017**, *12*, e1226453-2. [CrossRef]
30. Rubinovich, L.; Weiss, D. The Arabidopsis cysteine-rich protein GASA4 promotes GA responses and exhibits redox activity in bacteria and in planta. *Plant J.* **2010**, *64*, 1018–1027. [CrossRef]
31. Qu, J.; Kang, S.G.; Hah, C.; Jang, J.C. Molecular and cellular characterization of GA-Stimulated Transcripts GASA4 and GASA6 in Arabidopsis thaliana. *Plant Sci.* **2016**, *246*, 1–10. [CrossRef]
32. Yang, Q.; Niu, Q.; Tang, Y.; Ma, Y.; Yan, X.; Li, J.; Tian, J.; Bai, S.; Teng, Y. *PpyGAST1* is potentially involved in bud dormancy release by integrating the GA biosynthesis and ABA signaling in 'Suli' pear (*Pyruspyrifolia* White Pear Group). *Environ. Exp. Bot.* **2019**, *162*, 302–312. [CrossRef]
33. Moyano-Cañete, E.; Bellido, M.L.; García-Caparrós, N.; Medina-Puche, L.; AmilRuiz, F.; González-Reyes, J.A.; Caballero, J.L.; Muñoz-Blanco, J.; Blanco-Portales, R. FaGAST2, a strawberry ripening related gene, acts together with faGAST1 to determine cell size of the fruit receptacle. *Plant Cell Physiol.* **2013**, *54*, 218–236. [CrossRef] [PubMed]
34. Paterson, A.H.; Freeling, M.; Tang, H.; Wang, X. Insights from the comparison of plant genome sequences. *Annu Rev. Plant Biol.* **2010**, *61*, 349–372. [CrossRef] [PubMed]
35. Cannon, S.B.; Mitra, A.; Baumgarten, A.; Young, N.D.; May, G. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol.* **2004**, *4*, 10. [CrossRef] [PubMed]
36. Morgan, C.C.; Loughran, N.B.; Walsh, T.A.; Harrison, A.J.; O'Connell, M.J. Positive selectionneighboring functionally essential sites and disease-implicated regions of mammalian reproductiveproteins. *BMC Evol. Biol.* **2010**, *10*, 39. [CrossRef]
37. Huang, X.; Zhao, Y.; Wei, X.; Li, C.; Wang, A.; Zhao, Q.; Li, W.; Guo, Y.; Deng, L.; Zhu, C. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* **2012**, *44*, 32–39. [CrossRef]
38. Li, X.; Shi, S.; Tao, Q.; Tao, Y.; Miao, J.; Peng, X.; Li, C.; Yang, Z.; Zhou, Y.; Liang, G. OsGASR9 positively regulates grain size and yield in rice (*Oryza sativa*). *Plant Sci.* **2019**, *286*, 17–27. [CrossRef]
39. Altschul, S.F.; Warren, G.; Webb, M.; Eugene, W.M.; David, J.L. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]

40. Jaillon, O.; Aury, J.M.; Noel, B.; Policriti, A.; Clepet, C.; Casagrande, A.; Choisne, N.; Aubourg, S.; Vitulo, N.; Jubin, C.; et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **2007**, *449*, 463–467.
41. Ahmad, B.; Zhang, S.; Yao, J.; Rahman, M.U.; Hanif, M.; Zhu, Y.; Wang, X. Genomic Organization of the B3-Domain Transcription Factor Family in Grapevine (*Vitis vinifera* L.) and Expression during Seed Development in Seedless and Seeded Cultivars. *Int. J. Mol. Sci.* **2019**, *20*, 4553. [CrossRef]
42. Artimo, P.; Jonnalagedda, M.; Arnold, K.; Baratin, D.; Csardi, G.; De Castro, E.; Duvaud, S.; Flegel, V.; Fortier, A.; Gasteiger, E. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* **2012**, *40*, W597–W603. [CrossRef]
43. Horton, P.; Park, K.J.; Obayashi, T.; Fujita, N.; Harada, H.; Adams-Collier, C.J.; Nakai, K. WoLF PSORT: Protein localization predictor. *Nucleic Acids Res.* **2007**, *35*, 585–587. [CrossRef] [PubMed]
44. Tamura, K.; Peterson, D.; Peterson, N.; Stecher, G.; Nei, M.; Kumar, S. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **2011**, *28*, 2731–2739. [CrossRef] [PubMed]
45. Bailey, T.L.; Williams, N.; Misleh, C.; Li, W.W. MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **2006**, *34*, W369–W373. [CrossRef] [PubMed]
46. Holub, E.B. The arms race is ancient history in Arabidopsis, the wildflower. *Nat. Rev. Genet.* **2001**, *2*, 516–527. [CrossRef] [PubMed]
47. Lee, T.H.; Tang, H.; Wang, X.; Paterson, A.H. PGDD: A database of gene and genome duplication in plants. *Nucleic Acids Res.* **2012**, *41*, D1152–D1158. [CrossRef]
48. Lynch, M.; Conery, J.S. The evolutionary fate and consequences of duplicate genes. *Science* **2000**, *290*, 1151–1155. [CrossRef]
49. Juretic, N.; Hoen, D.R.; Huynh, M.L.; Harrison, P.M.; Bureau, T.E. The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. *Genome Res.* **2005**, *15*, 1292–1297. [CrossRef]
50. Schmittgen, T.D.; Livak, K.J. Analyzing real-time PCR data by the comparative CT method. *Nat. Protoc.* **2008**, *3*, 1101–1108. [CrossRef]

International Journal of
Molecular Sciences

MDPI

Article

Molecular Characterization and Disease Control of Stem Canker on Royal Poinciana (*Delonix regia*) Caused by *Neoscytalidium dimidiatum* in the United Arab Emirates

Seham M. Al Raish [1], Esam Eldin Saeed [2], Arjun Sham [1], Khulood Alblooshi [1], Khaled A. El-Tarabily [1,2,3,*] and Synan F. AbuQamar [1,*]

[1] Department of Biology, College of Science, United Arab Emirates University, Al-Ain 15551, UAE; 200440261@uaeu.ac.ae (S.M.A.R.); arjunsham@uaeu.ac.ae (A.S.); 201404120@uaeu.ac.ae (K.A.)

[2] Khalifa Center for Genetic Engineering and Biotechnology, United Arab Emirates University, Al-Ain 15551, UAE; esameldin_saeed@uaeu.ac.ae

[3] College of Science, Health, Engineering and Education, Murdoch University, Murdoch, WA 6150, Australia

* Correspondence: ktarabily@uaeu.ac.ae (K.A.E.-T.); sabuqamar@uaeu.ac.ae (S.F.A.);
Tel.: +971-3-713-6518 (K.A.E.-T.); +971-3-713-6733 (S.F.A.)

Received: 15 December 2019; Accepted: 1 February 2020; Published: 4 February 2020

Abstract: In the United Arab Emirates (UAE), royal poinciana (*Delonix regia*) trees suffer from stem canker disease. Symptoms of stem canker can be characterized by branch and leaf dryness, bark lesions, discoloration of xylem tissues, longitudinal wood necrosis and extensive gumming. General dieback signs were also observed leading to complete defoliation of leaves and ultimately death of trees in advanced stages. The fungus, *Neoscytalidium dimidiatum* DSM 109897, was consistently recovered from diseased royal poinciana tissues; this was confirmed by the molecular, structural and morphological studies. Phylogenetic analyses of the *translation elongation factor 1-a* (*TEF1-α*) of *N. dimidiatum* from the UAE with reference specimens of Botryosphaeriaceae family validated the identity of the pathogen. To manage the disease, the chemical fungicides, Protifert®, Cidely® Top and Amistrar® Top, significantly inhibited mycelial growth and reduced conidial numbers of *N. dimidiatum* in laboratory and greenhouse experiments. The described "apple bioassay" is an innovative approach that can be useful when performing fungicide treatment studies. Under field conditions, Cidely® Top proved to be the most effective fungicide against *N. dimidiatum* among all tested treatments. Our data suggest that the causal agent of stem canker disease on royal poinciana in the UAE is *N. dimidiatum*.

Keywords: chemical fungicide; disease control; *Neoscytalidium dimidiatum*; royal poinciana; stem canker; UAE

1. Introduction

Royal poinciana (*Delonix regia* (Bojer ex Hook.) Raf.) is a beautiful flowering and shady branching tree. This member of the pea family (Fabaceae), which is also known as flamboyant, peacock or flame tree, can be recognized by the color of flowering cultivars, ranging from deep red to bright orange or yellow [1]. It is a rapid growing tree that can reach to 6–12 m height, and bears compound leaves that reach 30–60 cm length and flat woody pod fruits of about 60 cm long [2]. Despite it is native to Madagascar and tropical regions, this deciduous tree provides landscape with cooling shade during hot summers and warming-sunshine winters. In addition to the "umbrella" canopy it provides, royal poinciana can grow in a variety of soil conditions, and is highly tolerant to drought and salinity [3]. For that reason, there is a growing interest in the plantations of royal poinciana in the United Arab

Emirates (UAE), mainly in parks, sidewalks, streets, parking lots and open areas. Although, this tree does not often suffer from real problems, stem canker has currently become a serious disease affecting royal poinciana. Therefore, it is urgent to address this present threat to royal poinciana in the UAE and worldwide.

Like other ornamental and stone fruit trees, fungi can attack different parts or tissues of royal poinciana under certain favorable conditions to cause canker diseases [3,4]. In general, cankers are destructive diseases which may cause damage to the whole or parts of trees such as branches, barks and woods. Fungi such as *Nectria galligena*, *Leptosphaeria maculans*, *Lasiodiplodia theobromae* and *Teratospheria zuluensis* are among those associated with canker diseases on sweet birch tree (*Betula lenta*), oilseed rape (*Brassica napus*), eucalypt and pine trees [5–8]. *Neoscytalidium dimidiatum* is another fungal pathogen that causes cankers and has a wide geographical and host range, including plum, almond (*Prunus dulcis*), mango (*Mangifera indica*), pitahaya (*Hylocereus undatus*), Citrus, Musa, Populus, and Ficus spp. in Australia, China, Egypt, Niger, Tunisia and the USA [9–15].

In Oman, stem canker has been reported on different trees including royal poinciana [16]. Symptoms can be recognized as branch wilt, dieback, canker, gummosis and death of infected trees. In general, severity of the disease caused by this fungus can be enhanced by stress factors such as water stress [9,16]. In the UAE, recent studies on tree diseases caused by fungi have reported black scorch disease and sudden decline syndrome (SDS) on date palm, and dieback disease on mango caused by *Thielaviopsis punctulata*, *Fusarium solani* and *L. theobromae*, respectively [17–19]. So far, there are no reports about royal poinciana-*N. dimidiatum* interaction causing stem canker disease in the UAE.

Plant disease management mainly relies on the life cycle of the pathogen. *N. dimidiatum* produces two types of spores, pycniospores which are formed in pycnidia embedded in mature lesions and phragmospores which are formed by the breaking up of individual or groups of cells of mature hyphae in dead tissues of the lesion [20,21]. In culture, only phragmospores are formed and produced. Typically, cultural and horticultural practices such as pruning and fertilization may lower the risk of the pathogen, increase the vigor of the tree and extend its life [3]. On the other hand, such practices can be harmful due to the improper timing, unsterile tools, inexperienced persons or advanced stages of the pathogen's life cycle. Regardless of its ecological problems and human health concerns, the use of chemical fungicides is yet the main disease management tactic to attenuate the threat of crop diseases [17–19,22]. In vitro treatment with the chemical, Beltanol-L (8-hydroxyquinoline), effectively inhibited the growth of *N. dimidiatum* in vitro [23]. The same fungicide also reduced symptoms of canker lesions on the seedlings of *Eucalyptus camaldulensis* under greenhouse conditions. Application of any of the systemic fungicides, Elsa® (carbendizim), Mizab® (mancozeb) or Curzate® (cymoxamil), showed a significant inhibition to this fungus that causes wilt and canker diseases on cypress trees [24]. Hence, one should take into consideration the timing for minimum effective dose of the fungicide application to control the disease.

Our long-term goal is to develop and implement integrated disease management (IDM) strategies using a combination of cultural, chemical and biological control with resistant cultivars of royal poinciana to manage stem canker disease. In the present investigation, an attempt was made to explore the feasibility of using efficient chemical fungicide(s) for the management of stem canker of royal poinciana. Therefore, our objectives were to: (1) isolate and identify the pathogen associated with infected plants; (2) evaluate the efficacy of fungicides against the causal agent of stem canker *in vitro*; (3) assess the potential fungicides against the pathogen in vivo under greenhouse conditions; and (4) manage disease of naturally infested plants in the field using the proper fungicide treatment. Here, we reported the assessment of systemic chemical fungicide treatments against *N. dimidiatum in vitro*, in the greenhouse as well as in the field. We also developed a short-term strategy to reduce the economic losses associated with stem canker disease. Future directions to employ research on biological control agents (BCAs) to suppress the damaging activities of the pathogen and to lower the risk of the disease on royal poinciana will further cooperate in the development of effective IDM programs.

2. Results

2.1. Symptoms of Stem Canker Disease on Royal Poinciana

Disease symptoms of stem/branch cankers associated with dieback were observed in the orchard of royal poinciana distributed in Dubai Festival City (DFC), UAE (Figure 1A). Apparently, the pathogen was able to attack different tissues of royal poinciana, and the trees were severely affected leading to progressive dieback. In general, cankers on branches were detected in young trees. Stem cankers were observed in old and mature trees, and were associated with pruning wounds and other wounds (Figure 1A).

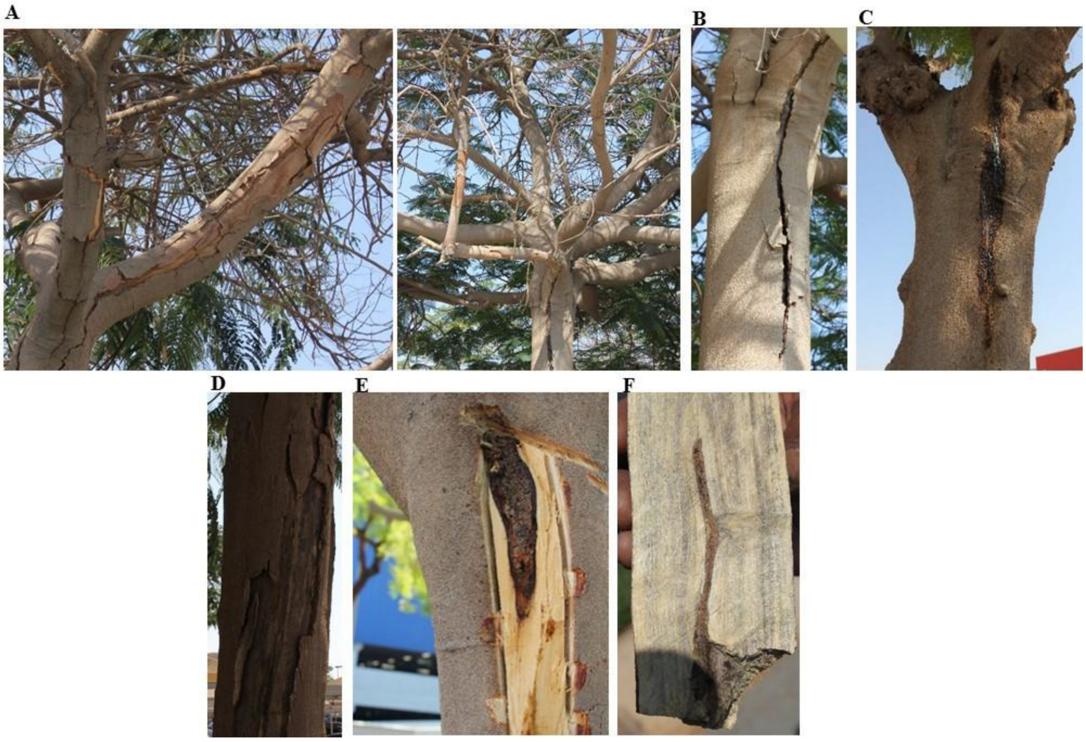


Figure 1. Symptoms of stem canker on trees of royal poinciana. (**A**) Severe symptoms of canker (left) and dieback (right); (**B**) typical longitudinal canker symptoms on stem; (**C**) gumming symptoms of the disease on the bark with fungal growth apparent beneath periderm; (**D**) main stem with the black stromata where the periderm has peeled away; (**E**) canker associated with internal symptoms in the trunk; and (**F**) affected vascular tissues. In (**A–F**), naturally infested royal poinciana trees with *N. dimidiatum* in DFC, UAE.

Cankers were developed longitudinally (Figure 1B), causing dark discoloration of xylem tissues and extensive gumming (Figure 1C). The main stem was often associated with black stromata, resulting the epidermis to peel away (Figure 1D). The discoloration continued outward, rotting symptoms led to spur and shoot blight was also observed. Sap was initially amber in color but later became dark. Internally, canker (Figure 1E) and affected vascular tissues (Figure 1F) were associated with this disease. Eventually, all royal poinciana trees were simultaneously found infected in the orchard (Figure 1A). These signs on royal poinciana are typical of stem canker that is known to be caused by a soil-borne wound pathogen. Therefore, attempts to isolate the putative pathogen from diseased royal poinciana was the first step in identifying the causal agent of this disease.

2.2. Identification and Molecular Characterization of Neoscytalidium Dimidiatum

First, we isolated the fungus from different symptomatic tissues on potato dextrose agar (PDA). From the cultural characteristics, the fungus grew and colonized the plate rapidly. It produced cream to white effuse, hairy to woolly colonies after 2 days of incubation (Figure 2A). The colonies turned olive green, greyish to ochraceous yellow color after 4 days. The fungus showed dark grey to black pigmentation at 8 and 12 days of incubation, respectively (Figure 2A). Microscopically, we observed mycelial growth (Figure 2B) and production of scytalidium-like anamorph of different maturity stages of arthoconidia segmenting from the hyphae (Figure 2C). We also noted that various conidial shapes ranging from ellipsoid to ovoid, rod shaped or round shaped, to hyaline with an acutely rounded apex, truncate base. Conidia were initially aseptate and brownish; at maturity, 0- to 2-septate, central cells were darker than the end cells, measuring 11.02 ± 0.33 x 4.98 ± 0.41 µm (Figure 2C). Conidiogenous cells, or pycnidial anamorph, were described as hyaline and intermingled with paraphyses, forming pycnidiospores after 25 days of incubation (Figure 2D). Cultures also produced fusicoccum-like conidia in pycnidia (Figure 2E). Together, the cultural and morphological characteristics suggest that this fungal isolate may belong to *Neoscytalidium* spp. [25]. Thus, molecular characteristics can identify the fungal specimen at the species level.

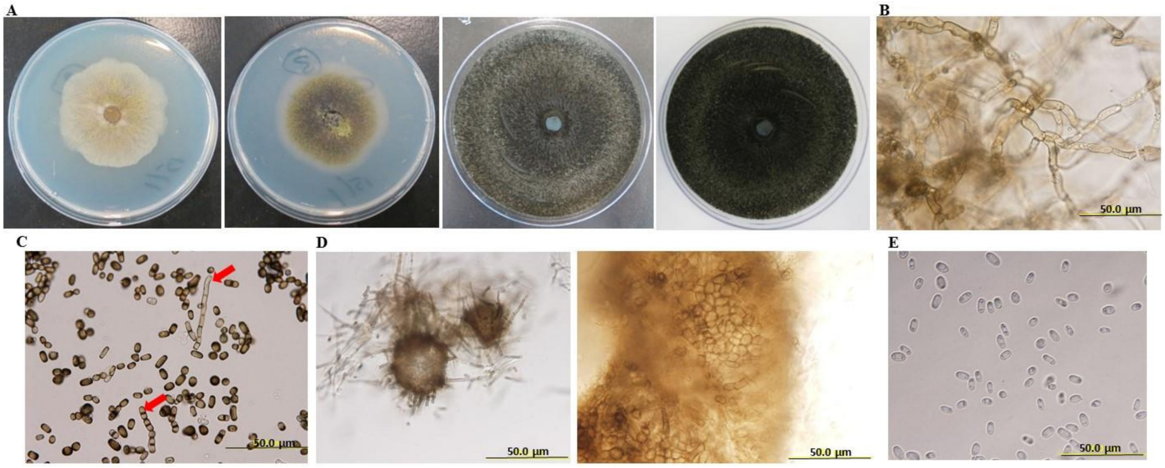


Figure 2. Cultural and morphological characteristics of *Neoscytalidium dimidiatum*. (**A**) Colonies on PDA (left to right: 2, 4, 8 and 12 days of incubation at 25 ± 2 °C); (**B**) mycelia; (**C**) scytalidium-like anamorph showing various shapes and maturity stages of arthroconidia (red arrows) segmenting from hyphae; (**D**) pycnidia formed on a 25-day-old colony (left) and pycnidiospores (right) on PDA; and (**E**) fusicoccum-like pycnidial conidia (immature).

DNA-based methods are widely used to detect and identify plant pathogens. First, we isolated the fungal DNA from the PDA-grown mycelium from each tissue (stems, branches and leaves) sample. Polymerase chain reaction (PCR) amplification using primers targeting the genomic regions of *internal transcribed spacer* (*ITS*), 28S rDNA region, *translational elongation factor 1-α* (*TEF1-α*) and *β-tubulin* was performed. The amplification product of all tested genes was clearly generated in all tested specimens (Figure 3A). Because there was no available DNA sequences about the strain isolated from the UAE, the *ITS* and *TEF1-α* genes [26] were further sequenced. Sequences obtained from *ITS/LSU* rDNA and *TEF1-α* genes were also deposited in GenBank under the accession number, MN371844 and MN447201, respectively. Our data suggest that *Neoscytalidium* spp. is probably the potential fungal pathogen commonly associated with stem canker disease symptoms on royal poinciana trees.

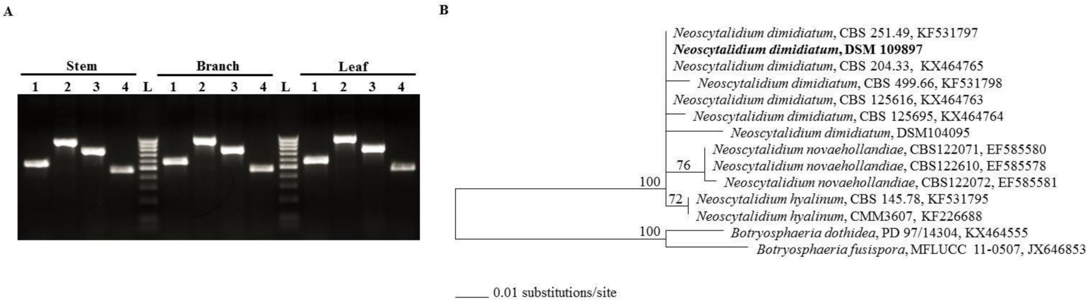

A
B
Stem
Branch
Leaf
1 2 3 4 L 1 2 3 4 L 1 2 3 4
Neoscytalidium dimidiatum, CBS 251.49, KF531797
Neoscytalidium dimidiatum, DSM 109897
Neoscytalidium dimidiatum, CBS 204.33, KX464765
Neoscytalidium dimidiatum, CBS 499.66, KF531798
Neoscytalidium dimidiatum, CBS 125616, KX464763
Neoscytalidium dimidiatum, CBS 125695, KX464764
Neoscytalidium dimidiatum, DSM104095
Neoscytalidium novaehollandiae, CBS122071, EF585580
76
Neoscytalidium novaehollandiae, CBS122610, EF585578
100
Neoscytalidium novaehollandiae, CBS122072, EF585581
72
Neoscytalidium hyalinum, CBS 145.78, KF531795
Neoscytalidium hyalinum, CMM3607, KF226688
100
Botryosphaeria dothidea, PD 97/14304, KX464555
Botryosphaeria fusispora, MFLUCC 11-0507, JX646853
0.01 substitutions/site

Figure 3. Molecular identification of *Neoscytalidium dimidiatum.* PCR amplification of specific genomic DNA regions of infected stem, branch and leaf tissues (**A**); and dendrogram showing phylogenetic relationships among *N. dimidiatum* (DSM 109897) identified in this study and other members of *Neoscytalidium* spp. prepared by the Maximum Likelihood (ML) method (**B**). In (**A**), lanes 1-4 correspond to amplifications of *ITS*, 28S rDNA region, *TEF1-α* and *β-tubulin*, respectively, in trunk (stem), branches and leaves. In (**B**), the ML tree was obtained from *TEFα-1* sequence data. The specimens used in this study carry GenBank accession number, *N. dimidiatum TEF1-α* (MN447201). Numbers at the nodes are bootstrap values after 100 replicates are expressed as percentages (LnL = −603.684353). Only values above 70% are indicated. The scale bar on the rooted tree indicates a 0.01 substitution per nucleotide position. The strain of *N. dimidiatum* from this report is indicated in bold. *Botryosphaeria dothidea*, PD 97/14304 (KX464555) and *B. fusispora* MFLUCC 11-0507 (JX646853) were used as outgroups. *ITS*, *internal transcribed spacer*; 28S rDNA, large subunit (LSU) of rDNA; *TEF1-α*, *translational elongation factor 1-α*; L, DNA ladder.

Second, a phylogenic tree using the obtained *TEF1-α* sequence was compared to other closely related sequences in order to determine the relationship with closely *TEF1-α* related sequences coming from other *Neoscytalidium* spp. The *TEF1-α* sequence of the strain isolated from the UAE grouped in a clade representing *N. dimidiatum* (Penz.) Crous & Slippers [27] (Figure 3B). Results of the Maximum Likelihood (ML) tree indicated that the isolate, in the current study, showed >99% identity with the other isolates of *N. dimidiatum*. These isolates have been collected from different plant species such as *Juglan regia* (CBS 251.49), *Prunus* sp. (CBS 204.33), pacific madrone (*Arbutus menziesii*; CBS 204.33 and CBS 499.66), mango (*Megnifera indica*; CBS 499.66) and others (CBS 125616, CBS 125695 and DSM 104095). The identified fungal species and the other *N. dimidiatum* separately clustered from the two other species of Botryosphaeriaceae, *N. novaehollandiae* and *N. hyalinum*; thus, this isolate was identified as *N. dimidiatum*. Together, this suggests that *N. dimidiatum* (DSM 109897) is most likely the causal species of stem canker disease on royal poinciana.

2.3. Pathogenicity Tests of Neoscytalidium Dimidiatum on Royal Poinciana Seedlings and Apple Fruits

Disease progress on one-year-old royal poinciana seedlings inoculated with 8-mm mycelial discs from 10-day-old pure culture of *N. dimidiatum* growing on PDA was regularly monitored in the greenhouse. Based on artificial inoculations, pathogenicity tests led to the development of disease symptoms on royal poinciana seedlings (Figure 4A–C). Typical symptoms of stem canker developed at the point of inoculation on the stem on plants following *N. dimidiatum* infection. At 2 weeks post inoculation (wpi), dark brown lesions formed on the surface of the stem, leaves became pale, turned yellowish in color and dropped off (Figure 4A). The disease progressed upward along the stem with black, necrotic lesions appeared at the site of inoculation; subsequently the infected stem rotted at 5 wpi. In addition, a general dryness in the plant was recognized forcing the leaves to fall (Figure 4B). In contrast, no symptoms were noticed in control seedlings. The pathogen was consistently re-isolated from all inoculated tissues and identified by conidial morphology, fulfilling Koch’s postulates (Figure 4C).

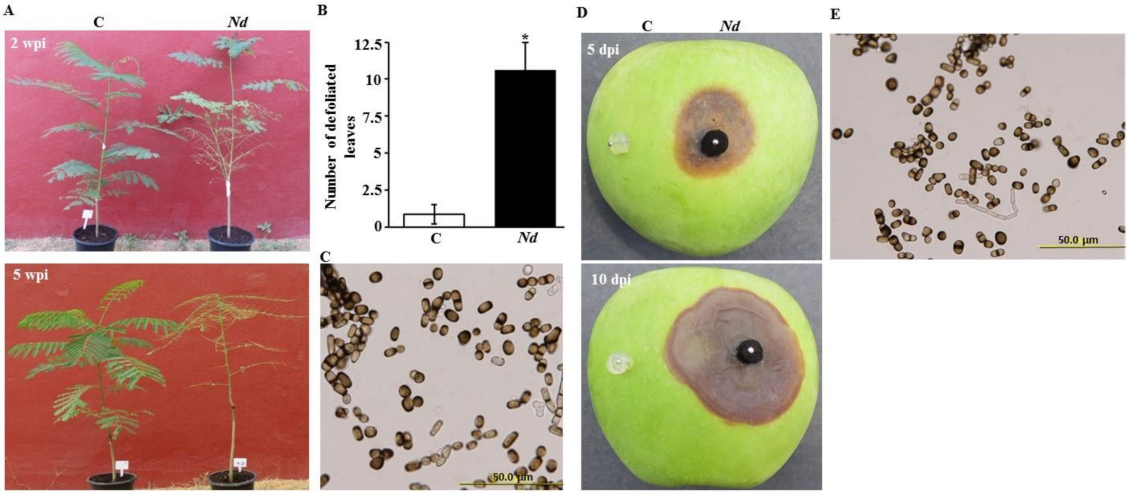


Figure 4. Development of canker on royal poinciana seedlings and apple fruits following artificial inoculation with *Neoscytalidium dimidiatum.* Pathogenicity test on royal poinciana seedlings inoculated (*Nd*; right) and non-inoculated (C; left) with *N. dimidiatum* at (**A**) 2 and 5 wpi; (**B**) number of defoliated leaves of inoculated and control seedlings; and (**C**) conidia after re-isolation of the pathogen from colonized stem tissues, at 5 wpi. Pathogenicity tests on (**D**) inoculated (right) and non-inoculated (left) apple fruits at 5 and 10 dpi; and (**E**) conidia of the pathogen from the inoculated apple fruits at 10 dpi. In (**B**), mean values followed by an asterisk are significantly different from control treatment at the tested time ($p < 0.05$). Experiments were repeated at least three times with similar results. C, control (no *N. dimidiatum*); *Nd*, *N. dimidiatum*.; dpi/wpi, days/weeks post inoculation.

Under laboratory conditions, apple fruits were also inoculated with the same pathogen. At 5 days post inoculation (dpi), we observed discoloration of apple tissues which expanded slowly underneath the PDA plugs containing the pathogen (Figure 4D). After 10 dpi, the fungus grew into apple tissues causing rapid spreading water-soaked lesions. By peeling away the skin from the discolored tissue and placing it on PDA Petri dishes, pure cultures recovered and conidia of *N. dimidiatum* were re-isolated (Figure 4E). No disease symptoms were evident on the same apple fruit under the control plug without the pathogen at 5 and 10 dpi (Figure 4D). Altogether, disease symptoms associated with the inoculated royal poinciana seedlings and apple fruits suggest that the Koch's postulates are fulfilled and that *N. dimidiatum* is most likely the causal agent of the stem canker disease on royal poinciana.

2.4. In Vitro Evaluation of Chemical Fungicides to Neoscytalidium Dimidiatum

To determine their effects on the mycelial growth of *N. dimidiatum*, PDA plates containing a final concentration of 0, 250, 500 and 1000 ppm of the chemical fungicides -available in the market- were evaluated in vitro (Figure S1). In general, we noticed varied response of *N. dimidiatum* to the selected fungicides. For example, application of the fungicides, Penthiopyrad®, Proxanil®, Protoplant® and Previcur® at 250 ppm (the lowest tested concentration) showed minimal or no effect on the mycelial growth of the fungus (Figure 5A). When the chemical fungicides Amistar Top®, Uniform®, Cidely® Top, Protifert® and Airone Liquido® were, however, supplied in PDA medium, there was greater inhibition in the mycelial growth of *N. dimidiatum* at all the concentrations examined in vitro (Figure S1) including the concentration of 250 ppm (Figure 5A). These promising fungicides were also statistically ($p < 0.05$) assessed at the concentration of 250 ppm for their efficacy to inhibit the growth of *N. dimidiatum in vitro*. Among the five fungicides, medium containing a final concentration of 250 ppm of either Cidely® Top or Protifert® demonstrated more than 85% inhibition in growth of *N. dimidiatum*, indicating that both fungicides were considered the most efficient fungicides (Figure 5B). Although the growth inhibition rate (M%) of *N. dimidiatum* at 5 dpi reached to 77–79% after the application of Amistar Top® and Airone Liquido® fungicides, Uniform® showed the lowest zone of inhibition (22%). This suggests that the latter fungicide is the least efficient; and therefore it is eliminated from further experiments.

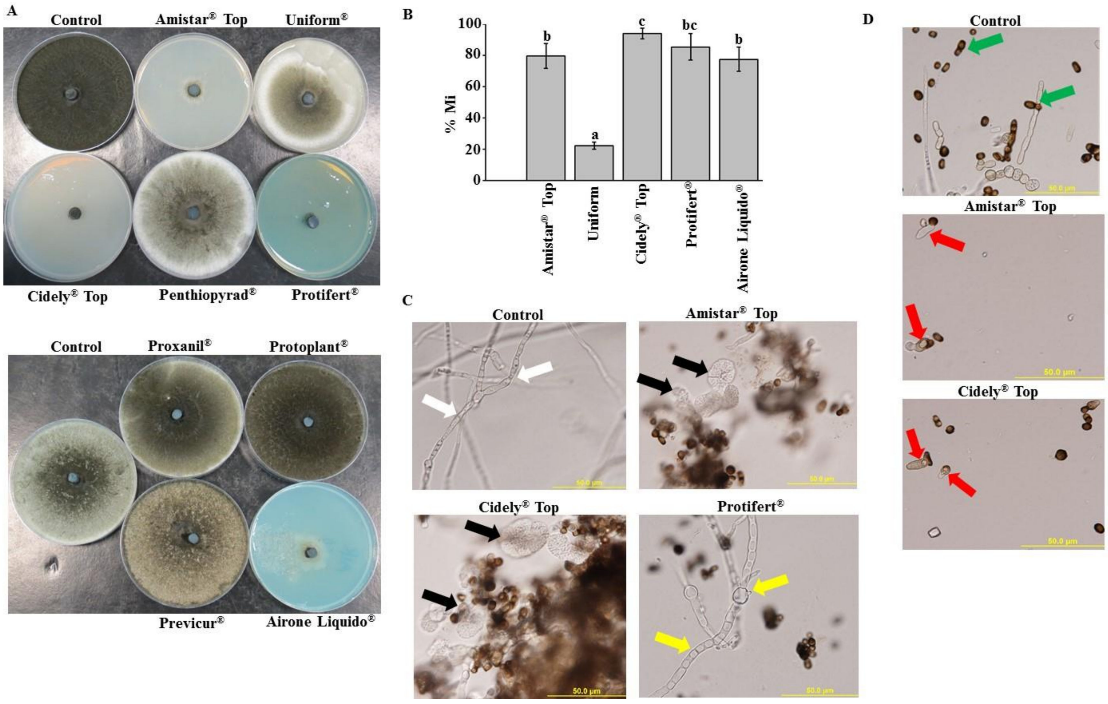


Figure 5. In vitro efficacy of fungicides against *Neoscytalidium dimidiatum*. (**A**) Effect of the fungicides Amistar Top®, Uniform®, Cidely® Top, Penthiopyrad®, Protifert® (top panel), Proxanil®, Proplant®, Previcur® and Airone Liquido® (bottom panel) at the concentration of 250 ppm on in vitro mycelial growth; and (**B**) growth inhibition rate (% Mi) of *N. dimidiatum* using 250 ppm of the fungicides after 5 days. (**C**) Abnormalities in hyphal morphology, septum formation and cytoplasmic contents; and (**D**) deformation of conidia of *N. dimidiatum* following Amistar Top® and Cidely® Top treatments compared to control. In (**B**)**,** values with different letters are significantly different from each other at $p < 0.05$; In (**C**)**,** white arrows indicate normal septate hyphal growth; black arrows indicate formation of non-septate hyphal formation and cytoplasmic coagulation; yellow arrows indicate lysis of hyphal wall and cytoplasm leakage. In (**D**), green arrows indicate normal formation of conidia and arthroconidia segmenting from hyphae; and red arrows indicate deformation of conidia and absence of arthroconidia.

We also examined the fungal pathogen microscopically in order to figure out the mode of action of the effective fungicides against *N. dimidiatum*. Results revealed that three fungicides caused significant alternations in the fungal morphology. In comparison to control treatment without any fungicide, application of either Amistar Top® or Cidely® Top at 250 ppm concentration to cultures led to lysis in hyphal wall and leakage in cytoplasm of *N. dimidiatum* (Figure 5C). We also noticed that Airone Liquido® caused not only unusual morphological abnormalities in cultures, but also septal defects and cytoplasmic deformations in hyphal cells. Surprisingly, we observed normal, septate hyphal morphology in cultures containing Protifert® similar to those in control treatment.

N. dimidiatum produced not only reduced numbers of deformed conidia, but also absences of arthroconidia in Amistar Top®- or Cidely® Top-treated cultures (Figure 5D). Similar to control, cultures of Airone Liquido and Protifert® showed normal conidial formation and well-formed arthroconidial segmentation produced by hyphae of *N. dimidiatum*. Altogether, the chemicals, Amistar Top®, Cidely® Top and Airone Liquido, had a direct effects on *N. dimidiatum* DSM 109897 through the inhibition of mycelial growth and induction of morphological abnormalities; thus, the former two fungicides shared a common mechanism of action. The mode of action of Protifert® in competently inhibiting the mycelial growth of *N. dimidiatum* was not determined. Because there are many reports in which chemical control against plant pathogens has proven successful only under laboratory conditions, more reliable in vivo studies are needed for the reproducibility of the results obtained from those of in vitro testing.

2.5. Assessment of Chemical Fungicides on Neoscytalidium Dimidiatum Using Apple Bioassay

To evaluate the most effective fungicides against *N. dimidiatum*, we developed the apple fruit bioassay method (Figure 6A). Placing the pathogen alone on apple fruits resulted in relatively large-sized, brown-colored lesions with distinct edges (Figure 6B). In contrast, none of the fungicides tested had negative effects against the pathogen. Excluding Airone Liquido®, when a plug containing any of the three fungicides paired with a plug of *N. dimidiatum* on the surface of the fruit, the particular fungicide completely suppressed the pathogen and no lesions were formed compared to the pathogen treatment alone (Figure 6B).

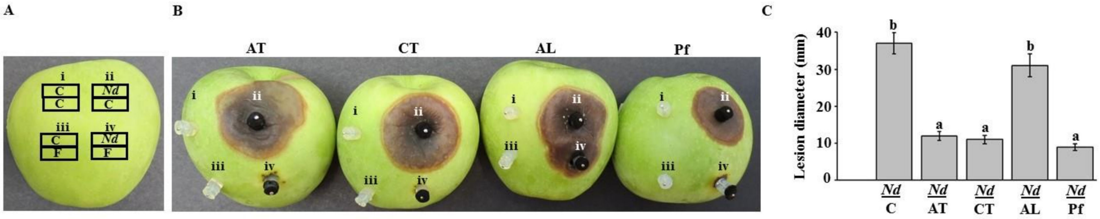


Figure 6. In vivo inhibitory effect of the chemical fungicides against *Neoscytalidium dimidiatum* using the "apple fruit bioassay". An illustration showing (**A**) inoculated-apple fruit with the chemical fungicides and/or *N. dimidiatum* agar plugs in combinations; (**B**) apple fruit bioassays using chemical fungicides; and (**C**) lesion diameter of *N. dimidiatum* using 250 ppm of the fungicides after 10 dpi. In (**A–B**), (i) two sterile non-inoculated PDA agar plugs; (ii) *N. dimidiatum* inoculum alone with a sterile agar plug below it; (iii) the fungicide (F) alone with a sterile agar plug above it; and (iv) pairing *N. dimidiatum* and the fungicide together, with the fungicide on the apple surface and *N. dimidiatum*-inoculated plug on top of the fungicide. In (**C**), values with different letters are significantly different from each other at $p < 0.05$. C, control (no *N. dimidiatum*); *Nd*, *N. dimidiatum;* F, fungicide; AT, Amistar Top®, CT, Cidely® Top; AL, Airone Liquido®; Pf, Protifert®; dpi, days post inoculation.

The fungicides Amistar Top®, Cidely® Top and Protifert® caused significantly ($p < 0.05$) smaller lesion sizes than the positive control (*N. dimidiatum*) treatment (Figure 6C). However, we did not notice any significant ($p < 0.05$) difference between the treatments of Airone Liquido® and the pathogen alone. Therefore, Airone Liquido® was excluded from further experiments. To greater extent, three chemical fungicides completely prevented lesion development on apple fruits. Overall, the novel apple fruit bioassay led to the selection of three prominent fungicides, Amistar Top®, Cidely® Top and Protifert®, which could have the potential to manage stem canker disease on royal poinciana seedlings.

2.6. Fungicide Effects on Royal Poinciana Infected with Neoscytalidium Dimidiatum

In the greenhouse experiment, we tested the efficacy of the most promising fungicides at 4 weeks post treatment (wpt) on *N. dimidiatum*-inoculated royal poinciana plants. Seedlings were artificially inoculated with the fungal pathogen for 2 weeks when symptoms of stem canker disease were easily recognized (Figure S2). Diseased plants were treated with a particular fungicide and this treatment was considered as 0 wpt. Disease progress or plant recovery of fungicide-treated plants was monitored until the end of the evaluation period of 4 wpt. In general, *N. dimidiatum*-inoculated plants that were sprayed with water only showed stem canker disease symptoms such as drying branches, falling leaves and discoloring stems, resulting in almost completely bare seedlings (Figure 7A). This was also clear in the longitudinal wood necrosis in these diseased plants (Figure 7B). In contrast, inoculated plants that were treated with Amistar Top®, Cidely® Top or Protifert® fungicide clearly showed vegetative growth recovery (Figure 7A) and developed relatively healthy wood (Figure 7B) at 4 wpt comparable to the negative control plants (no prior artificial infection). Affected plants treated with Airone Liquido® showed similar disease symptoms as diseased plants (Figure 7A). Peeling away the periderm of the inoculated plants that were treated with Airone Liquido® revealed the presence of a black layer of fungal growth from which *N. dimidiatum* was reisolated (Figure 7B).

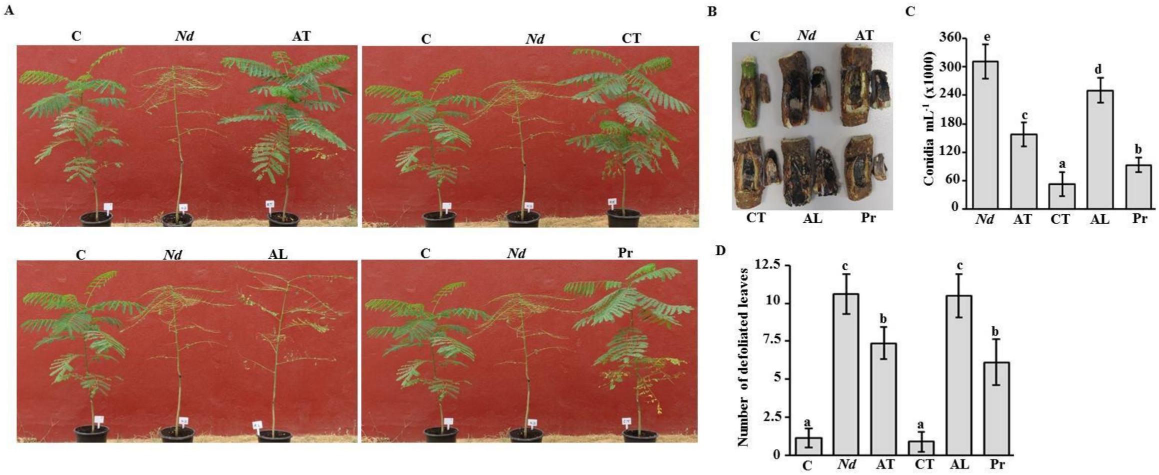


Figure 7. Effect of fungicide treatments on artificially inoculated royal poinciana seedlings with *Neoscytalidium dimidiatum* in the greenhouse. Fungicidal suppression of stem canker disease on royal poinciana seedlings using (**A**) potential chemical fungicides; (**B**) symptoms of inoculated regions; (**C**) number of conidia after recovery of the pathogen from stem tissues; and (**D**) number of defoliated leaves in inoculated seedlings sprayed with chemical fungicides at 4 wpt. In (**A–D**), seedlings were inoculated for 2 weeks with *N. dimidiatum* before the fungicide treatment. In (**C & D**), mean values with different letters are significantly different from each other at $p < 0.05$. Experiments were repeated at least three times with similar results. C, control (no *N. dimidiatum*); *Nd*, *N. dimidiatum*; AT, Amistar Top®: CT, Cidely® Top; AL, Airone Liquido®, Pr, Protifert®; wpt, weeks post treatment.

The effects of each of the chemical fungicides were also determined according to the number of conidia progressing on diseased- and treated-seedlings. In general, there was a significant ($p < 0.05$) difference between all treatments (Figure 7C). This was accompanied with a dramatic decrease in the number of conidia in Cidely® Top-treated seedlings that nearly reached to 6-fold reduction compared to that of untreated plants. We noticed that the number of conidia of *N. dimidiatum* recovered from the stems of royal poinciana treated with Protifert® and Amistar Top® fungicides was 3.3- and 2-fold less than in the control, respectively (Figure 7C). Airone Liquido® was marked the least spore counts; and thus it was considered the least effective among all tested fungicides.

The number of defoliated leaves was also assessed on diseased- and recovered-seedlings as an indication on the severity of disease symptoms on seedlings at 4 wpt. Based on our results, Cidely® Top treatment was comparable to the treatment without inoculation (Figure 7D). This was evident by the similar number of defoliated leaves per plant. On the other hand, the same plants showed significantly ($p < 0.05$) less falling leaves than inoculated-seedlings without fungicide treatment at the same period of evaluation. At 4 wpt, defoliated leaves demonstrated 31–42% reduction on seedlings sprayed with Amistar Top® and Protifert®, respectively, in comparison to *N. dimidiatum*-inoculated seedlings without any fungicide treatment (Figure 7D). It was also clear that Airone Liquido® was not efficient enough, confirming our previous results on the number of conidia recovered from inoculated seedlings using the same fungicide. Our data imply that Cidely® Top seems to be the most effective fungicide because the severity of stem canker disease is gradually suppressed and the pathogen is more or less restrained.

2.7. Effect of Cidely®Top on Royal Poinciana Trees Naturally Infected with Neoscytalidium Dimidiatum

We confirmed the results obtained from the in vitro and in vivo experiments by applying the promising fungicide Cidely® Top on royal poinciana trees naturally affected by stem canker under field conditions. Royal poinciana trees were sprayed with 250 ppm of Cidely® Top, and severity of symptoms or recovery of the trees was monitored for 32 weeks. Typical disease symptoms were observed on the day of fungicidal application (0 wpt; Figure 8A). After 16 weeks of spraying with

Cidely® Top, disease severity was remarkably decreased in the treated trees (Figure 8B). This was evident by diminishing trunk damage and developing new fresh shoots. It was also noted that trees treated with Cidely® Top fungicide increased their vegetative growth and were completely recovered at 32 wpt (Figure 8C). This suggests that the application with Cidely® Top results in disappearance of disease symptoms, ultimately leading to nice looking, healthy trees.

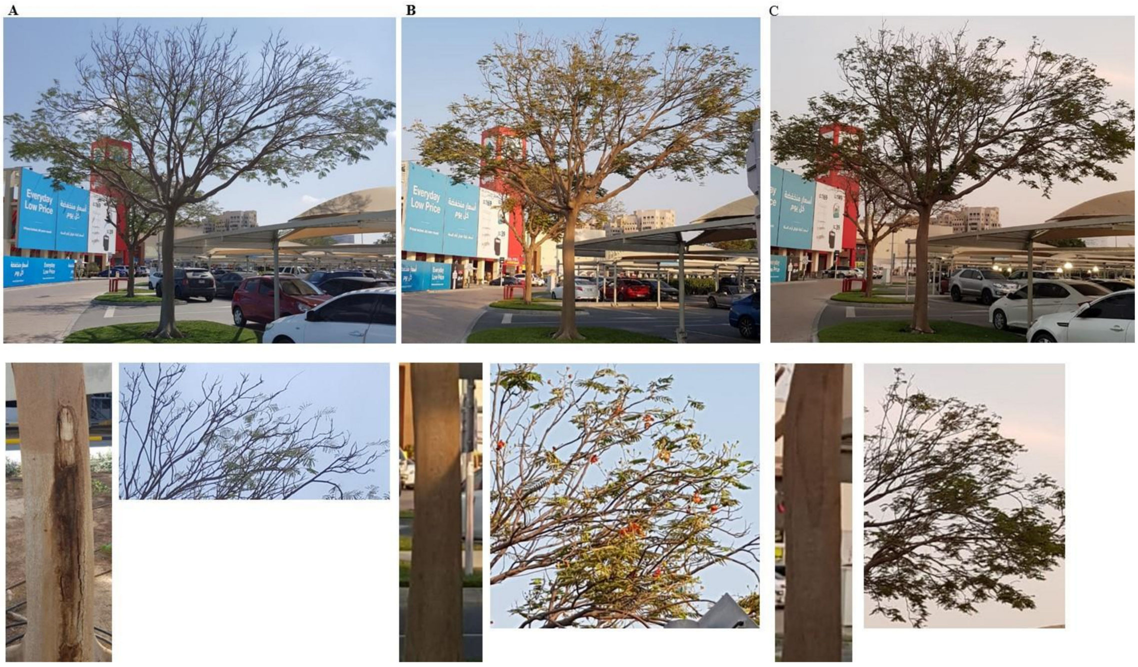


Figure 8. Effect of Cidely® Top on royal poinciana trees naturally infected with *N. dimidiatum* in the field in DFC, UAE. Fungicidal suppression of stem canker disease symptoms on (**A**) royal poinciana trees (n = 6); followed by treatment with the fungicide Cidely® Top at (**A**) 0 (**B**) 16 and (**C**) 32 weeks post treatment. In (**A–C**), photos showed the severe disease symptoms and the recovery of the same whole tree (upper panel), trunk (left, bottom panel) and branches (right, bottom panel).

3. Discussion

Royal poinciana is a large deciduous tree species prevalent in subtropical and tropical areas of the world. It is valued as a local street tree and is widely planted in open areas [1]. In the last decade, this beautiful flowering plant has become widespread in urban and agricultural areas of the UAE. Although it is known for its ability to withstand severe conditions, diseases are major factors that affect the health of royal poinciana [3,4]. Many of the phytopathogens can cause diseases on host plants, including royal poinciana [5–8]. Therefore, careful attention should be attained to the causal agent of stem canker disease on royal poinciana, taking into account the frequency of disease incidence, the geographical distribution and the environmental conditions favorable to the disease occurrence.

In our efforts to identify the pathogen linked with the diseased trees, we first detected the symptoms of stem canker on royal poinciana. In general, we noticed dieback, canker and gummosis, which ultimately led to complete dryness and death of royal poinciana trees (Figure 1). Although some studies have reported several fungi to cause cankers on plant species [5–8], others have recorded *N. dimidiatum* on almond, dragon fruit, eucalyptus, fig and plum, displaying disease symptoms of canker and dieback in different places of the world [13–15,28]. In general, environmental stress has negative impact on the severity of disease, depending on the level and duration of the stress, and the sensitivity and developmental stage of the plant species. In hot summers, sooty canker invades trees and ornamentals of mulberry, ash, walnut, fig, sycamore, apple, apricot, poplar, eucalyptus and olive in Iraq [29,30]. In Oman, significant damage due to dieback, witling and death of royal poinciana has been reported to be caused by *N. dimidiatum* and symptoms are even worsened when trees are exposed to heat (up to 45 °C) and shortage of water [16]. All previously mentioned reports are in agreement with the findings of the current study. Yet, there are no reports about the causal agent of the disease symptoms of stem canker on royal poinciana or any other ornamental woody tree in the UAE. Previously, the fungal pathogens *T. punctulata* and *F. solani* have been shown to cause black scorch disease and SDS on date palm, respectively [17,19,31] and *L. theobromae* to cause dieback disease on mango [18]. Therefore, accurate fungal identification was carried out, along with proper chemical fungicide treatment to manage the devastating damage of this disease on royal poinciana.

The fungal pathogen was constantly isolated from all symptomatic tissues examined from trees of royal poinciana, and it was characterized based on its morphology, phylogeny and pathogenicity assays. On PDA, a rapid growth of mycelia filling the entire plate was observed within 8 days. The culture was effuse, hairy to wooly, started as white with creamy, ochraceous-yellowish color that turned to dark greyish or blackish color by day 12. Similar observations have been previously reported on *N. dimidiatum* isolated from diseased trees of eucalyptus [23]. Microscopic examination of the pathogen demonstrated branched and septate hyphae with no conidiophores. Consistent with [21], arthroconidia were thick-walled and barrel-shaped that could be found individually or in chains, ranging 5–15 x 3–6 μm in size (Figure 2). Old cultures, of 25 days, developed hyaline pycnidial conidia when young, and dark brown central regions when aged. Cultures also produced fusicoccum-like conidia in pycnidia (Figure 2) [32]. Because *Neoscytalidium* spp. are very close and difficult to discriminate, molecular characterization was followed to avoid misleading conclusions about the pathogen. For that reason, phylogenetic analysis using *TEF1-α* sequence (MN447201) was generated and proved the identity of the fungus as *N. dimidiatum*. *N. dimidiatum* was closely related to both *N. novaehollandiae* [33] and *N. hyalinum* [34], confirming previous findings [13,35]. Our data indicated that the isolate of *N. dimidiatum* in the current study was morphologically and genetically similar to other isolates of *N. dimidiatum* from *Juglan regia*, *Prunus* sp., mango and others. Therefore, isolate DSM 109897 in the present study belonged to *N. dimidiatum* and was the main causal agent of stem canker on royal poinciana in the UAE. Our observations on the symptoms and the pathogen associated with stem canker disease on royal poinciana are similar to a previous report on the same tree in Oman [16]. This suggests that *N. dimidiatum* may possibly have been introduced from this neighboring country to the UAE.

The existence of the pathogen and the progression of the disease in tissues of the whole royal poinciana seedlings and apple fruits were further verified via pathogenicity tests. The results obtained from the greenhouse experiment on young healthy plants after inoculation were similar to the disease symptoms on trees of royal poinciana located in the field, and that was confirmed by Koch's postulates when *N. dimidiatum* was frequently recovered from the inoculated seedlings. Our data match those in other trials using artificial inoculation of the same pathogen on royal poinciana [16] or other plant species [13–15,20,28]. Pathogenicity assays on seedlings of royal poinciana (Figure 4), *F. benjamina* and *F. nitida* [13] and eucalyptus, poplar and olive [30] clearly described that discoloration of vascular tissues, and drying and defoliation of leaves, were symptoms associated with stem canker caused by *N. dimidiatum*. There has been a rise in reports about *N. dimidiatum* causing diseases on fruits of pitahaya, plum and almond [12,14,15]. Apple fruit bioassays have been conducted to determine the effects of the fungal pathogen associated with canker diseases [5,36]. Therefore, we performed pathogenicity tests on healthy apple fruits and monitored the disease progress.

There are some examples of using BCAs effective against *N. dimidiatum* or other pathogens [37–39]; yet these studies have not been assessed *in vivo*. For example, *Trichoderma harzianum* T3.13 revealed in vitro antagonistic activities to *N. dimidiatum* [39]. Although chemical fungicides have adverse effect on human health, food and environment [23,40], these agents are commonly used due to their relatively low cost, rapid acting, long lasting, high stability and ease of application [41]. Under laboratory conditions, four of the tested chemicals, Amistar Top®, Cidely® Top, Protifert® and Airone Liquido®, showed suppression in the growth of *N. dimidiatum*. This was evidenced by the abnormalities seen in hyphal morphology, septal formation, cytoplasmic contents and the deformation of conidia following fungicide treatments (Figure 5). Previously, Cidely® Top exhibited the strongest inhibition of mycelial growth of *T. punctulata* and *L. theobromae* in petri dish experiments [17,42]. The same fungicides were further evaluated in vivo using apple fruit bioassay (Figure 6). In general, Amistar Top®, Cidely® Top and Protifert®, significantly reduced the lesion size on apple fruits when 250 ppm of the fungicide was applied concurrently with the pathogen. On contrast, Airone Liquido® was not effective against this pathogen on apple and was carried out in further experiments as a negative control. We claim that the novel apple bioassay is a small-scale reference of what may occur in the greenhouse/field. In vivo experiments using carrot roots and mango fruits have previously been implemented to assess growth retardation of *Pythium coloratum* and L. theobromae by BCAs, respectively [22,43].

Recent reports have shown that in vitro tests along with greenhouse experiments are essential to determine the sensitivity of plant pathogens to chemical and/or biological treatments [17–19,22,42]. According to our greenhouse experiments, Cidely® Top, followed by Protifert® and then Airone Liquido® were effective on diseased seedlings of royal poinciana. It is known that the organic foliar fertilizer, Protifert®, is a good source of minerals, essential traces, amino acids and peptides necessary for plant growth and development. In this study, we also showed that Protifert® not only provided vigorous and healthy seedlings, but also it served as a protection to trees from fungal infections i.e., *N. dimidiatum*. Under greenhouse conditions, we noticed that the most significant reduction in disease symptoms of stem canker was found in Cidely® Top-treated seedlings of royal poinciana at 4 wpt. This was clear in seedlings sprayed with Cidely® Top possessing the lowest conidial counts and the least number of defoliated leaves, indicating that this fungicide could be a potent fungicide for the management of *N. dimidiatum* affecting royal poinciana trees. The result of Cidely® Top is in agreement with previous studies indicating high effectiveness of this fungicide against a number of fungal pathogens attacking trees such as *T. punctulata*, *L. theobromae* and *F. solani* that were almost completely inhibited [19,22,42]. To a lesser extent, Amistar Top® was not as effective as Protifert® or Cidely® Top in reducing the pathogenic activities of *N. dimidiatum* in greenhouse trials. Eventhough Amistar Top® and Cidely® Top were difenoconazole-based fungicides sharing the same concentration of the active ingredient; the superior efficiency of Cidely® Top over Amistar Top® could be attributed to the presence of cyflufenamid as an additional active ingredient leading to increased inhibition levels of *N. dimidiatum*. Difenoconazole was ineffective against *Fusarium magniferae* [44], but it was significantly capable for managing other diseases [17,18,42,45,46], including stem canker on royal poinciana in the current study (Figure 7). This can be disputed to the fungicide application methods, active ingredient concentrations, plant growth conditions or pathogen responses. Airone Liquido® (metal copper), on the other hand, is not recommended to manage the disease.

So far, there are no reports to evaluate Cidely® Top or any systemic fungicide on royal poinciana trees infected with *N. dimidiatum* under field conditions. Thus, the same fungicide was found to be highly effective against plant pathogenic fungi on date palm and mango [18,19,42]. Accordingly, a field experiment was carried out to assess the efficacy of Cidely® Top on naturally infested royal poinciana plants. Apparently, the entire trees showed “more or less” full recovery that was mainly observed in newly developed inflorescences (branches with flower clusters) and reduced disease symptoms on trunks of royal poinciana trees sprayed with Cidely® Top at 16 and 32 wpt (Figure 8). This suggests that Cidely® Top can possibly serve as a competent element of IDM of stem canker on royal poinciana. Here, we report the symptoms, the pathogen as well as the proper chemical treatment to manage stem canker as the first step toward planning IDM programs against this devastating disease on royal poinciana in the UAE or elsewhere. In the current study, the phenotype *i.e.*, symptoms associated with the disease can be considered as a starting point for future comparative ‘omic’ analyses including genomes and responses to environmental variation [47]. A combination of different methods to achieve suitable IDM practices is on top of our priorities. Investigations for cultural (pruning), chemical (Cidely® Top and Protifert®) and BCAs as IDM to manage stem canker on royal poinciana are in progress for environmental sustainability.

4. Materials and Methods

4.1. Fungal Culture and Isolation

Eight-year-old royal poinciana trees located in DFC, Dubai, UAE (latitude/longitude: 25.22/55.36) were associated with longitudinal cankers on stems (Figure 1). Cross-sections in trunks and branches were made and drying leaves were gathered from diseased trees. All collected tissues were then transferred to the Plant Microbiology Laboratory, Department of Biology, United Arab Emirates University in Al Ain city, UAE, for isolation and identification purposes. To isolate the pathogen, affected tissues were cut into small pieces (3–5 mm long), washed and surface-sterilized with mercuric chloride 0.1% for 1 min, and 1.05% NaOCl for 5 min; followed by three consecutive washings in sterile distilled water. They were then transferred onto PDA (Lab M Limited, Lancashire, UK) plates, supplemented with 25 mg/L penicillin-streptomycin (Sigma-Aldrich Chemie GmbH, Taufkirchen, Germany) to inhibit bacterial contaminants. Petri dishes were incubated for 5 days at 25 ± 2 °C. Once grown out of the plated tissue, mycelia were aseptically sub-cultured on fresh PDA and purified using hyphal-tip isolation technique [48]. To characterize fungal structures, mycelia and conidia were observed using Nikon-Eclipse 50i light microscope (Nikon Instruments Inc., Melville, NY, USA). The culture of the identified fungus, *N. dimidiatum* [27], was deposited in Leibniz-Institute DSMZ-German Collection of Microorganisms and Cell Cultures GmbH (Braunschweig, Germany) under the accession number 109897.

4.2. Molecular Identification of the Pathogen

DNA of the pathogen isolated from diseased of stem, branch and leaf tissues was extracted from mycelia cultured for 10 d at 25 °C on PDA plates, using the fungi DNA isolation kit (Norgen Biotek Corp., Thorold, ON, Canada). PCR was set up to amplify target regions of internal transcribed spacer (*ITS*) of the nuclear rDNA for *N. dimidiatum* using ITS1 and ITS4 primers [26], partial *28S rDNA* using LR0R and LR5 primers [49], partial *TEF1-α* using EF1-728F and EF1-986R [50] and partial *β-tubulin* using Bt1a and Bt1b [51]. PCR reactions (50 μL) contained 30-ng DNA template, 50 pmol of each primer, 200 μM of each dNTP, 2.5 unit of Taq DNA polymerase and 2.2 mM buffer ($MgCl_2$). Each cycle of PCR was set as the following: 94 °C for 1 min; 58 °C for 1 min; and 72 °C for 1 min (total of 32 cycles). All primer sequences can be found in Table S1. All protocols for amplification and sequencing were as described [26].

The sequence of *TEF1-α* gene of the fungal isolate from the UAE was deposited in GenBank (accession number: MN447201). The phylogenetic tree using *TEF1-α* sequence, obtained from DSMZ, was constructed against other sequences of *TEF1-α* belonging to *Neoscytalidium* spp. [27] retrieved from GenBank-NCBI (www.ncbi.nlm.nih.gov). ML analysis was performed for the estimation of the phylogenetic tree [52] after all sequences were aligned. Phylogenetic trees were validated with a statistical support of the branches with 100 bootstrap resamples. The following isolates used in the analysis belong to *N. dimidiatum*, *N. novaehollandiae*, *N. hyalinum*, *Botryosphaeria dothidea* and *B. fusispora*.

4.3. In Vivo Pathogenicity Tests and Koch's Postulates

Pathogenicity tests were conducted on one-year-old healthy royal poinciana seedlings ($n = 9$), purchased from the local market. Using sterile scalpels, the bark of the main stem was wounded and inoculations under the wounded bark were performed at 30–50 cm above the soil surface [13]. An agar plug (8-mm-diameter) colonized by mycelium of 10-day-old culture of *N. dimidiatum* was placed into the wound, where the mycelium facing inner parts, and wrapped using parafilm. Control royal poinciana seedlings were inoculated with sterile agar plugs (no pathogen). Plants were maintained in the greenhouse (15 h day/9 h night at 25 ± 2 °C) and were evaluated for symptoms and disease progression at 2 and 5 wpi. By the end of the experiment, the fungus was re-isolated from the point of infection on PDA and compared morphologically with the inoculated fungus.

Disease was assayed on disease-free apple fruits (cv Granny Smith), purchased from local fresh markets, to find out the effect of *N. dimidiatum*. Fruits (n = 8) were washed with sterile distilled water, surface-sterilized with 70% ethanol and wounded with a sterilized scalpel (2 mm diameter) according to [36] with some modifications. On each fruit, one agar plug (11 mm in diameter) containing mycelium of *N. dimidiatum* (colonized mycelium facing down) and one agar control plug without pathogen was applied. Inoculated fruits were maintained in dark (at 25 ± 2 °C and 80% relative humidity) and lesion size was rated for an interval of 5 d for 10 d. At 10 dpi, pieces from regions showing disease symptoms of inoculated fruit tissues were removed, surface sterilized, plated and incubated, as mentioned above. Structures of conidia and mycelium were morphologically compared with the inoculated fungus.

4.4. In Vitro Evaluation of Fungicides Against N. Dimidiatum

The fungicide experiment was carried out according to the previously described procedures [17–19]. The selected fungicides along with their active ingredients can be found in Table S2. Fungal growth was assessed on each fungicide with a final concentration of 0 (control), 250, 500, 750 and 1000 ppm aseptically introduced into sterilized PDA plates, supplied with penicillin-streptomycin antibiotics, at 25 ± 2 °C. The tested fungal pathogen was introduced to PDA plates using a sterile cork-borer (8 mm diameter). Cultures were incubated at 25 ± 2 °C for 10 days, and percentage of the mycelial growth inhibition was measured according to:

$$\%\ \mathrm{Mi} = (\mathrm{Mc}\ \text{-}\ \mathrm{Mt})/\mathrm{Mc} \times 100\% \qquad (1)$$

where Mi, inhibition of the mycelial growth; Mc, colony diameter (in mm) of control set; and Mt; colony diameter (in mm) of the target fungus on the medium with fungicide.

4.5. In Vivo Evaluation of Selected Fungicides

To determine the ability of fungicides to reduce lesion formation after *N. dimidiatum* inoculation under laboratory conditions, an apple fruit bioassay was developed. The apple fruit bioassay was modified according to previous bioassays on carrot and mango against *Pythium coloratum* and *L. theobromae*, respectively [18,22,43]. Healthy apple fruits (cv. Granny Smith) were washed with sterile distilled water, surface-sterilized with 70% ethanol and placed in plastic trays on wet, sterile filter papers. Apple fruits were then inoculated using agar plugs (11 mm) colonized by the selected fungicide and/or *N. dimidiatum*, as described above, onto each apple fruit according to the following combinations: (i) two sterile non-inoculated PDA agar plug (control; C); (ii) *N. dimidiatum* alone with a sterile PDA agar plug below it; (iii) the fungicide alone with a sterile PDA agar plug above it; and (iv) pairing *N. dimidiatum* and the fungicide together (the fungicide on the apple surface and *N. dimidiatum*-inoculated plug on top of the fungicide). All fungicides were introduced onto the apple surface 24 h before inoculation with the pathogen to have enough time for the active ingredients to disperse uniformly onto the apple surface. Each apple fruit was inoculated with the four combinations for each fungicide of five fruits/tray and was replicated three times. Trays were covered with aluminum foil and incubated in dark (at 25 ± 2 °C and 80% relative humidity) for 10 d. Lesion diameters were measured (in mm) and averaged.

In a greenhouse experiment, we assessed the impact of each fungicide on one-year-old royal poinciana seedlings. Seedlings were wounded and inoculated with agar plugs containing mycelium of *N. dimidiatum* in the stem of each plant as described above. Inoculated plants were maintained in the greenhouse at 25 °C until symptoms were evident. At 2 wpi, seedlings were either sprayed with 250 ppm fungicide or water (control); and these treatments were designated as 0 wpt. Symptoms on inoculated plants, conidia counts of the fungal pathogen and the number of falling leaves were recorded at 4 wpt [42]. The procedure of conidia counts involved homogenized weight of affected tissues placed in 5 mL of water, and the suspended material was assessed to estimate the number of conidia using haemocytometer (Agar Scientific Limited, Essex, UK).

Regarding the field experiments, trees were located in the same place described above. Cidely® Top (Syngenta International AG, Basel, Switzerland) was the only tested fungicide on six royal poinciana trees (8 years old). Each *N. dimidiatum* naturally infested tree was chosen so as to be surrounded by untreated corresponding trees to serve as a reservoir for recontamination. Trees were pruned and completely sprayed/treated with the recommended dose of the fungicide (250 ppm). Experiments were repeated twice in February 2018 and February 2019 with similar results.

4.6. Statistical Analysis

For the pathogenicity assays, fruits (n = 5) and seedlings (n = 9) for each treatment were used. For the in vitro evaluation of fungicides against *N. dimidiatum*, 6 plates for each treatment were used. For the fungal conidia counts and the number of falling leaves in the in vivo evaluation of fungicides under greenhouse conditions, a minimum of 4 plants for each treatment was used. Data represent the mean ± SD. Analysis of Variance (ANOVA) and Duncan's multiple range test were performed to determine the statistical significance at $p < 0.05$. All experiments were independently repeated three times with similar results. All statistical analyses were performed by using SAS Software version 9 (SAS Institute Inc., Cary, NC, USA).

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/21/3/1033/s1.

Author Contributions: K.A.E.-T. and S.F.A. conceived and designed the experiments; S.M.A.R., E.E.S., A.S. and K.A. performed the experiments; S.M.A.R., K.A.E.-T. and S.F.A. analyzed the data; K.A.E.-T. and S.F.A. contributed reagents/materials/analysis tools; K.A.E.-T. and S.F.A. wrote the paper. All authors have read and approved the manuscript.

Funding: This research was funded by Khalifa Center for Biotechnology and Genetic Engineering-UAEU, grant number 31R081) to S.F.AQ.

Acknowledgments: We'd like to thank Yousef A. BinZayed and Ahmed Basher from Al-Futtaim Group Real Estate for the technical assistance in the field.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gledhill, D. *The Names of Plants*, 4th ed.; Cambridge University Press: Cambridge, UK, 2008; p. 137.
2. Kirtikar, K.R.; Basu, B.D. *Indian medicinal plants*, 2nd ed.; International Book Distributors: Dehradun, India, 1999; Volume 2, p. 852.
3. Gilman, E.F.; Watson, D.G.; Klein, R.W.; Koeser, A.K.; Hilbert, D.R.; McLean, D.C. *Delonix regia: Royal Poinciana*; The Institute of Food and Agricultural Sciences (IFAS), University of Florida: Gainesville, FL, USA, 2019; Available online: edis.ifas.ufl.edu/pdffiles/ST/ST22800.pdf (accessed on 3 January 2020).
4. Cayley, D.M. Fungi associated with "die back" in stone fruit trees. *Ann. Appl. Biol.* **1923**, *10*, 253–275. [CrossRef]
5. Anagnostakis, S.L.; Ferrandino, F.J. Isolation of *Nectria galligena* from cankers on sweet birch. *Plant Dis.* **1998**, *82*, 440–441. [CrossRef]
6. Rouxel, T.; Balesdent, M.H. The stem canker (blackleg) fungus, *Leptosphaeria maculans*, enters the genomic era. *Mol. Plant Pathol.* **2005**, *6*, 225–241. [CrossRef]
7. Chungu, D.; Muimba-Kankolongo, A.; Wingfield, M.J.; Roux, J. Identification of fungal pathogens occurring in eucalypt and pine plantations in Zambia by comparing DNA sequences. *Forestry* **2010**, *83*, 507–515. [CrossRef]
8. Darge, W.A. First report of *Lasioplodia theobromae* causing needle blight and stem canker diseases on *Araucaria heterophylla* in Ethiopia. *J. Hortic. Res.* **2017**, *25*, 15–18. [CrossRef]
9. Reckhaus, P. Hendersonula dieback of mango in Niger. *Plant Dis.* **1987**, *71*, 1045. [CrossRef]
10. Farr, D.F.; Bills, G.F.; Chamuris, G.P.; Rossman, A.Y. *Fungi on Plants and Plant Products in the United States*; The American Phytopathological Society (APS) Press: St. Paul, MN, USA, 1989; Volume 8, p. 1252.
11. Ray, J.D.; Burgess, T.; Lanoiselet, V.M. First record of *Neoscytalidium dimidiatum* and *N. novaehollandiae* on *Mangifera indica* and *N. dimidiatum* on *Ficus carica* in Australia. *Australas. Plant Dis. Notes* **2010**, *5*, 48–50. [CrossRef]
12. Yi, R.H.; Lin, Q.L.; Mo, J.J.; Wu, F.F.; Chen, J. Fruit internal brown rot caused by *Neoscytalidium dimidiatum* on pitahaya in Guangdong province, China. *Australas. Plant Dis. Notes* **2015**, *10*, 1–4. [CrossRef]

13. Al-Bedak, O.A.; Mohamed, R.A.; Seddek, N.H. First detection of *Neoscytalidium dimidiatum* associated with canker disease in Egyptian *Ficus* trees. *Forest Pathol.* **2017**, 1–7. [CrossRef]
14. Hajlaoui, M.R.; Nouri, M.T.; Hamrouni, N.; Trouillas, F.P.; Ben Yahmed, N.; Eddouzi, J.; Mnari-Hattab, M. First record ofdieback and decline of plum caused by *Neoscytalidium dimidiatum* in Tunisia. *New Dis. Rep.* **2018**, *38*, 20. [CrossRef]
15. Nouri, M.T.; Lawrence, D.P.; Yaghmour, M.A.; Michailides, T.J.; Trouillas, F.P. *Neoscytalidium dimidiatum* causing canker, shoot blight and fruit rot of almond in California. *Plant Dis.* **2018**, *102*, 1638–1647. [CrossRef] [PubMed]
16. Elshafie, A.E.; Ba-Omar, T. First report of *Albizia lebbeck* caused by *Scytalidium dimidiatum* in Oman. *Mycopathologia* **2002**, *154*, 37–40. [CrossRef] [PubMed]
17. Saeed, E.E.; Sham, A.; El-Tarabily, K.A.; Abu Elsamen, F.; Iratni, R.; AbuQamar, S.F. Chemical control of dieback disease on date palm caused by the fungal pathogen, *Thielaviopsis punctulata*, in United Arab Emirates. *Plant Dis.* **2016**, *100*, 2370–2376. [CrossRef] [PubMed]
18. Saeed, E.E.; Sham, A.; AbuZarqa, A.; Al Shurafa, K.; Al Naqbi, T.S.; Iratni, R.; El-Tarabily, K.A.; AbuQamar, S.F. Detection and management of mango dieback disease in the United Arab Emirates. *Int. J. Mol. Sci.* **2017**, *18*, 2086. [CrossRef] [PubMed]
19. Alwahshi, K.J.; Saeed, E.E.; Sham, A.; Alblooshi, A.A.; Alblooshi, M.M.; El-Tarabily, K.A.; AbuQamar, S.F. Molecular identification and disease management of date palm sudden decline syndrome in the United Arab Emirates. *Int. J. Mol. Sci.* **2019**, *20*, 923. [CrossRef]
20. Chuang, M.F.; Ni, H.F.; Yang, H.R.; Shu, S.L.; Lai, S.Y.; Jiang, Y.L. First report of stem canker disease of pitaya (*Hylocereus undatus* and *H. polyrhizus*) caused by *Neoscytalidium dimidiatum* in Taiwan. *Plant Dis.* **2012**, *96*, 906–907. [CrossRef]
21. Mohd, M.H.; Salleh, B.; Zakaria, L. Identification and molecular characterizations of *Neoscytalidium dimidiatum* causing stem canker of red-fleshed dragon fruit (*Hylocereus polyrhizus*) in Malaysia. *J. Phytopathol.* **2013**, *161*, 841–849. [CrossRef]
22. Kamil, F.H.; Saeed, E.E.; El-Tarabily, K.A.; AbuQamar, S.F. Biological control of mango dieback disease caused by *Lasiodiplodia theobromae* using streptomycete and non-streptomycete actinobacteria in the United Arab Emirates. *Front Microbiol.* **2018**, *9*, 829. [CrossRef]
23. Al-Tememe, Z.A.M.; Lahuf, A.; Abdalmoohsin, R.G.; Al-Amirry, A.T. Occurrence, identification, pathogenicity and control of *Neoscytalidium dimidiatum* fungus, the causal agent of sooty canker on *Eucalyptus camaldulensis* in Kerbala Province of Iraq. 2019. *Plant Arch.* **2019**, *19*, 31–38.
24. Murad, N.Y.; Al-Dabagh, M.N. Evaluation some of pesticides in control of *Neoscytalidium dimidiatum* (Penz) Crous and Slippers causing wilt and canker on cypress trees in Iraq. *Iraqi J. Agric. Sci.* **2014**, *19*, 25–38.
25. Farr, D.F.; Elliott, M.; Rossman, A.Y.; Edmonds, R.L. *Fusicoccum arbuti* sp. nov. causing cankers on pacific madrone in western North America with notes on *Fusicoccum dimidiatum*, the correct name for *Scytalidium dimidiatum* and *Nattrassia mangiferae*. *Mycologia* **2005**, *97*, 730–741. [CrossRef] [PubMed]
26. White, T.J.; Bruns, T.; Lee, S.; Taylor, J. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR Protoc.* **1990**, *91*, 315–322.
27. Crous, P.W.; Slippers, B.; Wingfield, M.J.; Rheeder, J.; Marasas, W.F.O.; Philips, A.J.L.; Alves, A.; Burgess, T.; Barber, P.; Groenewald, J.Z. Phylogenetic lineages in the Botryosphaeriaceae. *Stud. Mycol.* **2006**, *55*, 235–253. [CrossRef] [PubMed]
28. Du, B.D.; Ngoc, D.T.B.; Thang, N.D.; Tuan, L.N.A.; Thach, B.D.; Hien, N.Q. Synthesis and in vitro antifungal efficiency of alginate-stabilized Cu_2O-Cu nanoparticles against *Neoscytalidium dimidiatum* causing brown spot disease on dragon fruit plants (*Hylocereus undatus*). *Vietnam J. Chem.* **2019**, *57*, 318–323.
29. Hassan, W.A.; Pasha, A.A.; Mohammad, M.B. Sooty canker on some thin bark trees caused by *Nattrassia mangiferae*. *Egypt. J. Agric. Res.* **2009**, *87*, 443–456.
30. Hassan, W.A.; Haleem, R.A.; Hassan, P.H. Effect of heat-stress predisposition on the development of sooty canker caused by *Neoscytalidium dimidiatum* (Penz.) Crous & Slippers. *Acta Agrobot.* **2011**, *64*, 207–212.
31. Alhammadi, M.S.; Al-Shariqi, R.; Maharachchikumbura, S.; Al-Sadi, A.M. Molecular identification of fungal pathogens associated with date palm root diseases in the United Arab Emirates. *J. Plant Pathol.* **2018**, *99*, 1–7.
32. Pavlic, D.; Wingfield, M.J.; Barber, P.; Slippers, B.; Hardy, G.E.S.; Burgess, T.I. Seven new species of the Botryosphaeriaceae from baobab and other native trees in Western Australia. *Mycologia* **2008**, *100*, 851–866. [CrossRef]

33. Polizzi, G.; Aiello, D.; Vitale, A.; Giuffrida, F.; Groenewald, Z.; Crous, P.W. (2009) First report of shoot blight, canker, and gumumosis caused by *Neoscytalidium dimidiatum* on citrus in Italy. *Plant Dis.* **2009**, *93*, 1215. [CrossRef]
34. Madrid, H.; Ruı´z-Cendoya, M.; Cano, J.; Stchigel, A.; Orofino, R.; Guarro, J. Genotyping and in vitro antifungal susceptibility of *Neoscytalidium dimidiatum* isolates from different origins. *Int. J. Antimicrob. Agents* **2009**, *34*, 351–354. [CrossRef]
35. Alwan, S.L.; Hussein, H.N. Efficacy of ecofriendly biocontrol *Azotobacter chroococcum* and *Lactobacillus rhamnosus* for enhancing plant growth and reducing infection by *Neoscytalidium* spp. in fig (*Ficus carica* L.) saplings. *J. Kerbala Agric. Sci.* **2019**, *6*, 16–25.
36. Hortova, B.; Novotny, D.; Erban, T. Physiological characteristics and pathogenicity of eight *Neofabraea* isolates from apples in Czechia. *Europ. J. Hort. Sci.* **2014**, *79*, 327–334.
37. AbuQamar, S.; Moustafa, K.; Tran, L.-S. Mechanisms and strategies of plant defense against *Botrytis cinerea*. *Crit. Rev. Biotechnol.* **2017**, *37*, 262–274. [CrossRef] [PubMed]
38. Mengiste, T.; Laluk, K.; AbuQamar, S. Mechanisms of induced resistance against B. cinerea. In *Post-harvest Pathology*; Chapter 2; Prusky, D., Gullino, M.L., Eds.; Springer Science + Business Media: Berlin, Germany, 2010; Volume 2, pp. 13–30.
39. Rusmarini, W.; Shah, U.K.D.; Abdullah, M.P.; Mamat, S.; Hun, T.G. Identification of *Trichoderma harzianum* T3.13 and its interaction with *Neoscytalidium dimidiatum* U1, a pathogenic fungus islated from dragon fruit (*Hylocereus polyrhizus*) in Malaysia. *Int. J. Environ. Agric. Res.* **2017**, *3*, 3205–3228.
40. Budzinski, H.; Couderchet, M. Environmental and human health issues related to pesticides: From usage and environmental fate to impact. *Environ. Sci. Pollut. Res.* **2018**, *25*, 14277. [CrossRef]
41. Kuai, X.; Barraco, C.; Després, C. Combining fungicides and prospective NPR1-based "just-in-time" immunomodulating chemistries for crop protection. *Front. Plant Sci.* **2017**, *8*, 1715. [CrossRef]
42. Saeed, E.E.; Sham, A.; Salmin, Z.; Abdelmowla, Y.; Iratni, R.; El-Tarabily, K.A.; AbuQamar, S.F. *Streptomyces globosus* UAE1, a potential effective biocontrol agent for black scorch disease in date palm plantations. *Front. Microbiol.* **2017**, *8*, 1455. [CrossRef]
43. El-Tarabily, K.A.; Hardy, G.E.St.J.; Sivasithamparam, K.; Hussein, A.M.; Kurtböke, D.I. The potential for the biological control of cavity spot disease of carrots caused by *Pythium coloratum* by streptomycete and non-streptomycete actinomycetes in Western Australia. *New Phytol.* **1997**, *137*, 495–507. [CrossRef]
44. Iqbal, Z.; Pervez, M.A.; Ahmad, S.; Iftikhar, Y.; Yasin, M.; Nawaz, A.; Ghazanfar, M.U.; Dasti, A.A.; Saleem, A. Determination of minimum inhibitory concentrations of fungicides against fungus *Fusarium mangiferae*. *Pak. J. Bot.* **2010**, *42*, 3525–3532.
45. Khan, S.H.; Idrees, M.; Muhammad, F.; Mahmood, A.; Zaidi, S.H. Incidence of shisham (*Dalbergia sissoo* Roxb.) decline and in vitro response of isolated fungus spp. to various fungicides. *Int. J. Agric. Biol.* **2004**, *6*, 611–614.
46. Yanase, Y.; Katsuta, H.; Tomiya, K.; Enomoto, M.; Sakamoto, O. Development of a novel fungicide, penthiopyrad. *J. Pestic. Sci.* **2013**, *38*, 167–168. [CrossRef]
47. AbuQamar, S.F.; Moustafa, K.; Tran, L.S. 'Omics' and plant responses to *Botrytis cinerea*. *Front. Plant Sci.* **2016**, *7*, 1658. [CrossRef]
48. Kirsop, B.E.; Doyle, A. *Maintenance of microorganisms and cultured cells, a manual of laboratory methods*, 2nd ed.; Academic Press: London, UK, 1991; p. 308.
49. Vilgalys, R.; Hester, M. Rapid genetic identification and mapping of enzymatically amplifed ribosomal DNA from several *Cryptococcus* species. *J. Bacteriol.* **1990**, *172*, 4239–4246. [CrossRef]
50. Carbone, I.; Kohn, L.M. A method for designing primer sets for speciation studies in filamentous ascomycetes. *Mycologia* **1999**, *91*, 553–555. [CrossRef]
51. Glass, N.L.; Donaldson, G.C. Development of primer sets designed for use with the PCR to amplify conserved genes from filamentous Ascomycetes. *Appl. Environ. Microbiol.* **1995**, *61*, 1323–1330. [CrossRef]
52. Tamura, K.; Stecher, G.; Peterson, D.; Filipski, A.; Kumar, S. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **2013**, *30*, 2725–2729. [CrossRef]

International Journal of
Molecular Sciences

MDPI

Article

Genome-Wide Identification, Expression Profile and Evolution Analysis of Karyopherin β Gene Family in *Solanum tuberosum* Group Phureja DM1-3 Reveals Its Roles in Abiotic Stresses

Ya Xu [1,2], Lu Liu [2], Pan Zhao [2], Jing Tong [3], Naiqin Zhong [2,4], Hongji Zhang [1,*] and Ning Liu [2,3,*]

[1] College of Plant Protection, Yunnan Agricultural University, Kunming 650201, China; xuyapoppy@163.com

[2] State Key Laboratory of Plant Genomics, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China; liulu183@mails.ucas.ac.cn (L.L.); zhaop@im.ac.cn (P.Z.); nqzhong@im.ac.cn (N.Z.)

[3] National Engineering Research Center for Vegetables, Beijing Vegetable Research Center, Beijing Academy of Agricultural and Forestry Sciences, Beijing 100097, China; tongjing@nercv.org

[4] The Enterprise Key Laboratory of Advanced Technology for Potato Fertilizer and Pesticide, Inner Mongolia Autonomous Region, Hulunbuir 021000, China

[*] Correspondence: liuning@im.ac.cn (N.L.); zhanghongji111@163.com (H.Z.)

Received: 3 December 2019; Accepted: 28 January 2020; Published: 31 January 2020

Abstract: In eukaryotic cells, nucleocytoplasmic trafficking of macromolecules is largely mediated by Karyopherin β/Importin (KPNβ or Impβ) nuclear transport factors, and they import and export cargo proteins or RNAs via the nuclear pores across the nuclear envelope, consequently effecting the cellular signal cascades in response to pathogen attack and environmental cues. Although achievements on understanding the roles of several *KPNβ*s have been obtained from model plant *Arabidopsis thaliana*, comprehensive analysis of potato *KPNβ* gene family is yet to be elucidated. In our genome-wide identifications, a total of 13 *StKPNβ* (*Solanum tuberosum* KPNβ) genes were found in the genome of the doubled monoploid *S. tuberosum* Group Phureja DM1-3. Sequence alignment and conserved domain analysis suggested the presence of importin-β N-terminal domain (IBN_N, PF08310) or Exporin1-like domain (XpoI, PF08389) at N-terminus and HEAT motif at the C-terminal portion in most StKPNβs. Phylogenetic analysis indicated that members of StKPNβ could be classified into 16 subgroups in accordance with their homology to human KPNβs, which was also supported by exon-intron structure, consensus motifs, and domain compositions. RNA-Seq analysis and quantitative real-time PCR experiments revealed that, except *StKPNβ3d* and *StKPNβ4*, almost all *StKPNβ*s were ubiquitously expressed in all tissues analyzed, whereas transcriptional levels of several *StKPNβ*s were increased upon biotic/abiotic stress or phytohormone treatments, reflecting their potential roles in plant growth, development or stress responses. Furthermore, we demonstrated that silencing of *StKPNβ3a*, a SA- and H_2O_2-inducible *KPNβ* genes led to increased susceptibility to environmental challenges, implying its crucial roles in plant adaption to abiotic stresses. Overall, our results provide molecular insights into *StKPNβ* gene family, which will serve as a strong foundation for further functional characterization and will facilitate potato breeding programs.

Keywords: karyopherin; solanum tuberosum; abiotic stress; expression analysis

1. Introduction

Unlike the prokaryotic ancestors, the nucleus of eukaryotic cells is surrounded by double layers of lipid membranes, called the nuclear envelope (NE), which provides a controlled barrier between nucleoplasm and cytoplasm [1–3]. The selective transportation of macromolecules across NE provides the eukaryotic cell with essential and additional benefits in regulating exchange of genetic information in response to the changing environments [4–6]. The nucleocytoplasmic transport machinery is composed of a variety of nuclear transport factors: (1) Karyopherin/Importin α (KPN α), which recognize cargo protein with nuclear localization signal (NLS) or nuclear export signal (NES); (2) Karyopherin/Importin β (KPNβ), which binds to KNPα and mediates cargo import into or export out of the nucleus; (3) A small GTPase Ran, which binds to KPNβ and drive directional nucleocytoplasmic transport of cargo-α/β/Ran complex by the RanGTP-RanGDP gradient across the NE [6–11]. In addition to collaborate with KPNα in nucleocytoplasmic transport, the KPNβ family of nuclear transport factors can mediate, by directly recognizing NLS/NES with cargos, most macromolecular transport across NE. Therefore, KPNβs are thought to be critical regulators of a set of cellular processes such as signal transduction, gene expression, immune response, etc. [12,13].

KPNβ is typically characterized with an importin-β N-terminal domain (IBN_N, PF03810) or Exporin1-like domain (XpoI, PF08389) at the N-terminus, and a series of tandemly repeated HEAT (Huntingtin, elongation factor 3, protein phosphatase 2A and yeast PI3-kinase TOR1) motifs at the C-terminal portion [13,14]. Based on the evolutionary and transcriptional analysis, KPNβ family is divided into 15 subfamilies which are named according to human nomenclatures [15,16]. Previous experiments have demonstrated that at least 11 human KPNβs and 10 yeast KPNβs can regulate nucleocytoplasmic transport [13].

Members of KPNβ gene family were identified in many eukaryotic organisms from yeast, plant, to mammal. It has been reported that there are 14 members in yeast and over 20 in human genomes, and *Arabidopsis* genome encodes 18 KPNβ proteins, suggesting individual members of KPNβ gene family might have their unique features [12,17,18]. Current knowledge on plant KPNβ genes were mostly obtained from functional analysis of *Arabidopsis* importin mutants [15,19]. For example, AtKPNB1, member of KPNβ1 subfamily, modulates abscisic acid (ABA) signaling and its loss-of-function mutant exhibits enhanced tolerance to dehydration stress due to the increase sensitivity of stomatal closure in response to ABA [20]. PAUSED, an ortholog of human LOS1/XPOT in *Arabidopsis*, is capable of rescuing the tRNA export defect of *los1* in *Saccharomyces cerevisiae* Meyen ex E.C. Hansen, indicating that their functions are highly evolutionarily conserved [21,22]. However, their genomic distribution and biological functions in plant species other than *Arabidopsis thaliana*, to our knowledge, has been largely uninvestigated yet.

Potato (*Solanum tuberosum*), grown on all continents except Antarctica, is the world's third most important staple crop after rice and wheat in terms of food consumption [23–25]. Although most cultivated potatoes are heterozygous autotetraploid and possess the huge genome, wild diploid potatoes with relatively smaller genome become the ideal targets of potato genome sequencing, which could adequately simplify the genome complexity [26]. Furthermore, wild diploid potatoes are widely used as sources of resistance by potato breeders because they are important reserves of genetic and phenotypic variation to biotic and abiotic stresses [27]. The diploid *S. tuberosum* Group phureja DM, cultivated in South America, was chosen to produce a homozygous double-monoploid clone (*S. tuberosum* group Phureja DM1-3 516 R44) using classical tissue culture techniques [25]. The annotated genome of *S. tuberosum* Group phureja DM1-3, was released in 2011 [26], and afterwards draft genome sequence of *Solanum commersonii*, a tuber-bearing wild potato, was also available in 2015 [28–30]. The genomic information released facilitates the researches on potato functional genomics, and provides an opportunity to conduct genome-wide analysis of nucleocytoplasmic transporters in potato. Here, we performed a genome-wide, comprehensive analysis of *KPNβ* genes. In total, 13 *KPNβ* genes were identified, and further confirmed by sequencing. The physical and chemical characteristics, genomic structures, chromosomal locations, evolutionary relationship, expression profiles of potato *KPNβ* gene family were analyzed in detail. Finally, VIGS (Virus-Induced Gene Silencing) approach was employed to investigate the role of potato *KPNβ3a*, demonstrating that *KPNβ3a* was associated with plant adaption to salt and oxidative stresses. This study provides the molecular information with respect to the *StKPNβ* gene family, paving the way to the further functional characterization of potato *KPNβ* genes.

2. Results

2.1. Genome-Wide Identification of KPNβ Genes from S. tuberosum

To identify KPNβ genes in potato, protein sequences of functionally validated KPNβs from *S. cerevisiae*, *Homo sapiens Linnaeus* and *A. thaliana* were used as the queries to perform BLASTP searches against the potato genome database in Phytozome as well as Potato Genomics Resource. After removing the non-representative splicing forms of same gene locus, 14 KPNβ-like genes were obtained from the genome sequences of *S. tuberosum* phureja DM1-3. Further, the presence of the IBN_N (or XpoI) and Heat repeats domains in these KPNβ-like proteins was scanned using the Conserved Domain Search (CD-search) with e-value <10^{-10}. One possible pseudogene (PGSC0003DMG400029568) was removed from our analysis because its expression could not be detected in all samples and conditions examined in subsequent expression analysis, although its protein sequence is identical to *KPNβ3d*. Eventually, only 13 genes were identified as *StKPNβ* genes (Table 1). According to the homologies against *Arabidopsis* and human *KPNβ*s, the nomenclature of these *StKPNβ* genes was listed in Table 1. The predicted proteins encoded by *StKPNβ* varied from 239 amino acids (StKPNβ3c) to 1111 amino acids (StKPNβ3a), with corresponding molecular weights from 27.2 kDa to 123.1 kDa. Of these putative StKPNβ proteins, the theoretical isoelectric points ranged from 4.22 (StPLANTKAP) to 6.10 (StKPNβ3d), indicating that, as weakly acidic proteins, they could participate biochemical processes under disparate in vivo environments.

2.2. Chromosomal Distribution and Duplication Events among StKPNβ Genes

The physical map position of *StKPNβ* genes on 12 potato chromosomes was established. The number of *StKPNβ*s are unevenly distributed on the potato chromosomes (Figure 1). Chromosome 1 contains the largest number of *StKPNβ* genes comprising six members, chromosome 3 and 9 each contain two members, whereas chromosome 6, 8 and 12 each contain a single *StKPNβ*.

The number of *StKPNβ* genes in potato genome is similar to its counterparts in yeast, human and *Arabidopsis*. Pairwise sequence comparison of StKPNβ proteins suggests that the homology broadly ranged from 4.58% (StXPO5 and StXPO2) to 91.71% (StKPNβ1b and StKPNβ1a). Strikingly, through the sequence similarity between StKPNβs, members in two subclades comprising StKPNβ1a/1b/1c share high sequence identity (64.4–91.7%), suggesting that these StKPNβs in KPNβ1 subfamily are likely to be originated from gene duplications while they are positioned to different chromosomes. A similar event was also found in *KPNβ3* subclade, which includes StKPNβ3a/3b/3c/3d with identity from 34.3% to 90.5%.

Table 1. List of putative *StImpβ* gene family members of *S. tuberosum* Group phureja.

Gene Name [a]	Locus ID [b]	Predicted Proteins	Chromosomal Location [c]			Gene Models [d]	Putative Proteins [e]		
			Chr	Chr_start	Chr_end		Length (aa)	pI	MW (kDa)
StKPNβ 1a	PGSC0003DMG400018525	PGSC0003DMP400032281	3	45710853	45713572	1	871	4.61	96.40
StKPNβ 1b	PGSC0003DMG400019597	PGSC0003DMP400034029	6	38386746	38389457	1	868	4.59	96.00
StKPNβ 1c	PGSC0003DMG400026641	PGSC0003DMP400046282	9	10350557	10353269	1	873	4.62	96.22
StKPNβ 3a	PGSC0003DMG400015862	PGSC0003DMP400027802	1	16901021	16911855	1	1111	4.74	123.08
StKPNβ 3b	PGSC0003DMG401004281	PGSC0003DMP400007618	12	61091176	61100591	1	1021	4.77	113.87
StKPNβ 3c	PGSC0003DMG400023766	PGSC0003DMP400055376	1	1281756	1283167	1	239	4.73	27.15
StKPNβ 3d	PGSC0003DMG400013325	PGSC0003DMP400051493	8	45306052	45311794	1	983	6.10	109.56
StKPNβ 4	PGSC0003DMG400032173	PGSC0003DMP400041127	1	80738183	80745787	5	1049	4.87	115.27
StKAP120	PGSC0003DMG401000117	PGSC0003DMP400011095	9	9658233	9662484	1	307	5.28	33.77
StPLANTKAP	PGSC0003DMG400006259	PGSC0003DMP400021286	1	34095216	34099108	1	438	4.22	49.23
StXOPT	PGSC0003DMG400012034	PGSC0003DMP400039669	3	61569820	61577676	4	990	5.44	111.51
StXPO2	PGSC0003DMG400022883	PGSC0003DMP400000259	1	73593165	73596092	1	975	5.52	109.63
StXPO5	PGSC0003DMG400022491	PGSC0003DMP400038992	1	75806958	75808695	1	320	6.00	35.01

[a] Name referred to systematic designation to members of *KPNβ* family in *S. tuberosum* according to the homology against *Homo sapiens*. [b] Gene accession number in PGSC database. [c] Chromosomal location of the St*KPNβ* genes in the DM1-3 potato genome (V4.3). [d] isomer numbers. [e] Length (number of amino acids), molecular weight(kilodaltons), and isoelectric point (pI) of the deduced polypeptides were calculated using Lasergene Molecular Biology Suite (Version 7.0).

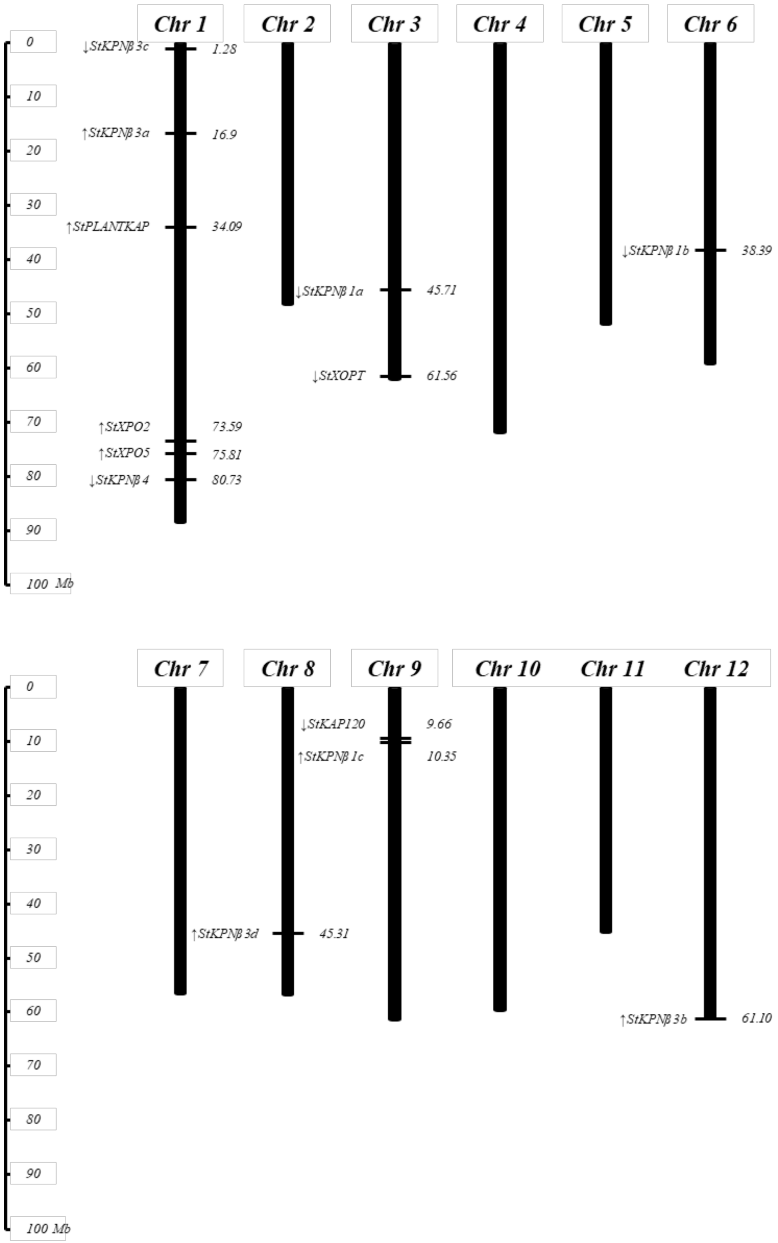


Figure 1. Genomic distribution of *StKPNβ* genes on *S. tuberosum* group phureja DM1-3 chromosomes. The chromosome numbers and size are indicated at the top and bottom of each bar, respectively. The arrows next to gene names show the transcription directions. The number on the right side of the bars designated the approximate physical position of the first exon of corresponding *StKPNβ* genes on potato chromosomes.

2.3. Gene Structure of *StKPNβs*

To better understand the gene structure of *StKPNβ*s, the exon-intron features among members of StKPNβ family were aligned via phylogenetic analysis. The phylogenetic analysis revealed three clusters in accordance with the group data presented in Figure 2. Gene structure analysis of all *StKPNβ* genes suggested that the number of exons ranged from 2 to 20, except that *StXPO2* is intronless gene. It is noteworthy that *StKPNβ* members in KPNβ1 subfamily shares identical intron-exon structure. Three members of *StKPNβ3* subclade were also exhibits similar gene structure while *StKPNβ3c* is a truncated gene. Although the exon-intron structure of *StKPNβ*s varies between subclades, it is similar within subclades, which was supported by the phylogenetic analysis of StKPNβ proteins.

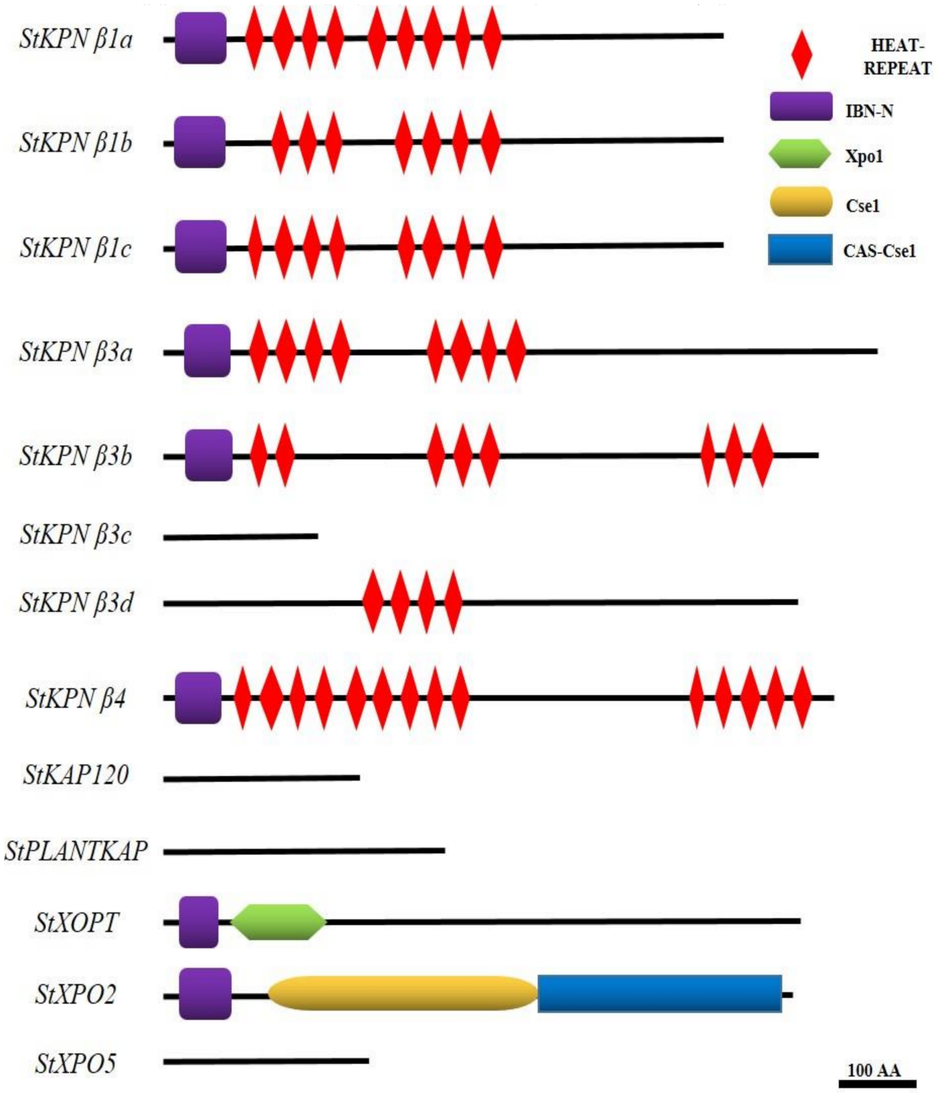


Figure 2. Analysis of conserved domains in StKPNβ proteins. Schematic organization of conserved domains in StKPNβ proteins. The IBN_N domain, HEAT repeats domain, XpoI amd CseI/CAS-CseI domain are shown in purple, red, green and yellow/blue, respectively.

2.4. Conserved Domains and Motif Analysis of StKPNβs

It is well-known that members of KPNβ proteins have common features—the IBN_N or XPO1 domains, and HEAT repeats. For members involved export of macromolecules, the conserved region was also called XPO1/CSE1 domains which contain HEAT repeats and a C-terminal domain. To better understand the structural similarity of potato Impβs, we analyzed the amino acid sequences of *StKPNβ* genes using CD-search available at NCBI with default configurations, and re-annotated the domains mentioned above. As shown on Figure 3, eight members of StKPNβs contain both Heat repeat motif and IBN_N motif, and two members possess IBN_N and XPO1/CSE1 domains. There were three potato KPNBs with high sequence similarity to functionally characterized *Arabidopsis* karyopherins, in which no conserved domains were identified by CD-search. StKPNβ3c, homologous to other StKPNβ3 genes, is truncated gene, which resulted in the loss of conserved domains aforementioned.

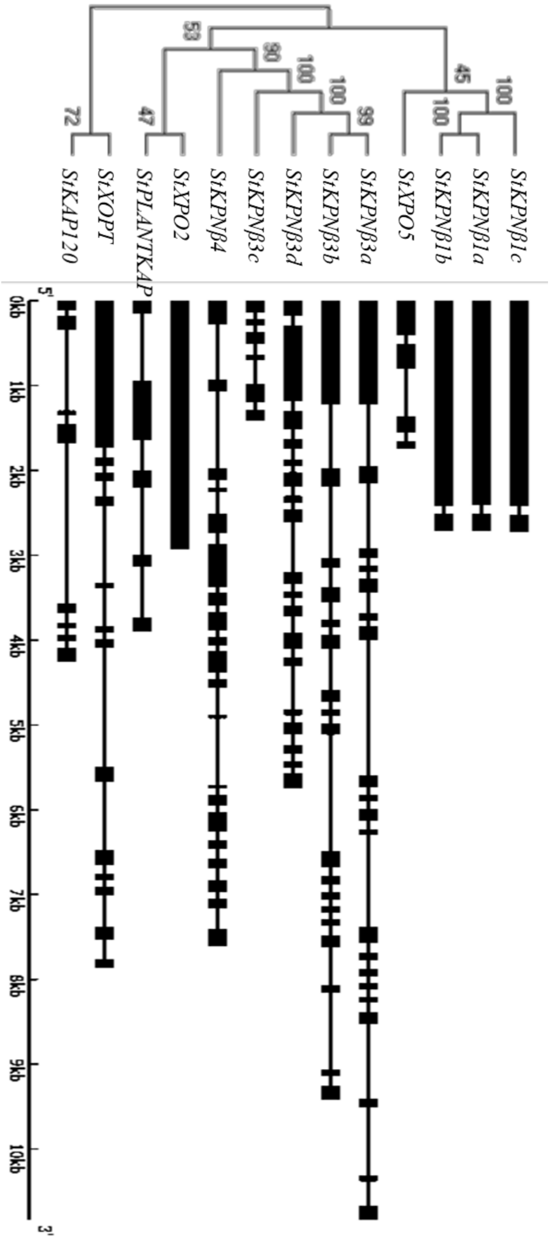


Figure 3. Classification of StKPNβ proteins. Neighbor-joining tree were generated using MEGA X to determine the phylogenetic relationship between StKPNβs (left). The intron-exon organization of *StKPNβ* genes was plotted using Gene Structure Display Server (Version 2.0). Black boxes represent exons and black lines represent introns (right).

In addition to the HEAT and IBN motifs, we searched the compositions of StKPNβs, which was evaluated using the MEME suite (http://meme-suite.org/tools/meme), an online motif discovery tool. In our analysis, four novel conserved motifs were identified, and among the four motifs, motif I was present in all StKPNβ proteins; Motif II was identified in eight StKPNβ members; and motif III was found in 10 StKPNβs (Figure 4), suggesting that these conserved regions might be essential to execute its biological functions. Furthermore, StKPNβs in the same subfamily shared similar patterns of motif composition, indicating that their functional similarities. Thus, distribution of the motifs also reveals that StKPNβs were likely conserved during the evolution.

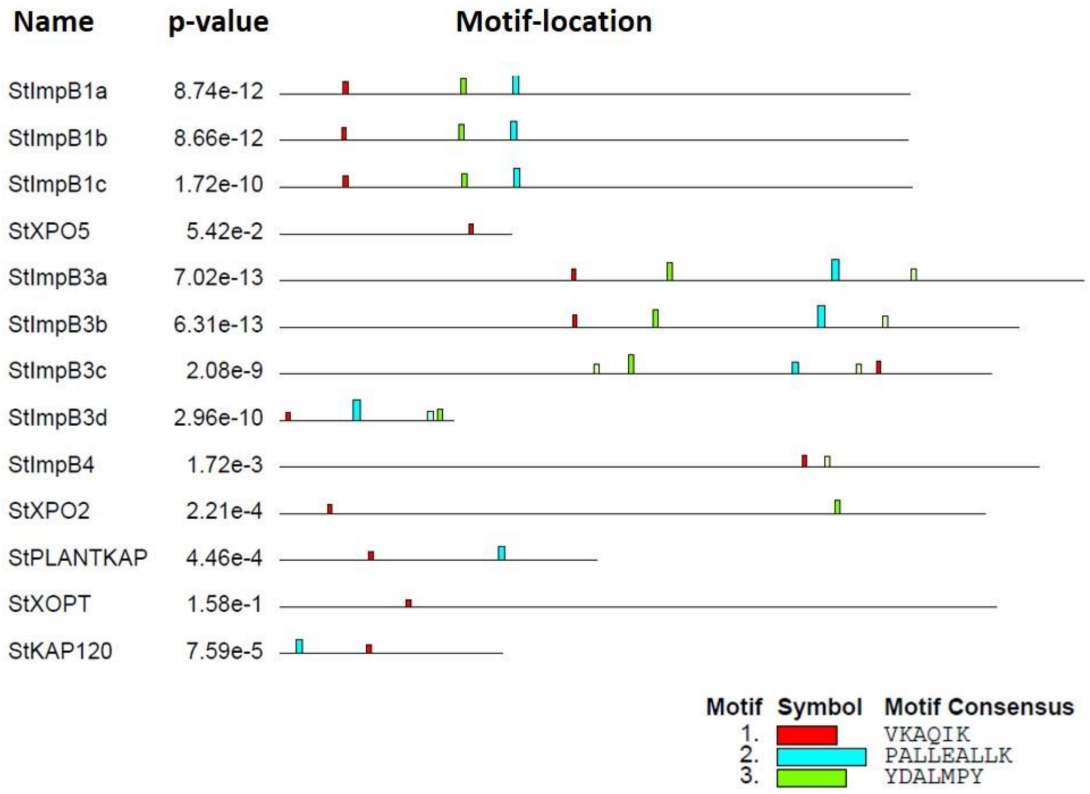


Figure 4. Conserved motifs embedded in the StKPNβ proteins. Conserved motif in StKPNβs was evaluated using the MEME, and the location of novel motifs identified were designated in different colors.

2.5. Phylogenetic Analysis of StKPNβs

To investigate the phylogenetic relationship between the members of *StKPNβ* gene family, a neighbor-joining tree was constructed based on the multiple alignment of karyopherin β protein sequences from *A. thaliana*, *S. tuberosum*, *H. sapiens* and *S. cerevisiae*. All these KPNβ proteins, in accordance with the human KPNβs, were allocated to 16 subfamilies with relatively high confidence (Figure 5). Multiple sequence alignment and phylogenetic analysis suggested that members of the KPNβ family were considerably diverged as the statistical support for some branches was relatively poor. Although yeast is a unicellular organism, at least 13 KPNβs were identified previously in *S. cerevisiae*. These yeast karyopherins included in our analysis actually represented 14 subfamilies of KPNβ nucleocytoplasmic transporters, strongly suggesting that the functional diversification of KPNβ had occurred. Moreover, two yeast KPNβs (NMD5-SXM1), in our phylogenetic tree (Figure 5), were clustered into a sister pair, probably implying they were evolved from a common ancestor. Taken together, the results reinforced that the establishment of KPNβ family predated the appearance of radiation of eukaryote organism, which agrees well with conclusion drawn by O'Reilly et al. [16].

As shown on Figure 5, StKPNβs were distributed into two sister pairs of paralogous Impβs (StKPNβ1a/1b/1c, StKPNβ3a/3b/3c/3d) with strong bootstrap support, while the other six form sister pairs with their *Arabidopsis* orthologs. Surprisingly, no potato ortholog could be detected in several KPNβ subfamilies including KPNB2/IMB2, KPNB5/IMB5, IPO8, XPO1, XPO4, XPO7 and TNPO3, whereas XOPT subfamily is the only one that was lost in *Arabidopsis*. The fact that, compared with yeast and *Arabidopsis*, there are fewer members in the potato KPNβ family reinforces that gene loss occurred after the divergence between Brassicaceae and Solanaceae. Notably, a lineage-specific subclade consisted with two KPNβ members from potato and *Arabidopsis* was detected, suggesting that they might represent a group of plant-specific nucleocytoplasmic transporters. The likely interpretation for the absence of XPO4 and XPO7 subclades from yeast genome indicates that in addition to PLANTKAP subclade, they were derived in multicellular organisms.

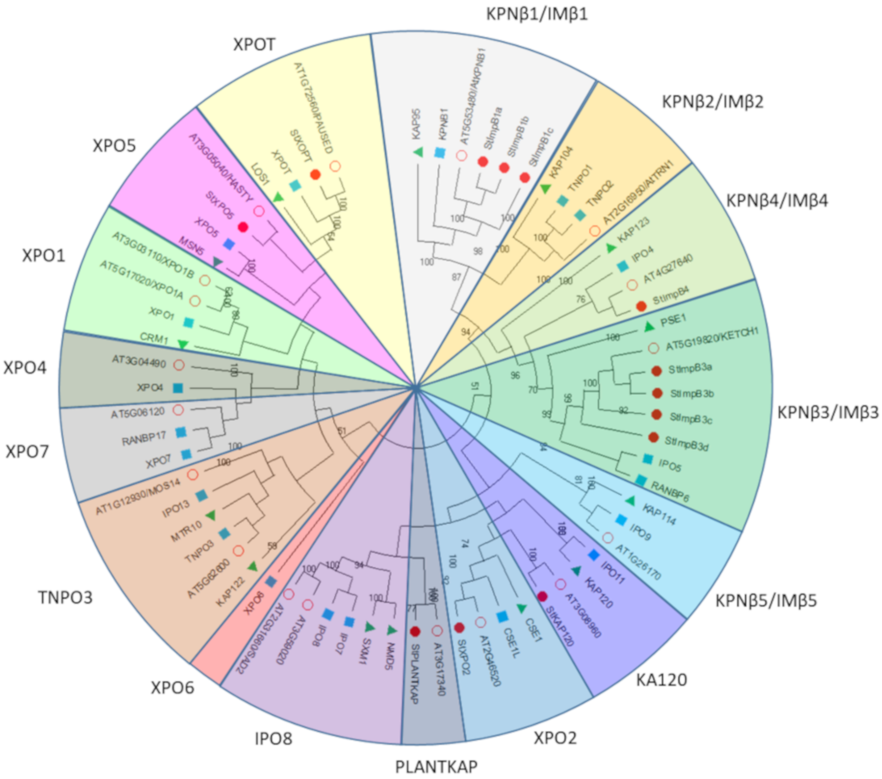


Figure 5. Phylogenetic analysis of StKPNβ proteins in *A. thaliana*, *S. tuberosum*, *S. cerevisiae* and *H. sapiens*. Neighbor-joining tree was constructed based on the alignment of KPNβ protein sequences from *S. cerevisiae* (Green triangle), *H. sapiens* (Blue square), *A. thaliana* (Red empty circle) and *S. tuberosum* (Red circle). The percent bootstrap support for 500 replicates is shown on each branch with >50% support.

2.6. Expression Profiles of StKPNβs among Various Tissues and Developmental Stages

To gain the insight into the tissue- or organ-specific expression preferences of *StKPNβ* genes, we analyzed the transcriptome data from Illumina RNA-Seq reads generated and stored by PGSC. The transcript abundance of 13 *StKPNβ* genes was determined from the RNA-Seq data as FPKM (Fragments per Kilobase of transcript per Million mapped reads) values. The RNA-seq database was generated from 16 tissues which could be divided into three major groups: floral (carpel, stamen, petal, sepal and mature flower), vegetative (leaf, leaflet, shoot, roots, tuber and stolon) and other tissues (callus) [31]. Digital gene expression analysis revealed that, among these 20 *StKPNβ* genes, *StKPNβ1a/1b/3a*, *StKAP120*, *StXPO5* were ubiquitously and robustly expressed in all tissues, suggesting that these StKPNβs might execute some universal roles and participate nucleocytoplasmic transport in various tissues and organs; conversely, the expression level of *StKPNβ3d* and *StKPNβ4*, compared to other *StKPNβ*s, was relatively lower, suggesting that these KPN-βs might be unnecessary in normal growth conditions (Figure 6). Strikingly, transcripts of *StKPNβ1a/3a/3c* were relatively abundant in tuber or stolon tissues, indicating that their possible association with potato tuber development. These results suggest that, as nucleocytoplasmic regulators, members of StKPNβ family have diverse roles of in potato floral and vegetative tissues.

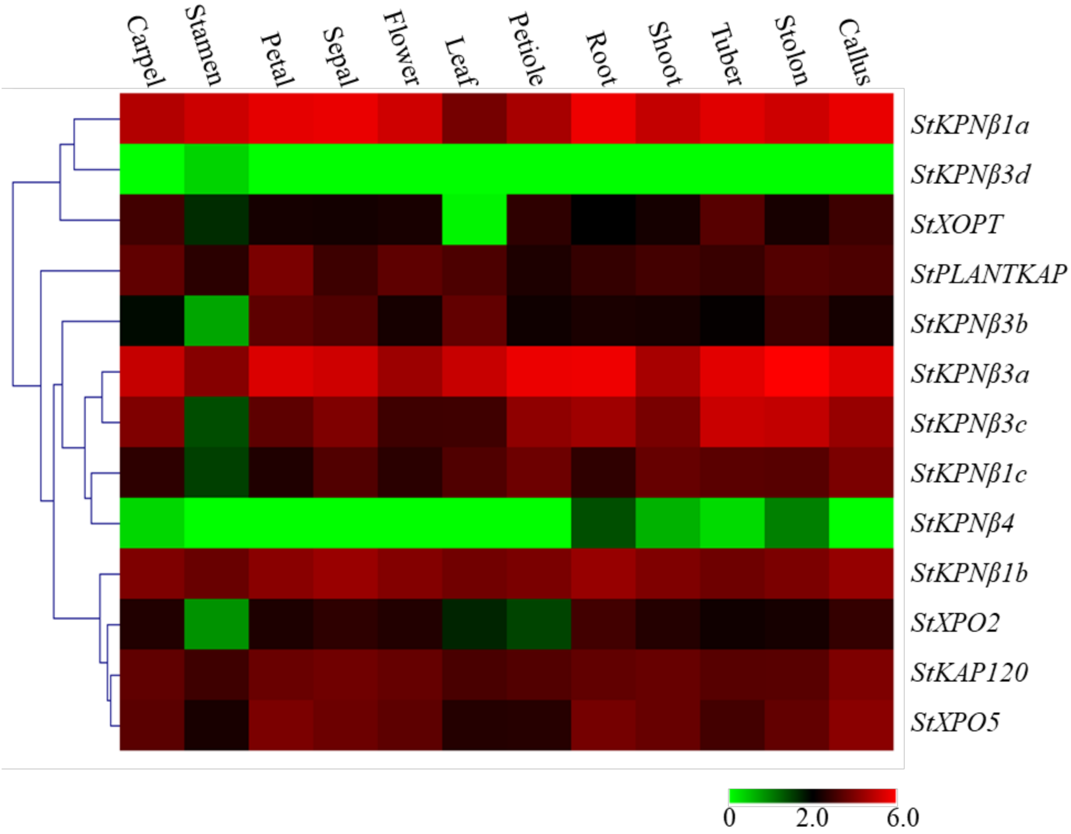


Figure 6. Expression profiles of StKPNβ genes with hierarchical clustering in different tissues. The Illumina RNA-Seq data were obtained from PGSC database, and the FPKM value of representative transcripts of StKPNβs were used to generate heatmap with hierarchical clustering based on the Manhattan correlation with average linkage using MeV software package. Color scale below heatmap shows the expression level; red indicates high transcript abundance while green indicates low abundance.

2.7. Expression Profiles of StKPNβs in Response to Biotic and Abiotic Stresses

To understand the functions of *StKPNβ* genes under various stresses, the transcript abundance of 13 StKPNβ genes was analyzed the log2 fold change between treatments and controls. RNA-Seq data revealed that most *StKPNβ*s were found to be significantly induced by at least one treatment, while the *StKPNβ3b* transcript was not affected by stress conditions (Figure 7a). Of these 13 *StKPNβ* genes, *StKPNβ4* increased by 2.63-fold under high salinity, and 2.21-fold in response to mannitol stress, while *StKPNβ3d* exhibited a high level of transcription abundance under mannitol and wounding stresses, with 1.90-fold and 1.81-fold increase, respectively. The expression of *StXPO2* and *StXPO5* was increased in response to both salt and wounding treatments. These results suggest that StKPNβs might be serve as core regulators in mediating the signaling transduction of abiotic stresses.

Several *StKPNβ* genes were found to be induced by at least one stress condition (Figure 7a). For example, *StPLANTKAP* were increased by 1.83-fold under salt stress, while in response to wounding treatment, *StKPNβ1c*, *StKPNβ3c* and *StXOPT* were highly increased by 1.94-, 2.15- and 2.30-fold, respectively. The expression specificity of these *StKPNβ*s indicates that they were functionally diverged and actively regulated trafficking of different responsive proteins across the nuclear membrane. It seems that most *StKPNβ*s did not respond to thermal and *Phytophtora infestans* (Mont.) de Bary challenges. The oomycetes *P. infestans* infection resulted in the 1.48-fold expression increase of *StXOPT*, suggesting it might involve the process of plant defense against the pathogen. Therefore, this result suggests that StKPNβs are associated with plant responses to abiotic and biotic stresses.

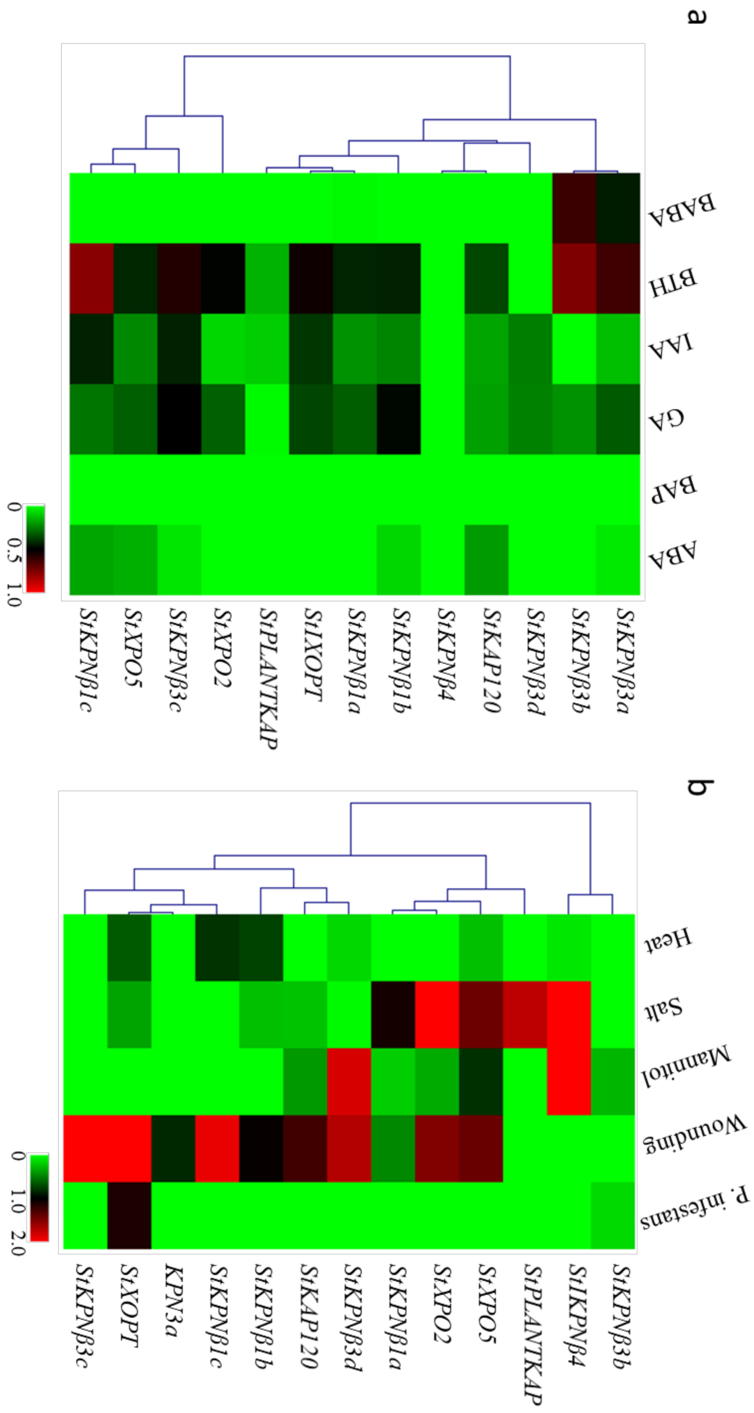


Figure 7. Heatmap representation and hierarchical clustering of *StKPNβ* genes under abiotic and biotic stresses (**a**) and phytohormone treatments (**b**). The Illumina RNA-Seq data were obtained from PGSC database, and the relative expression of *StKPNβ* genes was calculated with respect to control samples using FPKM values of representative transcripts corresponding to *StKPNβ* genes. Fold changes of *StKPNβ* expression were $\log_2$ transformed, and the normalized expression data was used to generate heatmap with MeV software package using the same parameters in Figure 6. Color scale below heatmap shows the expression level; red indicates high transcript abundance, while green indicates low abundance.

2.8. StKPNβs Response to Various Phytohormones

Similarly, we also examined the expression changes of *StKPNβ*s under different phytohormone or chemical analog treatments by RNA-Seq and quantitative real-time RT-PCR (qRT-PCR) analysis. An interesting observation from RNA-Seq analysis was that expression level of a majority of *StKPNβ* genes were decreased when potato plant being treated with phytohormones or their analogs (Figure 7b). When plants treated with benzothiadiazole S-methyl ester (BTH), a chemical analog of salicylic acid, transcript accumulation of *StKPNβ1c*, *StKPNβ3a* and *StKPNβ3b* genes was observed, suggesting their upregulation possibly contributes to the plant defenses to pathologies. Application of DL-β-amino-n-butyric acid (BABA), known as a disease resistance-priming agent, resulted in the weak induction of *StKPNβ3b.*

Although Illumina RNA-Seq data provides plenty of information on the expression profiles of *StKPNβ* genes, we still lack their expression behavior in response to some important signal molecules such as ethylene (ETH), jasmonic acid (JA), hydrogen peroxide (H_2O_2) and salicylic acid (SA). Thus, quantitative real-time RT-PCR (qRT-PCR) analysis was employed to determine the expression patterns of *StKPNβ* genes in these phytohormones or chemicals, and leaf tissues of potato treated with 50 μM SA, 1 mM JA, 1 mM ETH and 50μM H_2O_2, respectively, were used in the experiments.

Most *StKPNβ* genes considered in this study were upregulated upon SA, ETH or JA treatments. Compared to the controls, SA-feeding promoted the expression increase of *StKPNβ1a/3a* and *StXOPT* by at least 6.5-fold under 24 h SA treatment, and similarly *StKPNβ1b/3b/3c/3c/4* and *StPLANTKAP* also exhibited moderately increases, which suggested that they might be involved in the SA-signaling pathway. In JA-feeding experiments, all *StKPNβ* genes displayed an enhanced level of transcript abundance after 4 h treatment, indicating their potential roles in JA-mediated signal transduction. After 4 h ETH treatments, expression of *StKPNβ1b/3c/3d* were strongly activated by ethylene, with 24.0-, 25.8- and 39.9-fold expression increases, respectively; yet, other *StKPNβ*s were slightly induced (Figure 8).

Hydrogen peroxide, predominantly produced during photosynthesis, photorespiration or respiration processes, plays an essential role as signaling molecule in numerous physiological process. The members of *StKPNβ* gene family were simply classified into two groups according to their responsive behavior in response to H_2O_2 upregulated and downregulated. The first group represents StKPNβ genes that were induced by H_2O_2 and correspond to *StKPNβ1a/1c/3a*, *StPLANTKAP* and *StXOPT*, while the second group includes the remaining *StKPNβ*s, of which the expression negatively responded to H_2O_2 (Figure 8). The observations imply that they may be important components of the Reactive oxygen species (ROS) signal cascade in plants. Collectively, these results indicate that StKPNβs were associated with diverse signaling pathways and probably were one of major players in environmental stress and immunity system.

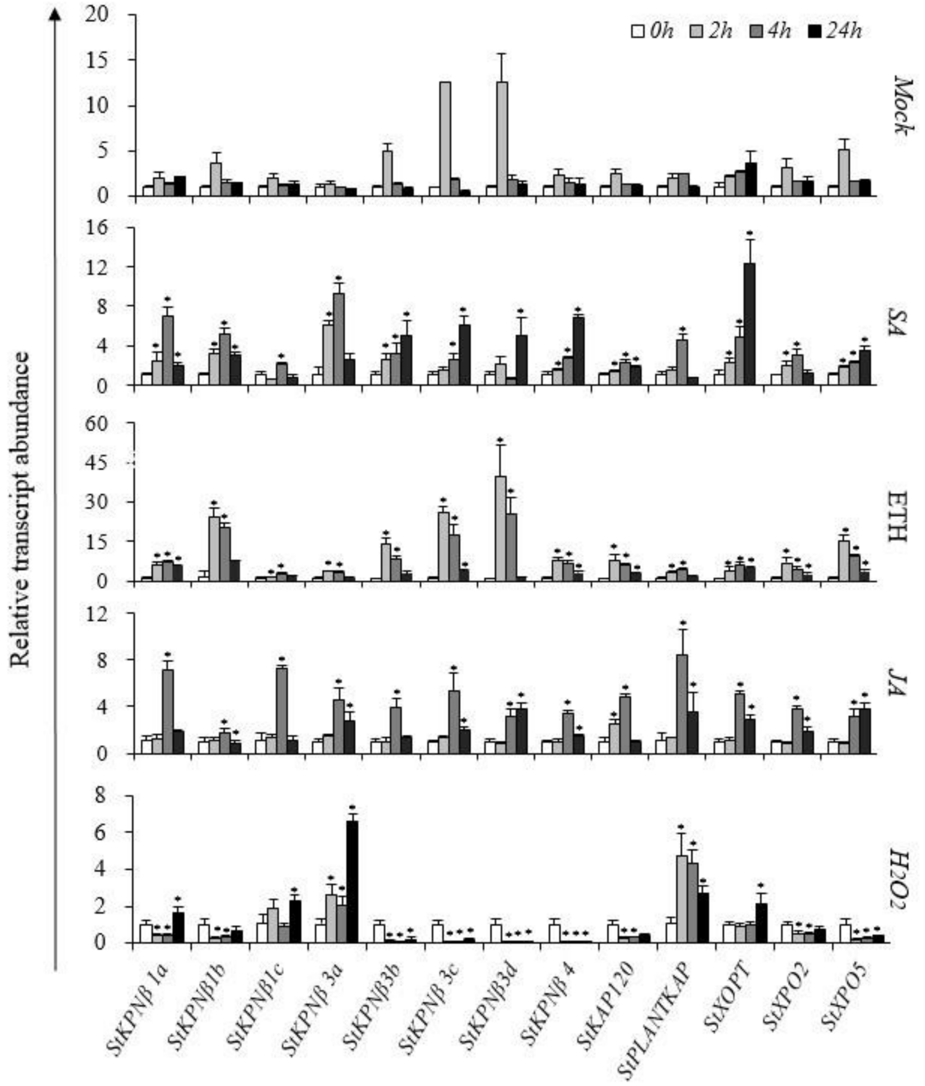


Figure 8. qRT-PCR analysis of *StKPNβ* genes in response to salicylic acid (SA), ethylene (ETH), jasmonic acid (JA) and hydrogen peroxide (H_2O_2). *StKPNβ* transcript levels measured by real-time RT-qPCR from the various tissues or under phytohormone treatments at indicated time points. Data are means of three biological replicates (eight pooled plants each), and error bars denote SE. The *StACT* gene was used as an internal control. Stars above the error bars indicate significant differences between treatments and controls (according to student's t-test). qRT-PCR primers for each *StKPNβ* genes were provided in Table S1.

2.9. Knockdown of StKPNβ3a Expression Results in Increased Susceptibility to Environmental Stresses

Considering that expression of some *StKPNβ* was activated by various stress or hormone treatments, it is plausible that silencing of positively responsive KPNβs would impair the plant tolerance to environmental stresses. Thus, VIGS approach was employed to investigate the role of potato KPNβs. As *StKPNβ3a* was one of highly expressed, H_2O_2- and SA-inducible genes, it was chosen for the insertion into the viral vector pGR107 (PVX), and the resulting plasmid *PVX-StKPNβ3a* was introduced into *Agrobacterium* containing the helper plasmid pJIC SA-Rep. The *Agrobacterium* lines harboring *PVX-StPDS* and empty PVX vector (*PVX00*) were served as controls. Potato plants were transformed by leaf-injection with *Agrobacterium* lines aforementioned, and after one month, all silencing lines were verified by qRT-PCR method. We found that leaves of *PVX-StPDS* lines exhibited photo-bleaching phenotypes, which was agreed with the reduction of *StPDS* genes. Compared to the control plants, transcript accumulation of *StKPNβ3a* was decreased in *StKPNβ3a*-sliciencing lines, whereas expression of *StKPNβ3b*, *StKPNβ3c* and *StKPNβ3d* were not significantly affected (Figure 9b), suggesting that *StKPNβ3a* expression was specifically turn down. Under normal conditions, *StKPNβ3a*-sliciencing lines did not exhibit any morphological changes compared to the control plants (Figure 9c). Subsequently, the leaf discs of *StKPNβ3a*-sciliencing as well as experimental controls were floated on the distilled water supplemented with 300 mM NaCl or 100μM H_2O_2. After 48-hr salt or H_2O_2 treatments, we observed that, compared to the *PVX00* controls, leaf discs of *PVX-StKPNβ3a* lines suffered severe damages (Figure 9a), while there were no evident morphological changes in leaf discs of non-silenced controls. The results illustrated that repression of *StKPNβ3a* could lead to the increased susceptibility to abiotic stresses.

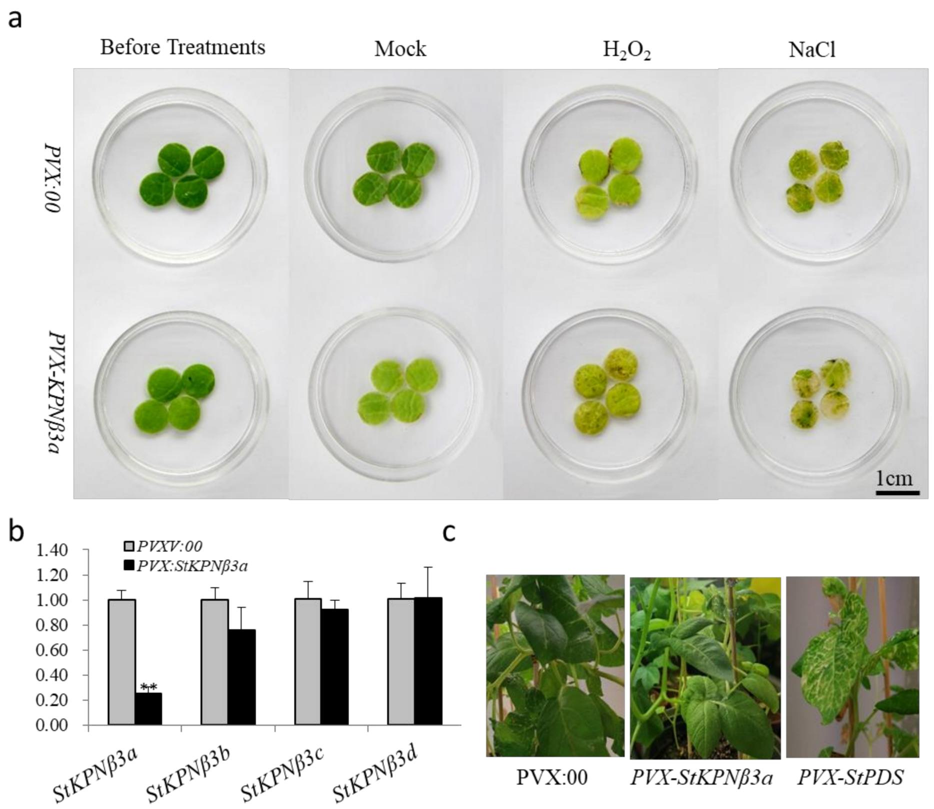


Figure 9. *StKPNβ3a*-silenced potato plants exhibit reduced resistance to salt and H_2O_2 treatments. Potato plants were infiltrated with Agrobacterium carrying VIGS-control vector (PVX:00) and PVX-*StKPNβ3a*, and after 2–3 weeks, the *StKPNβ3a*-silencing lines confirmed by qRT-PCR were used for leaf-disk assay. (**a**) Leaf-disk assay for plant tolerance to different abiotic stresses. (**b**) Expression analysis of *StKPNβ3* members in *StKPNβ3a*-silencing and control lines. (**c**) Phenotype of *PVX-StKPNβ3a*-Silencing and control potato plants. The photographs were taken before or after 48-hrs salt (300 mM) or H_2O_2 (100μM) treatments, respectively. qRT-PCR analysis of *StKPNβ3a* expression in cotton plants infiltrated with VIGS-control vector (PVX:00) and PVX-*StKPNβ3a*. Error bars indicate SD from three technical replicates of three biological experiments, and asterisks indicate statistically significant differences, as determined by the Student's t test (**, $p < 0.01$). The experiments were repeated three times with similar results.

3. Discussion

Karyopherin/Importin β, as an essential nucleocytoplasmic transport receptor, is considered to be a global regulator of diverse cellular functions, ultimately affecting the growth, development and stress adaptions of the eukaryotes [32]. However, current knowledge on its characteristics of was largely obtained from functional characterization of animal and yeast *KPNβ* genes. In the past two decades, achievements have been made in understanding the role of KPNβ in model plant *A. thaliana*, and several *KPNβ* genes, including *Hasty*, *SAD2/EMA1*, *AtKPNB1*, *MOS14* and *KETCH1*, were investigated in detail, demonstrating their vital roles involved in the *Arabidopsis* development, biotic and abiotic stresses [21,22,33–36]. However, the identification and functional analysis of KPNβ homologs still limited in plants other than *Arabidopsis*. Hence, analyses of *KPNβ* gene family in *S. tuberosum* become indispensable in understanding of its gene structure, protein function and evolution.

The number of *KPNβ* genes varies among organisms. In the study, 13 *KPNβ* genes were identified from potato genome, whereas previous search identified 18 *KPNβ*s in *Arabidopsis* [20,31]. Considering that potato has undergone two rounds of whole-genome duplication (WGD) events so that the genome size of DM1-3 potato was nearly five times larger than *Arabidopsis* [26], the observations on *StKPNβ* gene family contradicted with genome complexity between potato and *Arabidopsis*. Therefore, it is interesting that the number of *StKPNβ* genes was much less than that of *Arabidopsis*. Our phylogenetic analysis revealed that eight KPNβ subfamilies, namely KPNβ1/IMβ1, KPNβ3/Impβ3, KPNβ4/Impβ4, KA120, PLANTKAP, XPO2, XPO5 and XOPT, were represented by at least one KPNβ ortholog in potato genome, and duplication events occurred only in KPNβ1 and KPNβ3 subfamilies, perhaps due to the independent, small-scale, segmental duplication events and chromosome rearrangements in the two loci. Nevertheless, in comparison to yeast and *Arabidopsis*, it seems that homologs to other seven KPNβ subfamilies (KPNβ2/Impβ2, KPNβ5/Impβ5, IPO8, XPO1, XPO6, XPO4, XPO7 and TNPO3) were lost completely during the evolution in potato genome, consequently resulting in the fewer members of KPNβ in potato genome.

Functional redundancy and diversification were observed in potato KPNβ1/Impβ1 and KPNβ3/Impβ3 gene subfamilies. With respect to KPNβ1 subfamily, phylogenetic analysis and sequence alignment revealed the existence of three genes, namely *StKPNβ1a*, *StKPNβ1b* and *StKPNβ1c*, homologous to *AtKPNB1*, which raises the possibility that these *StKPNβ1*s might execute similar functions. Consistent with the assumptions, we found that expression patterns under stress or phytohormone treatments, to a large extent, resembled among members of *StKPNβ1*, implying that members of *StKPNβ1s* might share some conserved and overlapping functions. Nevertheless, it was noteworthy that some expression discrepancies between *StKPNβ1* genes, because expression analysis demonstrated that only *StKPNβ*1a could not respond to wounding stress, while expression of *StKPNβ*1c, instead of *StKPNβ*1a and *StKPNβ*1b, was able to be strongly activated by wounding treatment, which reflects that members of *StKPNβ*1 subfamily might have acquired its unique roles through functional diversifications.

Recent investigations have reported that a few *Arabidopsis* KPNβ/Impβs, as nucleo-cytoplasmic transport receptors, are involved in stress adaption under abiotic and biotic stresses, while they are not stress-inducible genes [20,37]. *AtKPNB1* encodes an ortholog of human KPNB1 in *Arabidopsis*, and *kpnb1* loss-of-function mutant exhibits increased sensitivity to ABA [20]. It was proven that AtKPNB1, functioning as negative regulator, could regulate the ABA responses and drought tolerance via ABI1- and ABI5-independent pathways, though ABA treatment only slightly boosted the transcript accumulation of *AtKPNB1* [24]. In addition, absence of SAD2 (Super sensitive to ABA and Drought 2), member of IPO8 subfamily, led to the enhanced sensitivity to ABA, H_2O_2 or drought in *Arabidopsis*, whereas its expression was independent from phytohormone or stress treatments [37,38]. In agreement with previous findings, our RNA-Seq analysis also suggested that many members of *StKPNβ* gene family did not show any transcriptional responses to hormones or stresses examined, and only several *StKPNβs*, such as *StKPNβ3d, StKPNβ3b, StPLANTKAP, StXOPT* etc., were able to respond to environmental cues or phytohormone inductions.

It is tempering to analyze the roles of responsive StKPNβs, especially whose expression could be activated by hormones, environmental cues or pathogen infections. Expression of *StKPNβ3a* was strongly induced by SA, JA or H_2O_2 treatments, suggesting its involvements in the phytohormone cascades. Thus, using VIGS approach, we demonstrated that silencing of *StKPNβ3a* resulted in the increased susceptibility to salt or oxidative stresses, supporting that its function is indispensable in the stress signaling transductions. However, due to the lack of stable transgenic lines of *StKPNβ3a*-overexpression or -RNAi, the biological functions of *StKPNβ3a* still need to be investigated in detail. Phylogenetic analysis supported that *StKPNβ3a* were orthologous to yeast *PSE1/Kap121*, human IPO5 and RANBP6. Yeast strains with disruption of PSE1 functions exhibit delayed mitosis and enhance sensitivity to temperature stress, while overexpression *PSE1* contributes to the three-fold increase of cellulose production [39–41]. The import of histone H2A/H2B and H3/H4 is mainly mediated by PSE1 in *S. cerevisiae*, suggesting its essential roles in intranuclear transport [42,43]. It has been demonstrated that human IPO5 also functions in the nuclear import of essential histones as well as some ribosomal proteins [44]. Given that members of *Impβ3* subclades play key roles in nucleocytoplasmic trafficking, it is reasonable that *StKPNβ3a* might execute the similar roles by regulating the import of positive regulatory protein(s) under abiotic stresses. Further investigations will be still required to identify its cargo(s) and to articulate the molecular mechanism of Impβ-mediated signaling pathway in plants.

4. Materials and Methods

4.1. Plant Material and Treatments

S. tuberosum Phureja DM1-3 or cultivar "Shpedy" plants were in vitro micropropagated on Murashige and Skoog (MS) medium plus 30 gL^{-1} sucrose and 0.8% agar (Sigma-Aldrich, USA), with pH adjusted to 5.8. Potato seedlings were routinely subcultured as two-node segments every 3–4 weeks and incubated at 23 °C with 16 h photoperiod under cool with fluorescent lamps (~70 μmol m^{-2} s^{-1} photon flux idensity). 3-week old potato plants were subjected to IAA (50 μM), SA (1 mM), ethylene (1mM) or H_2O_2 (1mM) treatments. The plant tissues were collected at designated points and immediately frozen in liquid nitrogen. Sample collections were performed on separate days for the replicates.

4.2. Identification of KPNβ Genes in S. tuberosum Group Phureja

To investigate the KPNβ gene family in in *S. tuberosum* Group phureja DM1-3, all members of KPNβ/Impβ sequences from Human (*H. sapiens*), yeast (*S. cerevisia*e) and *Arabidopsis* were used as queries for BLAST search against Phytozome (https://phytozome.jgi.doe.gov/), NCBI (http://blast.ncbi.nlm.nih.gov/), Potato Genomics Resource (http://solanaceae.plantbiology.msu.edu/) and other online resources with default parameters. The StKPNβ candidates were confirmed the presences of IBN_N (PF08310) or XpoI (PF08389) domain, and HEAT repeats using SMART (http://smart.embl-heidelberg.de/smart/batch.pl) and CDD-search. In order to obtain non-redundancy KPNβ sequences, potato KPNβ sequences were used as queries to blast against Phytozome database, and any redundancy was manually removed. The representing gene model per *StKPNβ* locus were identified and their corresponding information on chromosomal location, locus ID, transcript ID were obtained simultaneously.

4.3. Analysis of Gene Structure and Conserved Domains

Based on the genome annotation of DM assembly available in Phytozome, the intron-exon structure of individual *StKPNβ* genes was predicated, and its genomic organization was visualized using Gene Structure Display Server 2.0 (GSDS, http://gsds.cbi.pku.edu.cn/) [45]. Conserved domains in protein sequences were verified using ScanProsite (http://pro-site.expasy.org/scanprosite/), which provides information about positions of different domains in the protein sequence. This information was used to draw visual representation of distribution of domains in the deduced amino acid sequences of proteins using Microsoft Office PowerPoint 2016.

4.4. Sequence Alignment and Phylogenetic Construction

Multiple alignment of KPNβ protein sequences from *A. thaliana*, *S. tuberosum*, *H. sapiens* and *S. cerevisiae* was conducted using ClustalW [46]. Neighbor-joining method was used to conduct a phylogenetic tree analysis in MEGA X, with 500 bootstrap replicates and randomized sequence input order.

4.5. Expression Profiling of StKPNβ Genes in Different Tissues or Under Various Stresses

The RNA-Seq data corresponding to *StKPNβ* genes was downloaded from the Potato Genomics Resource [31], and the corresponding FPKM (fragments per kilobase per million reads) values for *StKPNβ* genes were obtained for 12 tissues representing major organs and developmental stages, including floral (carpel, petals, sepals, stamens and mature flower), leaf (whole leaf, leaflet and petiole), tuber (tuber and stolon), and other organs (shoot, root and callus). As described, biotic and abiotic treated tissues included potato plants exposed to heat (35 °C), NaCl (150 mM) or Mannitol (260 mM), and leaves challenged by *P. infestans*, BABA (DL-β-amino-n-butyric acid), BTH (6-benzylaminopurine) or hormones [31]. Similarly, FPKM values for abiotic or biotic stress-treated potato plants were analyzed by calculating the fold change of expression levels between treatments and the corresponding controls. The normalized expression data was used to generate heatmap by using the MeV software package (http://mev.tm4.org) available at the Institute for Genomic Research, and hierarchical clustering analysis (HCA) was built on the basis of the Manhattan correlation with average linkage method.

4.6. RNA Extraction and Quantitative Real-Time RT-PCR

Total RNA was extracted with Trizol (Invitrogen Inc., Madison, WI, USA) as described previously [47,48]. RNA quantity and quality were assessed using a NanoDrop8000 (Thermo Scientific™, Wilmington, DE, USA). Total RNA isolation and reverse transcription with oligo $(dT)_{18}$ (18418-012; Invitrogen, Madison, WI, USA) were performed as described previously. The amounts of individual genes were measured with gene-specific primers by real-time PCR analysis with a cycler IQ real-time PCR instrument CFX96 and SYBR Green mixture (Bio-Rad, Foster City, CA, USA). The relative expression of specific genes was quantitated with the $2^{-\Delta\Delta Ct}$ calculation method [49], where ΔΔCt is the difference in the threshold cycles and the reference housekeeping gene, which was potato *StACT* (PGSC0003DMG400027746) for expression analyses. The sequences of specific primers are shown in Table S1.

4.7. Virus-Induced Gene Silencing (VIGS) of Potato

The potato virus X (PVX)-induced gene silencing is conducted as described previously [50,51]. Briefly, *PVX-StKPNβ3a* were generated by cloning a PCR fragment amplified by *S. tuberosum* phureja DM1-3 potato leaf cDNA template using specific oligonucleotide primers incorporating *SalI* and *ClaI* restriction sites, respectively, at the 5′- and 3′-ends for cloning into virus vector pGR107. The *Agrobacterium tumefaciens* (Smith & Townsend, 1907) strain GV3101 harboring the recombinant plasmids *PVX-StKPNβ3a* and help plasmid pJIC SA_Rep were used for in vitro agroinoculation by leaf-injecting of 4-week-old potato plants. The *Agrobacterium* lines carrying with *PVX-StPDS* and the PVX vectors were used as positive and negative controls, respectively. Primers used for RT-PCR amplifications are listed in Table S1.

5. Conclusions

In this study, the systematic characterization of *KPNβ/Impβ* gene family was performed in the *S. tuberosum*. A total of 13 *StKPNβ* genes were identified through searching potato genome, and their chromosomal distribution, conserved domain, motif composition and intron-exon structure were studied in detail. Expression analysis based on the RNA-Seq and qRT-PCR analysis suggested that several *StKPNβs* was responsive to biotic and/or abiotic stresses. Furthermore, the function of *StKPNβ3a* was characterized through VIGS approach, illustrating that it might be a promising candidate gene for molecular breeding. In summary, our results provide valuable insights of *StKPNβs* gene family, which will facilitate further functional analysis of *StKPNβs* and will also benefit genetic engineering of potato.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/21/3/931/s1.

Author Contributions: Conceptualization, N.L.; Methodology, N.L.; Investigation, Y.X., L.L., P.Z. and N.L.; Data analysis, Y.X., J.T. and N.L.; Project administration, N.Z. and H.Z.; Funding acquisition, N.L. and N.Z..;

Supervision and Manuscript preparation, N.L. All authors have read and agreed to the published version of the manuscript.

Funding: This project was supported by the National Key Research and Development Program of China (2017YFD0200708), the Key Innovation of Science and Technology of Shandong province in China (2018CXCC0303), the Science Transfer and Service (KFJ-STS-ZDTP-056) of Chinese Academy of Sciences, the Key Research and Development Program of Ningxia Autonomous Region in China (2018BBF02021), the National Natural Science Foundation of China (31601622) and the Young Talent Supporting Program (YCXTD0000210) of the Beijing Agricultural and Forestry Sciences.

Acknowledgments: We thank Prof. Guixian Xia (Institute of Microbiology, Chinese Academy of Sciences) and Prof. Zhanhui Wu (Vegetable Research Center, Beijing Academy of Agricultural and Forestry Sciences) for fruitful discussions on this manuscript. We also thank Prof. Jiahe Wu (Institute of Microbiology, Chinese Academy of Sciences) for kindly providing DM1-3 potato seedlings.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baum, D.A.; Baum, B. An inside-out origin for the eukaryotic cell. *BMC Biol.* **2014**, *12*, 76. [CrossRef] [PubMed]
2. Lane, N.; Martin, W. The energetics of genome complexity. *Nature* **2010**, *467*, 929–934. [CrossRef] [PubMed]
3. Xu, X.M.; Meier, I. The nuclear pore comes to the fore. *Trends Plant Sci.* **2008**, *13*, 20–27. [CrossRef] [PubMed]
4. Fahrenkrog, B.; Aebi, U. The nuclear pore complex: Nucleocytoplasmic transport and beyond. *Nat. Rev. Mol. Cell Biol.* **2003**, *4*, 757–766. [CrossRef]
5. Fried, H.; Kutay, U. Nucleocytoplasmic transport: Taking an inventory. *Cell Mol. Life Sci.* **2003**, *60*, 1659–1688. [CrossRef] [PubMed]
6. Gorlich, D.; Mattaj, I.W. Nucleocytoplasmic transport. *Science* **1996**, *271*, 1513–1518. [CrossRef]
7. Goryaynov, A.; Yang, W. Role of molecular charge in nucleocytoplasmic transport. *PLoS ONE* **2014**, *9*, e88792. [CrossRef]
8. Oka, M.; Yoneda, Y. Importin alpha: Functions as a nuclear transport factor and beyond. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* **2018**, *94*, 259–274. [CrossRef]
9. Ullman, K.S.; Powers, M.A.; Forbes, D.J. Nuclear export receptors: From importin to exportin. *Cell* **1997**, *90*, 967–970. [CrossRef]
10. Yeon, S.I.; Youn, J.H.; Lim, M.H.; Lee, H.J.; Kim, Y.M.; Choi, J.E.; Lee, J.M.; Shin, J.S. Development of monoclonal antibodies against human IRF-5 and their use in identifying the binding of IRF-5 to nuclear import proteins karyopherin-alpha1 and -beta1. *Yonsei Med. J.* **2008**, *49*, 1023–1031. [CrossRef]
11. Yoneda, Y.; Hieda, M.; Nagoshi, E.; Miyamoto, Y. Nucleocytoplasmic protein transport and recycling of Ran. *Cell Struct. Funct.* **1999**, *24*, 425–433. [CrossRef] [PubMed]
12. Kimura, M.; Imamoto, N. Biological significance of the importin-beta family-dependent nucleocytoplasmic transport pathways. *Traffic* **2014**, *15*, 727–748. [CrossRef] [PubMed]
13. Strom, A.C.; Weis, K. Importin-beta-like nuclear transport receptors. *Genome Biol.* **2001**, *2*, REVIEWS3008. [CrossRef] [PubMed]
14. Vetter, I.R.; Arndt, A.; Kutay, U.; Gorlich, D.; Wittinghofer, A. Structural view of the Ran-Importin beta interaction at 2.3 A resolution. *Cell* **1999**, *97*, 635–646. [CrossRef]
15. Tamura, K.; Hara-Nishimura, I. Functional insights of nucleocytoplasmic transport in plants. *Front. Plant Sci.* **2014**, *5*, 118. [CrossRef]
16. O'Reilly, A.J.; Dacks, J.B.; Field, M.C. Evolution of the karyopherin-beta family of nucleocytoplasmic transport factors; ancient origins and continued specialization. *PLoS ONE* **2011**, *6*, e19308.
17. Goldfarb, D.S.; Corbett, A.H.; Mason, D.A.; Harreman, M.T.; Adam, S.A. Importin alpha: A multipurpose nuclear-transport receptor. *Trends Cell Biol.* **2004**, *14*, 505–514. [CrossRef]
18. Huang, J.G.; Yang, M.; Liu, P.; Yang, G.D.; Wu, C.A.; Zheng, C.C. Genome-wide profiling of developmental, hormonal or environmental responsiveness of the nucleocytoplasmic transport receptors in *Arabidopsis*. *Gene* **2010**, *451*, 38–44. [CrossRef]
19. Yang, Y.; Wang, W.; Chu, Z.; Zhu, J.-K.; Zhang, H. Roles of Nuclear Pores and Nucleo-cytoplasmic Trafficking in Plant Stress Responses. *Front. Plant Sci.* **2017**, *8*, 574. [CrossRef]

20. Luo, Y.; Wang, Z.; Ji, H.; Fang, H.; Wang, S.; Tian, L.; Li, X. An *Arabidopsis* homolog of importin beta1 is required for ABA response and drought tolerance. *Plant J.* **2013**, *75*, 377–389. [CrossRef]
21. Hunter, C.A.; Aukerman, M.J.; Sun, H.; Fokina, M.; Poethig, R.S. PAUSED encodes the *Arabidopsis* exportin-t ortholog. *Plant Physiol.* **2003**, *132*, 2135–2143. [CrossRef] [PubMed]
22. Li, J.; Chen, X. PAUSED, a putative exportin-t, acts pleiotropically in *Arabidopsis* development but is dispensable for viability. *Plant Physiol.* **2003**, *132*, 1913–1924. [CrossRef] [PubMed]
23. Musarella, C. *Solanum torvum* Sw. (Solanaceae): A new alien species for Europe. *Genet Resour. Crop Evol.* **2020**, *67*, 515–522. [CrossRef]
24. FAOSTAT. Available online: http://www.fao.org/faostat/en/$\backslash$#data/QC (accessed on 6 June 2018).
25. Knapp, S.; Vorontsova, M.S.; Särkinen, T. Dichotomous keys to the species of *Solanum L.* (Solanaceae) in continental Africa, Madagascar (incl. the Indian Ocean islands), Macaronesia and the Cape Verde Islands. *PhytoKeys* **2019**, *127*, 39–76. [CrossRef]
26. The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature* **2011**, *475*, 189–195. [CrossRef]
27. Hawkes, J. *The Potato: Evolution, Biodiversity and Tenetic Resources*; Belhaven Press: London, UK, 1990.
28. Aversano, R.; Contaldi, F.; Ercolano, M.R.; Grosso, V.; Iorizzo, M.; Tatino, F.; Xumerle, L.; Dal Molin, A.; Avanzato, C.; Ferrarini, A.; et al. The *Solanum commersonii* Genome Sequence Provides Insights into Adaptation to Stress Conditions and Genome Evolution of Wild Potato Relatives. *Plant Cell* **2015**, *27*, 954–968. [CrossRef]
29. Esposito, S.; Aversano, R.; Bradeen, J.M.; Di Matteo, A.; Villano, C.; Carputo, D. Deep-sequencing of *Solanum commersonii* small RNA libraries reveals riboregulators involved in cold stress response. *Plant Biol. (Stuttg)* **2020**, *22* Suppl. 1, 133–142. [CrossRef]
30. Esposito, S.; Aversano, R.; D'Amelia, V.; Villano, C.; Alioto, D.; Mirouze, M.; Carputo, D. Dicer-like and RNA-dependent RNA polymerase gene family identification and annotation in the cultivated *Solanum tuberosum* and its wild relative *S. commersonii*. *Planta* **2018**, *248*, 729–743. [CrossRef]
31. Massa, A.N.; Childs, K.L.; Lin, H.; Bryan, G.J.; Giuliano, G.; Buell, C.R. The transcriptome of the reference potato genome *Solanum tuberosum* Group Phureja clone DM1–3 516R44. *PLoS ONE* **2011**, *6*, e26801. [CrossRef]
32. Harel, A.; Forbes, D.J. Importin beta: Conducting a much larger cellular symphony. *Mol. Cell* **2004**, *16*, 319–330.
33. Bollman, K.M.; Aukerman, M.J.; Park, M.Y.; Hunter, C.; Berardini, T.Z.; Poethig, R.S. HASTY, the *Arabidopsis* ortholog of exportin 5/MSN5, regulates phase change and morphogenesis. *Development* **2003**, *130*, 1493–1504. [CrossRef] [PubMed]
34. Wang, W.; Ye, R.; Xin, Y.; Fang, X.; Li, C.; Shi, H.; Zhou, X.; Qi, Y. An importin beta protein negatively regulates MicroRNA activity in *Arabidopsis*. *Plant Cell* **2011**, *23*, 3565–3576. [CrossRef] [PubMed]
35. Xu, S.; Zhang, Z.; Jing, B.; Gannon, P.; Ding, J.; Xu, F.; Li, X.; Zhang, Y. Transportin-SR Is Required for Proper Splicing of Resistance Genes and Plant Immunity. *PLoS Genet.* **2011**, *7*, e1002159. [CrossRef] [PubMed]
36. Zhang, Z.; Guo, X.; Ge, C.; Ma, Z.; Jiang, M.; Li, T.; Koiwa, H.; Yang, S.W.; Zhang, X. KETCH1 imports HYL1 to nucleus for miRNA biogenesis in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 4011–4016. [CrossRef] [PubMed]
37. Zheng, Y.; Zhan, Q.; Shi, T.; Liu, J.; Zhao, K.; Gao, Y. The nuclear transporter SAD2 plays a role in calcium- and H2 O2 -mediated cell death in *Arabidopsis*. *Plant J.* **2020**, *101*, 324–333. [CrossRef] [PubMed]
38. Verslues, P.E.; Guo, Y.; Dong, C.H.; Ma, W.; Zhu, J.K. Mutation of *SAD2*, an importin beta-domain protein in *Arabidopsis*, alters abscisic acid sensitivity. *Plant J.* **2006**, *47*, 776–787. [CrossRef]
39. Kroukamp, H.; den Haan, R.; van Wyk, N.; van Zyl, W.H. Overexpression of native *PSE1* and *SOD1* in *Saccharomyces cerevisiae* improved heterologous cellulase secretion. *Appl. Energy* **2013**, *102*, 150–156. [CrossRef]
40. Makhnevych, T.; Lusk, C.P.; Anderson, A.M.; Aitchison, J.D.; Wozniak, R.W. Cell cycle regulated transport controlled by alterations in the nuclear pore complex. *Cell* **2003**, *115*, 813–823. [CrossRef]
41. Ueta, R.; Fukunaka, A.; Yamaguchi-Iwai, Y. Pse1p mediates the nuclear import of the iron-responsive transcription factor Aft1p in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **2003**, *278*, 50120–50127. [CrossRef]
42. Mosammaparast, N.; Jackson, K.R.; Guo, Y.; Brame, C.J.; Shabanowitz, J.; Hunt, D.F.; Pemberton, L.F. Nuclear import of histone H2A and H2B is mediated by a network of karyopherins. *J. Cell Biol.* **2001**, *153*, 251–262. [CrossRef]

43. Thiriet, C.; Hayes, J.J. Histone dynamics during transcription: Exchange of H2A/H2B dimers and H3/H4 tetramers during pol II elongation. *Results Probl. Cell. Differ.* **2006**, *41*, 77–90. [PubMed]
44. Yaseen, N.R.; Blobel, G. Cloning and characterization of human karyopherin beta3. *Proc. Natl. Acad. Sci. tUSA* **1997**, *94*, 4451–4456. [CrossRef] [PubMed]
45. Hu, B.; Jin, J.; Guo, A.Y.; Zhang, H.; Luo, J.; Gao, G. GSDS 2.0: An upgraded gene feature visualization server. *Bioinformatics* **2015**, *31*, 1296–1297. [CrossRef] [PubMed]
46. Thompson, J.D.; Higgins, D.G.; Gibson, T.J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994**, *22*, 4673–4680. [CrossRef] [PubMed]
47. Liu, N.; Fromm, M.; Avramova, Z. H3K27me3 and H3K4me3 chromatin environment at super-induced dehydration stress memory genes of *Arabidopsis thaliana*. *Mol. Plant* **2014**, *7*, 502–513. [CrossRef]
48. Song, S.; Hao, L.; Zhao, P.; Xu, Y.; Zhong, N.; Zhang, H.; Liu, N. Genome-wide Identification, Expression Profiling and Evolutionary Analysis of Auxin Response Factor Gene Family in Potato (*Solanum tuberosum* Group Phureja). *Sci Rep.* **2019**, *9*, 1755. [CrossRef]
49. Livak, K.J.; Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{(\text{-Delta Delta C(T)})}$ Method. *Methods* **2001**, *25*, 402–408. [CrossRef]
50. Faivre-Rampant, O.; Gilroy, E.M.; Hrubikova, K.; Hein, I.; Millam, S.; Loake, G.J.; Birch, P.; Taylor, M.; Lacomme, C. Potato virus X-induced gene silencing in leaves and tubers of potato. *Plant Physiol.* **2004**, *134*, 1308–1316. [CrossRef]
51. Lacomme, C.; Chapman, S. Use of potato virus X (PVX)-based vectors for gene expression and virus-induced gene silencing (VIGS). *Curr. Protoc. Microbiol.* **2008**. [CrossRef]

International Journal of
Molecular Sciences

Review

RNA-seq and ChIP-seq as Complementary Approaches for Comprehension of Plant Transcriptional Regulatory Mechanism

Isiaka Ibrahim Muhammad [1], Sze Ling Kong [1], Siti Nor Akmar Abdullah [1,2,*] and Umaiyal Munusamy [1]

[1] Laboratory of Plantation Science and Technology, Institute of Plantation Studies, Universiti Putra Malaysia, Selangor 43400, Malaysia; muhammadii108@gmail.com (I.I.M.); szeling0923@gmail.com (S.L.K.); yalyagu@gmail.com (U.M.)

[2] Department of Agriculture Technology, Faculty of Agriculture, Universiti Putra Malaysia, Selangor 43400, Malaysia

* Correspondence: snaa@upm.edu.my; Tel./Fax: +603-9769-1044

Received: 8 December 2019; Accepted: 23 December 2019; Published: 25 December 2019

Abstract: The availability of data produced from various sequencing platforms offer the possibility to answer complex questions in plant research. However, drawbacks can arise when there are gaps in the information generated, and complementary platforms are essential to obtain more comprehensive data sets relating to specific biological process, such as responses to environmental perturbations in plant systems. The investigation of transcriptional regulation raises different challenges, particularly in associating differentially expressed transcription factors with their downstream responsive genes. In this paper, we discuss the integration of transcriptional factor studies through RNA sequencing (RNA-seq) and Chromatin Immunoprecipitation sequencing (ChIP-seq). We show how the data from ChIP-seq can strengthen information generated from RNA-seq in elucidating gene regulatory mechanisms. In particular, we discuss how integration of ChIP-seq and RNA-seq data can help to unravel transcriptional regulatory networks. This review discusses recent advances in methods for studying transcriptional regulation using these two methods. It also provides guidelines for making choices in selecting specific protocols in RNA-seq pipelines for genome-wide analysis to achieve more detailed characterization of specific transcription regulatory pathways via ChIP-seq.

Keywords: RNA-sequencing; ChIP-sequencing; transcriptome; transcriptional regulatory mechanism; data integration

1. Introduction

The transcriptome defines the functional element in a genome as it encompasses the complete set of coding and non-coding RNA molecules present in a single cell or a population of cells [1]. The ultimate expression of a subset of genes into complementary RNA transcripts would designate a cell's identity and the control of the biological activities within the cell [2]. Transcriptome profiling therefore can greatly facilitate the understanding of a functional genome via characterization of the gene structures, identification of the alternative splicing events, as well as detection of the dynamic regulation of transcripts in various tissues during development, diseased, or stressed conditions [3].

Ever since they were first introduced in 2005, high throughput next-generation DNA sequencing (NGS) technologies have revolutionized the transcriptomics field through massively parallel sequencing of complementary DNA (cDNAs) derived from a transcript population. This important application of NGS termed RNA-sequencing (RNA-seq) [4,5] has overcome several limitations posed by generally used microarray technologies, including not requiring prior knowledge of the genome or sequence of interest, which enables genome-wide unbiased detection of both known and novel transcripts [3]. Single nucleotide-resolution RNA-seq data can also enhance the detection of alternative splicing events and isoform expression. Reanalyzing RNA-seq data in relation to any new genome or datasets that become available in future also can be easily achieved [6]. On the other hand, microarrays inherently exhibit cross-hybridization results in high background noise and have a limited dynamic range of detection, for example in identification of low-abundance transcripts [7]. Due to its distinct advantages and rapid decrease in the per-base costs, together with the application of multiplexing strategies, RNA-seq methods have mostly displaced hybridization-based methods as the preferred option for gene expression studies [6]. With constantly improving RNA-seq techniques and platforms for bioinformatics analysis, RNA-seq has been widely adopted in the analysis of both prokaryotic and eukaryotic transcriptomes as in the studies of bacterial pathogens [8–10], livestock [11–14], and human cancer and disease [15–18].

Since the initiation of the oneKP project, which aims to sequence 1000 of plant transcriptomes, RNA-seq has been extensively applied to transcriptome studies of a wide range of economically important crop plants [19–22]. Moreover, integration of RNA-seq with different molecular biology and biochemical techniques has allowed deeper exploration of various aspects of the transcriptome in plants, such as miRNA-seq [23], Ribo-seq [24], HITS-CLIP/CLIP-seq [25], and GRO-seq [26].

Protein–DNA binding interactions play key roles in gene regulatory and expression processes such as replication, splicing, transcription, and DNA repair. To predict the accuracy of modified histones and bound proteins, functional assays were developed, such as electrophoretic shift mobility assays (EMSA) [27,28], DNA microarrays [29], yeast one-hybrid studies [30,31], and chromatin immunoprecipitation, followed by microarray, also termed ChIP-chip [32]. Chromatin immunoprecipitation, followed by sequencing (ChIP-seq) assays, have become an indispensable next generation technique for detecting in vivo interactions of DNA target sites against their corresponding transcription factors (TFs), epigenetic histone modifications, as well as chromatin remodeling. Chromosome structure and function is largely determined by nucleic acids interactions with specific proteins [33]. ChIP-seq is, so far, the best technique to study these interactions because of its improved signal-to noise ratio and genomic sequence information [34]. ChIP-seq nomenclature has been reported in different forms to suit different investigators' research goals. For instance, ChIP quantitative polymerase chain reaction (ChIP-qPCR) was developed to be a robust method to analyze ChIP-data via different normalization strategies [35]. In contrast, Nano-ChIP-seq has been used to study protein DNA interactions where little source of DNA is available [36], which is necessary because ChIP-seq was originally proposed to use a large number of cells.

The 'big data' generated by many high-throughput technologies often tend to be noisy and contains various sources of unwanted variance and procedural artifacts. It is a challenge for the accurate analysis of extraordinary data volumes to identify true signals, combine variable data types, and understand their relationships [37]. When designing integrated Omics (ChIP-seq and RNA-seq) experiments, RNA-seq can be performed prior to ChIP-seq. In this way, the most enriched TF in differentially expressed genes (DEGs) revealed by RNA-seq, such as in studies involving biotic or abiotic stress treatments, are considered as targets for ChIP-seq assay either by raising custom antibodies against the TFs [38] or through transgenic expression against the tagged TFs (tags like FLAG, green fluorescent protein (GFP), and Glutathione S-transferase (GST), etc.) in model plants [39]. Moreover, independent ChIP-seq can be carried out based on TFs that have substantial literature information and subsequently comparing with an independent RNA-seq assay under the same biological treatment on the same plant. Combining ChIP-seq and RNA-seq assays can show agreement between both findings, revealing more information about a TF by either discovering a new function or a new set of genes for the same function [40]. Assays of transposase accessible chromatin [41] (ATAC-seq) can also measure how much chromatin can be accessed for peak enrichment from ChIP-seq assay [42] and therefore, it can be accompanied with ChIP-seq assay.

In this review, we will discuss how ChIP-seq can strengthen information generated from RNA-seq in elucidating the role of transcription factors. To be precise, we discuss how a combination of ChIP-seq and RNA-seq data can help to unravel the transcriptional regulatory network. RNA-seq essentially serves as the gene discovery tool for identifying specific transcription factors based on their expression profiles and the profile of potential target genes. The ChIP-seq is potentially useful to validate transcription factor target (downstream) genes interaction with potential link with certain physiological or biochemical processes. This review also discusses the transformation of RNA-seq and ChIP-seq assays over time, together with a review of the basic steps required for plant system, highlighting the most recent applications in different plant species. We also review the basic characteristics of RNA-seq and ChIP-seq data analysis pipelines. We then provide examples of genome-wide identification of transcription factor co-regulated genes by RNA-seq and ChIP-seq, which highlight the potential of such studies in elucidating transcriptional regulatory network in important biological processes in plants. These examples will show how combining these tools will help in addressing hormonal response like jasmonic acid in Arabidopsis [1], gibberellic acid in rice [2], and the developmental stage effect in maize [3] to reveal some important insights on their transcriptional regulatory mechanisms. We also introduce the third-generation sequencing, which expands the application of sequencing technology due to the longer read length offering higher capability in sequence assembly and identifying sequence variance in RNA-seq.

2. RNA-seq Platform Selections

There are several commercially available deep sequencing platforms for RNA-seq, such as Ion Torrent, PacBio, and Illumina [43]. Currently, the HiSeq series of sequencers from Illumina is the most widely deployed sequencing platform due to its ability to produce a high data output with low sequencing errors. In view of the variation in data quality and quantity achieved from different deep sequencing platforms and the downstream interpretation processes, the selection of a suitable sequencing platform based on the research goals is an initial key step before starting an RNA-seq experiment. For instance, Illumina Hiseq can produce short reads (50–250 bp), while PacBio generates longer reads (4200–8500 bp). Longer reads will ease the de novo transcriptome assembly process and the detection of alternative splice isoforms compared with short reads. Additionally, paired-end reads (sequencing from both ends of a fragment) are attainable with Illumina instruments but not with Ion Torrent [2,43,44]. Paired-end reads uncover sequence from both ends of the cDNA fragment and accelerate the inspection of splicing variants, chimeric transcripts, and indels [45]. Figure 1 shows the summarized RNA-seq workflow comprising of the wet laboratory works (RNA extraction, library preparation, and sequencing) and the dry laboratory works (in silico RNA-seq data analysis). We will go through each of these steps in more detail in the text below.

In general, RNA-seq experiments start with total RNA isolation and selection of a specific RNA population, such as messenger RNA (mRNA) or microRNA (miRNA), before subjecting samples to a fragmentation step. Next, the short RNAs are converted to a cDNA library and each cDNA fragment is ligated with platform-specific adaptors at one or both ends in order to capture the fragments on a solid support. Millions of reads from one end (single-end sequencing) or both ends (pair-end sequencing) are retrieved by parallel sequencing of millions of cDNA fragments in different NGS platform. Read lengths can vary depending on the sequencing chemistry and technology. The resulting reads will either align to a reference genome or transcriptome or de novo assembled to produce a genome-wide transcription landscape [6,46]. The comprehension of the generated data to resolve the primary research questions characterize the success of an RNA-seq study. Thus, it is crucial to consider upfront several key points before conducting an RNA-seq experiment, such as the selection of the library type, the number of biological and technical replicates, and the depth of sequencing across the transcriptome [47].

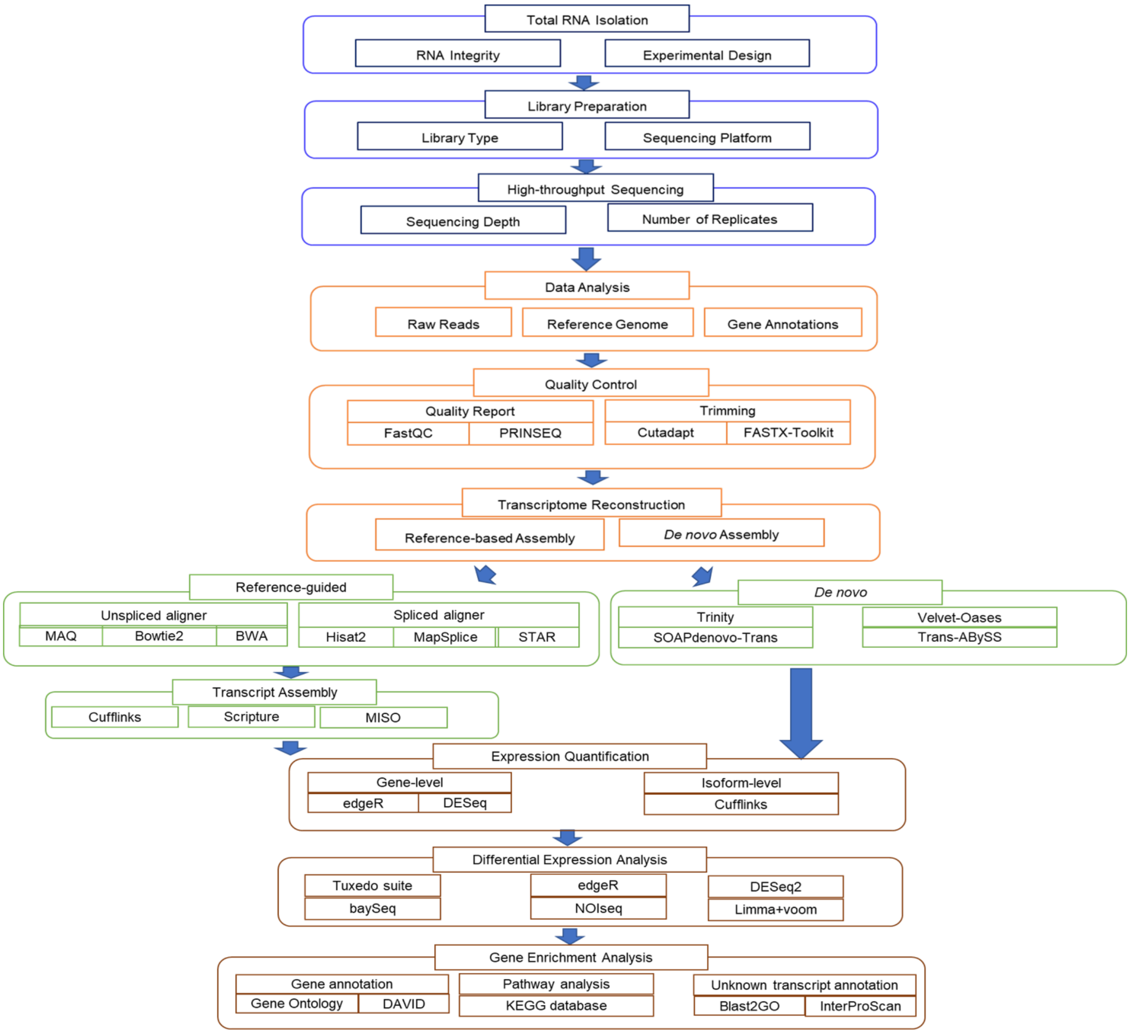


Figure 1. General RNA-seq analysis pipeline. The workflow typically starts with total RNA extraction depending on experimental design and RNA integrity. The library preparation step relies on the selection of sequencing platform and library type, while sequencing depth and number of replicates can impact the downstream sequencing output analysis processes. RNA-seq data analysis generally requires inputs such as raw sequencing reads, reference genome sequences, and gene annotations. Next is examination of raw data quality and perform poor read trimming, transcriptome assembly, and expression quantification. Finally, differential expressed genes (DEGs) must be identified and interpreted through gene enrichment analysis. Each step in the data analysis has several representative tools, as highlighted.

Determination of the number of biological and technical replicates required in an RNA-seq experiment varies with the technical biases and the heterogeneity of each experimental system. While reproducibility of RNA-seq data across lanes and flow cells is generally high, biological replication is mandatory for population inferential analysis [48]. Optimal data interpretation can be achieved by reducing the data variability with duplicate or triplicate experimental datasets [49]. The optimal sequencing depth, which means the number of sequencing reads for a given sample, is strongly governed by the aims of the study. Generally, the required sequencing coverage depends on several factors, including reference genome size, gene expression level, and specific application of interest using the data generated. In this review, we will provide an overview of a typical RNA-seq experiment and steps involved in the bioinformatics analysis.

3. RNA-seq Workflow (Wet Laboratory)

3.1. Total RNA Isolation

High quality RNA is a prerequisite for a successful RNA-seq experiment. RNA integrity number (RIN), a measuring unit produced by Agilent Bioanalyzer, is an unofficial standard used to estimate the integrity of RNA before proceeding with library preparation step. RIN ranges from 1 to 10, with 10 being the highest score for samples with minimal degradation. RIN < 6 indicates low quality RNA that can introduce substantial biases into the final sequencing results [2]. But for plant materials, a good RIN number can be lower depending on species and tissue types. For fluorometric quantification of the RNA input, Thermo Fisher Scientific Qubit or Nanodrop is the most commonly used fluorometer [50].

3.2. Library Preparation

The second step in RNA-seq is the construction of an RNA-seq library. It starts with enrichment or depletion of the total RNA pool for that desired RNA species. In most cell types, RNA can be divided into different populations comprised of ribosomal RNA (rRNA), transfer RNA (tRNA), non-coding RNA (ncRNA), and messenger RNA (mRNA), which is the common interest in most transcriptome studies. Deep sequencing without removal of rRNAs that occupy 80% of the RNA population will reduce the depth of sequence coverage and limit the detection of lowly expressed transcripts [47]. The common practice before library construction is to enrich the mRNA by selecting the poly(A) RNAs. Poly(A) RNA can be isolated using magnetic or cellulose beads coated with poly-T oligos. Alternatively, rRNA depletion can be carried out through duplex-specific nucleases treatment or commercial kits such as Ribo-Zero (Illumina), NEBNext® (New England BioLabs) or RiboMinus (Thermo Fisher). The technical limitations and biases of each approach need to be discerned in order to choose the most appropriate method for library preparation. For example, if one aimed at the exploration of noncoding RNA including pre-mRNA, then ribo-depletion libraries are a more appropriate choice than poly(A) libraries. In recent years, small RNA which includes microRNA (miRNA), small interfering RNA (siRNA), and piwi-interacting RNA (piRNA) have gained great interest among plant researchers to profile and characterize their functions in post-transcriptional regulation. Therefore, several commercially available isolation kits have been developed to capture these short and lowly abundant transcripts based on size fractionation method through gel electrophoresis or silica spin columns [2].

Following poly(A) RNA selection or rRNA removal, RNA molecules need to be fragmented into appropriate sizes (120–200 bp) by enzymatic digestion or chemical hydrolysis under an elevated temperature. In the case of small RNAs, no fragmentation step is needed, and one can directly proceed with adaptor ligation. Once the RNA is cleaved, samples are reverse transcribed into first strand complementary DNA (cDNA) using random primers. After synthesis, the second strand cDNA using DNA polymerase I and RNase H, a single 'A' base is added to the end of each cDNA fragments before ligating with the sequencing adapters. The cDNA pool is then purified and amplified to form the final sequencing-ready cDNA library [47]. By having the sequencing adapters ligated to both ends of the cDNA, researchers may perform paired-end sequencing, which sequences the cDNA from both directions (forward and reverse) to produce more reliable sequencing data compared to single-end sequencing.

In the standard library construction protocol described above, the information about the strand orientation of each transcript is lost, which could complicate the identification of overlapping genes transcribed from the opposite strand and particularly in de novo transcript discovery. Subsequently, this could mislead the quantification of global expression of both sense and antisense RNAs [51]. The preferred approach to retain the strand origin is by incorporating deoxy-UTPs (dUTPs) instead of dTTPs during the second strand cDNA synthesis step, which can be selectively digested using uracil-*N*-glycosylase (UDG). Eventually, the remaining first strand cDNA is amplified to yield a strand-specific cDNA library. Zhao et al. (2015) [52] demonstrated that stranded RNA-seq could provide better resolution in estimating the relative abundance of overlapping transcripts expression as compared with the conventional non-stranded RNA-seq.

Through the implementation of barcoding strategy, one can carry out multiplexing of several samples in an analysis which could significantly reduce the per sample cost for large scale projects. Van Nieuwerburgh et al. (2011) [53] compared three different barcoding methods, including pre-PCR, TruSeq, and PALM. For pre-PCR method, the barcode is associated with the 5′ RNA adapter and ligated to the RNA template before performing RT-PCR, while the barcode is incorporated into one of the RT-PCR primers during the library amplification step in TruSeq method. On the contrary, PALM barcoding method ligated the T-tailed barcode adapter to the A-tailed RT-PCR products after the library amplification step to produce a library that is free of barcode-induced PCR bias.

4. RNA-seq Workflow (Data Analysis)

After completing the sequencing, the next challenge is dealing with millions of reads generated from each experiment. The conventional analysis pipeline of the RNA-seq data starts with quality checks and preprocessing of the raw sequencing short reads, followed by mapping of the filtered reads to a reference genome sequence or de novo assembly using different de novo transcriptome assemblers. Gene expression levels of all mapped transcripts are then quantified and normalized to define the differential expressed genes. Further analysis of the listed genes, such as alternative splicing analysis, functional annotation, and pathway enrichment analysis, can be carried out using a range of bioinformatic programs. Specialized data analysis workflows can be designed according to individual experimental setups and research aims. In this section, we will briefly discuss the routine RNA-seq data analysis pipeline and related bioinformatics tools in each step.

4.1. Quality Control

The preliminary sequencing output is supplied in FASTQ format and is generally contaminated with sequencing artefacts and errors which may arise in library preparation, sequencing, or imaging steps that can ultimately lead to misinterpretation and erroneous conclusions. Therefore, pre-processing and quality control of the raw reads data is mandatory to improve downstream assembly quality and computational efficiency [54]. A Phred quality score (Q score) was assigned to estimate the base call accuracy of the sequencing output. Q30 corresponds to an incorrect base call of 1 in 1000 (99.9%) and serves as the gold standard for quality in read data. Publicly available tools such as FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and PRINSEQ (http://prinseq.sourceforge.net/faq.html) can be used to generate summary statistical reports for sequencing outputs including GC content percentage, base quality and content, level of duplication, sequence quality scores, presence of ambiguous bases, etc. Based on the quality report, further removal or trimming of poor-quality reads, adapter sequences, or demultiplexing can be performed using software tools likes Cutadapt (https://cutadapt.readthedocs.org/en/stable) and FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit) [1,3].

4.2. Transcriptome Reconstruction

The process of identifying all of the transcripts and isoforms that are expressed in a specimen through assembly filtered short reads or read alignments into transcription units is defined as transcriptome reconstruction. Transcriptome assembly can be done using two different strategies: reference-based assembly or de novo assembly. The reference-based approach involves mapping the filtered sequencing reads to an annotated genome or transcriptome followed by transcript assembly. This is relatively less computationally intensive compared with de novo assembly. However, the de novo assembly approach is particularly beneficial when a reference genome or transcriptome is not available. The analysis starts with assembling the sequencing reads into contigs, which will be used as a novel reference transcriptome to align with the raw reads again [3].

4.3. Reference-Guided Assembly

Once aberrant reads are eliminated, the RNA-seq data is ready for alignment, with the condition that a reference genome or transcriptome is available. The national centre of biotechnology information (NCBI; https://www.ncbi.nlm.nih.gov/genome/), Ensembl (www.ensembl.org/index.html), and UCSC genome browser (https://genome.ucsc.edu/) are the three most well-known publicly available resources for retrieving the reference genomes and annotation files for a variety of species. There are two major categories of computational programs that have been developed to map the millions of short query sequences to a reference genome or transcriptome precisely with appropriate parameter values at each step of the analysis. The first group is referred as unspliced aligners which includes MAQ [55], Bowtie2 [56], and Burrow Wheelers alignment (BWA) [57], which is a better option for prokaryotic RNA-seq analysis, and in reference, transcriptome mapping as splicing event detection is unnecessary. In contrast, spliced aligners such as HiSat2 [58], MapSplice [59], and STAR [60] are extensively applied in mapping query sequences to the reference genome of eukaryotes. This group of aligners possesses the ability to identify the exon boundaries and align the query sequences that span across introns, which consequently increases the possibility for alternative splicing detection. Despite understanding the intrinsic alignment algorithms, computational infrastructure requirements for each mapping tools to complete the tasks should also be taken into consideration. Typically, the aligned read data is presented in SAM file format and then compressed into binary of SAM (BAM) file format. The alignment file can be viewed and manipulated using SAMtools (samtools.sourceforge.net/) and Picard (https://broadinstitute.github.io/picard/). Integrative genomic viewer (IGV) (http://software.broadinstitute.org/software/igv/) is a high-performance viewer that supports diverse file formats, e.g., SAM, BAM, and Goby. In addition to the ability to display varying level of alignment details depending on the resolution scale, IGV is also able to simultaneously display multiple genomic regions in an adjacent panel [61]. Towards the completion of the alignment step, one can assess the quality of the mapping result using tools like Qualimap 2 (http://qualimap.bioinfo.cipf.es/) and RSeQC (http://rseqc.sourceforge.net/) considering several metrics including percentage of mapped reads, error distributions, and 3′–5′ coverage ratio [2,3,47]. Subsequently, the overlapping reads can be assembled into full length transcripts using RNA-seq analysis packages such as Cufflinks [62], Scripture [63], and MISO [64]. This assembly method is more advantageous in the discovery of low expressed transcripts and alternatively spliced isoforms. However, the success of the assembly is dependent on the quality of the reference sequence being used. Large genomic deletions and mis-assembly of a genome will sequentially propagate into a misassembled or partially assembled transcriptome [54].

A reference-based assembly strategy was extensively being applied in RNA-seq analysis, especially for plant species with an established genome sequence available. *Arabidopsis thaliana* reference genome (TAIR10) has contributed in the transcriptome data alignment and also in the mapping of ChIP-seq data [65,66]. By mapping the RNA-seq reads against Arabidopsis genome (TAIR10), Pajoro et al. (2017) [4] have successfully identified the temperature-induced differentially spliced events in Arabidopsis plants after being exposed to different temperatures. Subsequently, they were able to detect a total of 59,736 regions to be enriched in H3K36me3 after using similar reference genome for the mapping of FASTQ files generated in ChIP-seq. Integration of the RNA-seq and ChIP-seq datasets revealed that the H3K36me3 histone mark was overrepresented in differentially spliced event genes, and reduction in the H3K36me3 mark deposition could affect the temperature-induced alternative splicing.

4.4. De Novo Assembly

For species lacking a sequenced genome, de novo assembly of the overlapping reads can be employed using one of the several assemblers, including Trinity [67], SOAPdenovo-Trans [68], and Trans-ABySS [69]. All the de novo assemblers listed above are developed by referring to de Bruijn graph algorithms, which broke the reads into k-mer seeds to construct a unique de Bruijn graph and then parsed into consensus transcripts. Annotation of the consensus transcripts can be achieved by mapping to a genome or alignment to a gene or protein database [70]. There are several general metrics for assessment of the de novo assembled transcriptome quality, such as assembly statistics, contigs statistics, mis-assembly statistics, number of contigs matching with the closest related genome, and number of hybrid transcripts [3]. Typically, de novo assembly of large transcriptome is challenging and requires much higher sequencing depth for better assembly output [54]. Nevertheless, the de novo assembly method still possesses certain merits against reference-guided assembly method in discovery of novel transcripts caused by missing genes or structural variants, identification of transcripts with long introns, and in detection of rare events like trans-splicing and chromosomal rearrangements [71].

4.5. Expression Quantification and Normalization for Differential Expression Analysis

Following transcriptome assembly, transcript expression can be quantified by counting the reads mapped to each coding unit including exon, gene, or transcript [72]. For single-end reads, the reads per kilobase of transcript per million mapped reads (RPKM) metric is introduced to remove the feature-length and library-size effects through dividing the number of read counts by both its length and total number of mapped reads. Fragments per kilobase of transcript per million mapped reads (FPKM) is the metric derived from RPKM which is applicable for paired-end reads data and considers a fragment (not reads). Together with transcripts per million (TPM), RPKM and FPKM are the most frequently reported values for transcript abundances in RNA-seq [3,47,70]. Although RPKM/FPKM is a popular choice in place of read count, its value in a sample can be significantly altered by the presence of several highly expressed genes which will "consume" many reads and subsequently underestimated the remaining genes, particularly lowly expressed genes [3]. Wagner et al. (2012) [73] demonstrated that RPKM has the potential to cause inflated statistical significance values due to its inconsistency between samples, which arises from the normalization by the total number of reads. HTSeq (https://pypi.python.org/pypi/HTSeq) is a Python library that contains a stand-alone script *htseq-count* which can count the number of aligned reads mapped to a single gene while discarding multi-mapping reads. These counts can then be used as input data for gene-level quantification using methods such as edgeR or DESeq [74]. The major challenges in read quantification is to quantify multi-mapping reads because of genes with multiple isoforms or close paralogs. In order to address this problem, several algorithms were developed to allow isoform-level quantification. Alternative expression analysis by sequencing (ALEXA-seq) estimates isoform abundances by counting the reads that mapped uniquely to a single isoform, but this method is not suitable for genes lacking unique exons [70]. Alternatively, Cufflinks will quantify isoform abundances by constructing a likelihood function that models the sequencing process to estimate the maximum likelihood that the read maps to an isoform and reports in FPKM or RPKM values [2].

Throughout the RNA-seq experiment, there will be various biases and variances being incorporated which involves intra-sample differences such as differences in length, GC content, or inter-sample differences, for example, differences in sequencing depth, sampling time, and so on [3]. These variations should be eliminated to improve the accuracy of the statistical analysis applied for inferring differential expression. Previous studies have demonstrated that the choice of normalization procedure can impact on the result of differential expression analysis and emphasizes the requirement for normalization [48,75]. Sequencing depth of a sample is one of the major sources of biases in RNA-seq data, therefore trimmed mean of M-values (TMM) and median of ratio approach by assuming most genes are not differentially expressed have been proposed [76]. A comparison study involving seven normalization methods demonstrated that TMM and median of ratio are the two most robust normalization approaches for library size normalization after testing with simulated and real RNA-seq data [77]. The TMM approach has been implemented in R/Bioconductor packages edgeR [78], while the median of ratio approach has been implemented in R/Bioconductor packages DESeq [79], DESeq2 [80], and in Cuffdiff2 [81].

One of the most routinely used analyses conducted using RNA-seq data is to identify differentially expressed genes (DEG) among phenotypes and experimental conditions, and hence, a number of complex statistical methods have been designed to perform this task. Tuxedo suite, a suite of tools for transcript assembly and quantification comprises of Bowtie, TopHat, and Cufflinks packages. TopHat utilizes Bowtie as an alignment "engine" to map millions of RNA-seq reads to the genome and these read alignments serve as input for Cufflinks to produce a transcriptome assembly for each condition. The assembly files are then merged together using the Cuffmerge and fed to Cuffdiff to detect DEGs and genes that are differentially spliced or differentially regulated via promoter switching across multiple conditions. Data generated by Cuffdiff analysis can be visualized and explored with CummeRbund [82]. Additionally, there are several other software tools that support DEG analysis such as edgeR, DESeq2, baySeq [83], and NOIseq [84]. EdgeR software uses an over-dispersed Poisson model to account for biological and technical variations in replicated data, and subsequently applied an empirical Bayes method to alleviate the degree of overdispersion across genes. Lastly, differential expression analysis is performed using either quasi-likelihood (QL) F-test or likelihood ratio test [85]. While DESeq2 is adapted from DESeq with the critical enhancement by incorporating empirical Bayes shrinkage estimators for dispersion and fold change, which facilitates a sound and statistically well-founded differential expression analysis across a wide dynamic of RNA-seq experiments. Besides, through the implementation of shrinkage of fold change on a per-sample basis termed as rlog transformation eases the visualization of differences in heatmap and the application of numerous downstream techniques, including principal component analysis and clustering, in which homoscedastic input data is needed [80].

Costa-Silva et al. (2017) [86] evaluated the impact of six mapping and nine differential expression analysis methodologies on real RNA-seq data and adopted qRT-PCR data as reference. The results indicated that mapping methods have minimal impact on the expression analysis result and highlighted that the DEGs identification method is the main choice for differential expression analysis. Based on the adopted experimental model, NOIseq, DESeq2, and limma + voom [87] are the most balanced DEGs identification software by considering the precision, accuracy, and sensitivity. However, there is no consensus on the best-suited differential analysis method for all circumstances.

4.6. Functional Annotation and Pathway Analysis

The final step in a standard transcriptome analysis pipeline is often the interpretation of the gene expression data through gene set enrichment analysis. The analysis would favour the characterization of the functional annotation of the listed DEGs and their associated biological pathways or molecular function in order to infer biological insights from these genes. Publicly available resources, such as Gene Ontology [88] and DAVID [89], containing annotation databases of gene products for most model species are commonly used for gene annotation purposes and also would allow identification of functional information across orthologs [47]. On the other side, multiple listed DEGs may have interactions with each other and be involved in certain biological pathways. KEGG (Kyoto encyclopedia of genes and genomes) pathway database provides a valuable resource for investigating significantly enriched biological pathways associated with the listed DEGs [90]. MapMan4 [91], which is the latest version of MapMan framework coupled with the revised Mercator4 online tool, provides another option for protein classification and annotation task of any land plant. With the triple increased total number of bin categories, MapMan4 has been improved to perform more precise protein descriptions for all assignments through a leaf node category. The prediction of protein–protein interaction network would facilitate the understanding of cellular processes and annotation of structural and functional properties of proteins. This analysis can be performed using the STRING database (http://string-db.org/), which is a web resource of known and computational predicted protein interactions [92]. During the de novo transcriptome reconstruction, there will be a number of unknown transcripts being discovered, and Blast2GO can be used for homologous gene identification through GenBank BLAST or InterProScan and assigning gene ontology terms to each locus [93].

5. ChIP-seq Workflow (Wet Laboratory)

ChIP-seq assay is a powerful tool used to determine nuclear protein interactions with DNA that is usually applied in the context of disease diagnostics, gene expression, and cell differentiation in animal systems for personalized medicine development. Plant scientists have now adopted the technique to better understand various in vivo epigenetics changes and discover genes expressed in a certain biotic/abiotic stress response (that is protein–DNA interaction) in plant systems. The encyclopedia of DNA elements (ENCODE) is now the largest database of sequencing-based techniques, including ChIP-seq. The database has a massive amount of information limited to four different animal species, namely; human, mouse, worm, and fly. ENCODE encodes ChIP-seq overview information covering experimental design to data analysis and contains some published standards to achieve each step in ChIP-seq analysis. Plant researchers can employ the same ENCODE standards. For instance, the ENCODE manual [94] reports step-by-step methods for primary and secondary characterization of protein/antibody, which can also be applicable to plants. Western blot and immunoprecipitation are amongst the primary methods, while secondary methods are subordinates to the primary methods and they use previously characterized antibodies for ChIP-seq, epitope-tagged expression pattern, motif analysis, etc.

For plants, the ChIP-seq protocol usually takes about 3–7 days to completion, starting from nuclei extraction to immunoprecipitation. There might be changes in some steps which are geared towards reducing time consumption and simplifying the tedious nature of the technique. In this section, major steps of ChIP-seq will be reviewed starting from formaldehyde fixing of plant sample, chromatin isolation, and to data analysis (as shown on Figure 2).

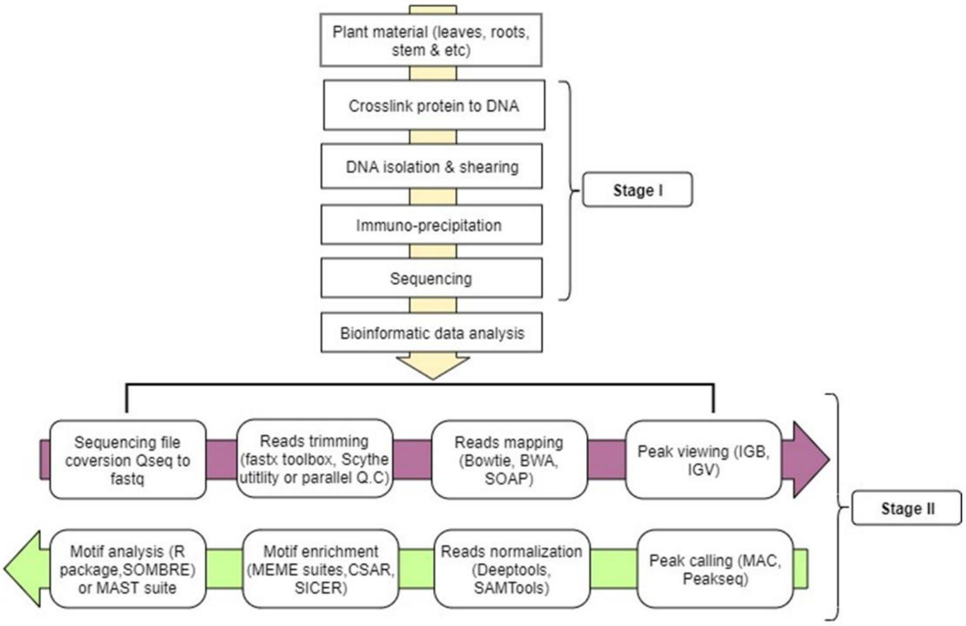

Plant material (leaves, roots, stem & etc)
Crosslink protein to DNA
DNA isolation & shearing
Immuno-precipitation
Sequencing
Stage I
Bioinformatic data analysis
Sequencing file coversion Qseq to fastq
Reads trimming (fastx toolbox, Scythe utitlity or parallel Q.C)
Reads mapping (Bowtie, BWA, SOAP)
Peak viewing (IGB, IGV)
Stage II
Motif analysis (R package,SOMBRE) or MAST suite
Motif enrichment (MEME suites,CSAR, SICER)
Reads normalization (Deeptools, SAMTools)
Peak calling (MAC, Peakseq)

Figure 2. Basic steps involved in ChIP-seq: stage I starts from crosslinking to sequencing and stage II involves steps for gene mining. Sequenced file Qseq are converted to fastq format using fastx tools, reads undergo trimming and filtering using Scythe utility or parallel Q.C, then reads alignment using Bowtie or Burrow Wheelers alignment (BWA), matched reads viewing is aided by integrative genome browser (IGB). Peaks are called using any available software like MAC/Peakseq, reads are normalized by removing duplicate reads and searching for tag densities in a window of reads per kilobase per million reads (RPKM) around the reference peak, mostly 1 kb upstream of transcription start site (TSS) to transcription end site (TES), with SAMTools, motif search using MEME suites, and finally predicts gene through MAST suite or R statistic package SOMBRE with the aid of GO and transcription factor databases like JASPAR.

5.1. Crosslinking in Plant Samples

Formaldehyde is a small (2Å) dipolar compound that can entrap protein–protein and protein–DNA complexes in vivo. Its small size makes it the best candidate for capturing macromolecular interaction that are close to one another [95]. Its carbon atom plays a role as a nucleophilic center. Amino and imino functional groups of DNA (adenine and cytosine) and of some amino acids (arginine, histidine, and lysine) readily react with formaldehyde to form a Schiff base intermediate. This can also react to another amino group to form the final crosslinked protein–DNA complex [96].

Formaldehyde is now the crosslinking agent of choice in ChIP-seq protein–DNA binding due to its robustness, reversibility, and less hazardous nature compared to use of ultraviolet (UV) radiation as a cross linker method [97,98]. For example, Haring et al. (2007) [35] used 3% formaldehyde to crosslink protein to DNA region in Maize (*Zea mays*) while subsequent publications used 1% formaldehyde in *Arabidopsis thaliana* [99,100]. Hoffman et al. (2015) [97] has reviewed formaldehyde binding chemistry and showed a two-stage stoichiometry mechanism of crosslinking protein–DNA with formaldehyde and quenching with glycine. Figures 3 and 4 show the two-step forward chemical reactions of crosslinking and quenching of excess crosslinking agent respectively. Reversal of protein–DNA crosslinking is typically achieved by heating (usually 65 °C) in the presence of high salt concentrations (example 5 M NaCl and 20% SDS) [35,101] or treating with proteinase K at 37 °C [100].

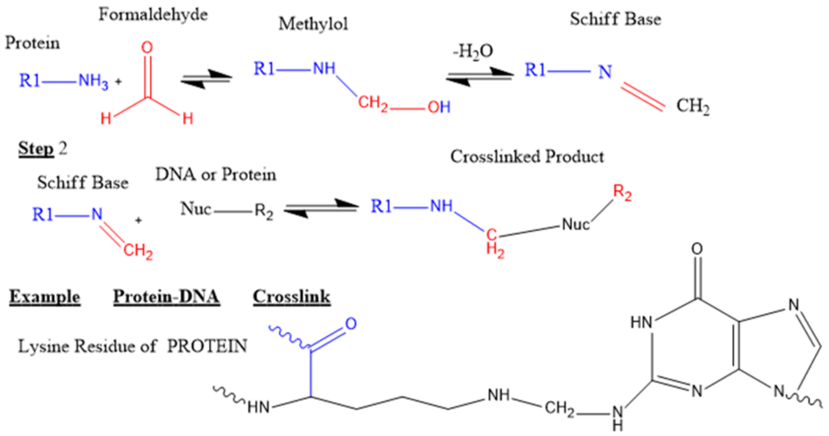


Figure 3. Chemical reactions of protein–DNA crosslinking by formaldehyde: Crosslinking of protein–DNA by formaldehyde occurs in two steps. Firstly, a strong nucleophile, commonly a lysine ϵ-amino group from a protein, reacts with formaldehyde to form a methylol intermediate which will lose water to give a Schiff base (an imine). Secondly, the Schiff base reacts with another nucleophile amine of a DNA to generate a crosslinked product. The latter nucleophile might also be from another protein or the same protein as the first nucleophile. All the reactions in this stoichiometric process are reversible. Modified from Hoffman et al. (2015) [97].

The reversal of protein–DNA crosslinking is normally achieved by heating (at 65 °C) in the presence of high salt concentration (example: 5 M NaCl and 20% SDS) [99,102] or by treating with proteinase K at 37 °C [100].

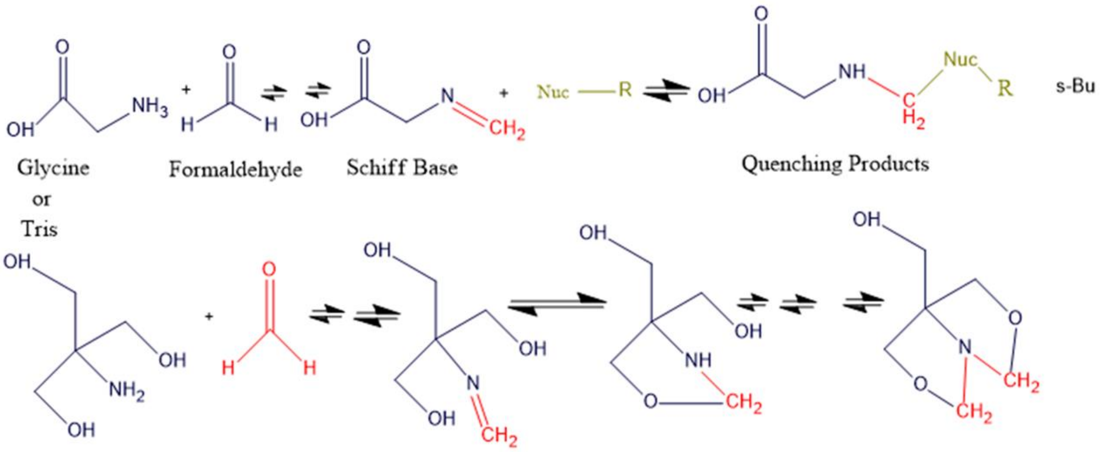


Figure 4. Glycine and Tris quenching reactions of formaldehyde: The chemical reactions are like those shown in Figure 3 above with the amino group of glycine or Tris serving as the principal nucleophile. The Schiff base formed from glycine is not necessary to react with a second nucleophile, but regardless, the crosslinking between protein–DNA will be quenched. The Tris molecule has another available nucleophile (hydroxyl group) that creates stable intramolecular penta-membered rings. Tris can also react with two molecules of formaldehyde, resulting in the last product shown. The tendency of Tris forming some stable intramolecular products allows it to search for formaldehyde from other molecules and thus enable crosslink reversal. Here, mint green color represents DNA/protein. Reconstructed from Hoffman et al. (2015) [97].

5.2. Chromatin Isolation

The method of nuclei isolation is dependent on the source of plant sample and quantity (required mostly 1 to 5 g). For example, whether it is a high phenolic and carbohydrate content like oil palm and Jatropha, which may need higher concentration or a longer treatment period with cell wall degradation components like Triton x-100 detergent in nuclei extraction buffer [103], and if physical shearing or if the sample is from a delicate plant such as Arabidopsis, in which nuclei can be isolated with mild extraction buffer. A protocol developed by Saleh et al. (2008) [99], also by Kaufmann et al. (2010) [100] and many more, explains the laboratory procedure for chromatin extraction. On the other hand, DNA shearing optimization is quite similar across different plant biology laboratories and it is achieved by sonicating the nuclei in a probe sonicator, or water bath ultrasonicator five times (more or less), 30 s ON and one minute OFF on ice (keep the whole step in cold conditions) until a desired DNA fragment size is achieved, which should be within 100–800 bp [35,100,104]. Shearing can also be achieved using an endo-exonuclease MNase [105], but random shearing is mostly not achieved using MNase due to the present of specific cut sites, and this makes sonication the most preferred shearing method since its DNA defragmentation is random.

Immunoprecipitation is a pulldown assay which involves an antibody designed against protein of interest or against a tagged DNA fragment (tags like FLAG, GFP, yellow fluorescent protein (YFP)) coupled with protein of interest, which is used to pull all DNA bound to the protein tag [98]. Conventionally, chromatin is incubated with 1 to 5 μg antibody overnight [106–108] to appropriately pull down all DNA fragments.

Obviously, ChIP-seq is a difficult immunological assay in plants. The major problems plant scientists are facing includes cell wall complexity of plant cell that requires vigorous disruption to avoid sample loss; high level of polysaccharides and phenolic compounds in plant tissues may be a problem for PCR amplification prior to library preparation; ChIP-grade antibodies selection in plants is limited, and as a result of that, investigators will have to take several months to generate epitope-tagged transgenic lines before ChIP-seq experiments [34]. These problems are not yet solved with the plant researchers, but significant contributions have been made to address some part of the problems such as high DNA recovery [105,109] and production of customized protein-specific antibody [110], which takes a similar period of time as transgenic epitope-tagged antibodies do.

5.3. ChIPped-DNA Purification

After immunoprecipitation, antibody ChIPped isolated DNA is followed by a purification step preceded by de-crosslinking. There are several methods for ChIPped-DNA purification, amongst which Zhong et al. (2017) [105] compared ten commercial kits and observed that phenol-chloroform (Invitrogen; PC) method gives the best DNA recovery. Interestingly, they found PC to yield the best DNA recovery. DNA recovery is important because ChIP usually gives insufficient amount of DNA for library construction and qPCR, and sometimes, PCR of the recovering DNA is required.

5.4. Library Construction

Sequencing library construction is the last stage of bench work for ChIP-seq assay and the subsequent steps will only be in silico. Library preparation is often carried out with the use of commercially available kits. From the recent literature, researchers often use kits such as Illumina TruSeq library protocol [40,111], NEBNext ChIP-seq library preparation reagent for Illumina kit (New England Biolab) [40,112], ThruPLEX DNA-Seq kit (Takara Bio USA) [65], and Ovation ultralow library kit (NuGEN Tech, San Carlos, CA, USA) [113,114]. Usually, for sequencing, reads preparation are either single-ended or paired-ended and can also be both. From the names, sequencing one end or both ends of DNA fragment is referred to as single-end tags (SET) and paired-end tags (PET), respectively. SET is the commonly used option while PET is more precise, less ambiguous in genome alignment, and typically used for repetitive fragments of genome [115]. After all, reads sequencing is commonly supplied by HiSeq 2000–3500 Illumina machine (San Diego, CA. US).

6. ChIP-seq Workflow (Data Analysis)

ChIP-sequencing data contain millions of short nucleotide sequences based on sequencing depth. The depth of the sequence depends on the organism's genome size, the probable binding site's size, and frequencies [116]. For example, 43 million reads are adequate for studying TFs involved in stress response mediated by jasmonic acid (JA) signaling pathway in rice [39] and approximately 25 million reads for analyzing maize endosperm development [117]. To get a reliable result, an ENCODE consortium standard of using two independent biological controls should be followed. This will help to assess replicates' agreement and threshold with the use of irreproducible discovery rate (IDR) [94]. Complexity of ChIP-seq libraries is linked to several factors such as antibody quality, over-cross-linking, amount of material, sonication, or over-amplification by PCR. Hence, the last factor can be corrected by systematic identification and removal of redundant reads, which is implemented in many peak callers because it may improve the specificity of PCR [116].

Large ChIP-seq data output analyses usually employ a stage-wise bioinformatic software pipeline and webtools for proper data interpretation and visualization. The steps include sequence alignment to a reference genome (mapping), peak calling, motif discovery, and interpretation [118], as shown in Figure 2. There are several reviews on ChIP-seq computational analysis encompassing reads quality control to TF motif discovery [119,120].

6.1. Reads Mapping

Alignment of ChIP-seq reads signal in the genome region has three categories: small point source base pairs coverage (known as punctate region with few kilobases) of localized signals, such as transcription factor and broad region of several kilobases that covers a large epigenetic domain like H3K36me3, and a mixed region which covers both transcription site on upstream part of a gene [120,121] and within the downstream part of a gene like RNA polymerase [122,123]. Nevertheless, there is a need for a complex normalization procedure for significant variation distribution coverage among samples, not only relying on the sequencing depth but similarly on library preparation methodological differences and sample disparity [124], as well as chromatin condition of the samples. There are a few ChIP-seq analysis normalizations methods in the public domain to provide identical coverage distribution across samples [125]. A peak is considered if its number of readings is higher than a predetermined cut-off value or if a minimum enrichment value equated to the background signal, is frequently in a genome through a sliding window. Many peak calling algorithms give an approximate calculation of a *p* value for called peaks, height of the peaks, and/or background rank peaks enrichment and a FDR to provide peak list [126]. PeakSeq mapping analysis technique [123] compares control sample and IP threshold factor coverages for two linear regression, a quantile normalization technique proposed by [127], which uses statistical moments for the normalization process [125]. Bowtie [56] indexes a reference genome based on the Burrows-Wheeler transform (BWT) [128] and FM index [129]. Besides these, there are many sequence aligners, but the most widely used are SOAP2 [130], BWA [57], Hisat2 [131], and DANPOS2 suite with Dpeak [132]. Bowtie seems to be the most preferred [110,113,133,134] based on the literature, while DANPOS2 [132] is the least preferred. Likewise, numerous software packages are used for peak calling, but the most popular peak caller is MAC [135], as reported in several publications [136–138].

6.2. Enrichment of Genomic Region

In order to determine TF binding site on a plant's genome, special web-based tools and software packages are designed to help with motif finding analyses. Some of these tools are based on various kinds of algorithms which are statistically dependent. According to the most recent ChIP-seq reports, the following motif enrichment tools are commonly used for genome enrichment: MEME [139], MEME/MAST suite [140], and the most repetitive in literature [38,40,112]; DREME [141], RSAT [142], CSAR Cisgenome [125], SICER package [143], and BEDTools [144]. Table 1 provides a summary of the trend of some ChIP-seq publications from 2014 to 2019 series, highlighting the field of research and emphasizing the type of antibody used.

Table 1. A summarized ChIP-seq research findings highlighting the kind of research and type of antibody used.

Application	Findings	Antibody Type	Refernce
Abiotic factor	Abscisic acid stress (ASR5) a TF binds Sensitive To Aluminum Rhizotoxicity 1 (STAR1) promoter in other to positively response against Aluminum stress in rice.	Anti-ASR5	Arenhart et al. (2014) [145]
Developmental + immunity	ROS and defense responsive genes were repressed by HBI1 indicating defense function of HBI1 and also indirectly plays a role in repressing growth through activation of growth-inhibiting HLH genes. HBI1 was also learned to bind to positive activators brassinosteroids function.	Anti-YFP	Fan et al. (2014) [112]
Genetics	In genetic imprinting, some subsets of genes are expressed according to their parental origin. Paternally expressed genes (PEGs) were associated to maternal-specific H3K27me3.	Anti-H3K27me3	Zhang et al. (2014) [146]
Abiotic + developmental	Abscisic, Stress, Ripening (ASR1) from tomato is upregulated in drought stress which acts primarily in the cell wall.	Anti-ASR1	Ricardi et al. (2014) [134]
Developmental	SQUAMOSA Promoter Binding Protein-Like3 (SPL3) bind GTAC motif of phosphate starvation responsive gene promoters like *PLDZ2*, *miR399f*, and *Pht1;5*.	Anti-HA	Lei et al. (2016) [147]
Developmental	Combinatorial action affect MADS-box transcription factors FLC and SVP in flowering shows gibberellins' processing genes.	Anti-GFP	Mateos et al. (2015) [137]
Genetics	In circadian clock of Maize hybrids, expression of morning-phased genes from binding with ZmCCA1 encourages growth vigor and photosynthesis.	Anti-CCA1	Ko et al. (2016) [110]
Photosynthesis	Discovers E-box variant binding motif for Phytochrome interacting factor 4/5 (PIF4 and PIF5) in Cryptochromes (CRYs) during exposure to low blue light and CRY2 association with PIF4/5.	Anti-HA	Pedmale et al. (2016) [148]
Cellular	Shows chromatin domain organization at the nuclei periphery of Arabidopsis. The domain is a clear translation of a repressed environment that contains jumping genes, heterochromatic marks and silenced coding genes.	Anti-GFP	Bi et al. (2017) [149]
Immunity	Using both ChIP-seq and RNA-seq, 655 MYC2 binding were identified in response to Jasmonic acid genes. Also found MYC2 TFs that function in late defense stage.	Anti-GFP	Du et al. (2017) [40]

Table 1. *Cont.*

Application	Findings	Antibody Type	Refernce
Immunity	After flagellin (flg22) treatment, HD2B targets chromatin were hyperacetylated responsible in plant immune defense and phosphorylation while hypoacetylated marks function in metabolic regulation, plastid organization, and chloroplast.	Anti-GFP	Latrasse et al. (2017) [133]
Biochemical	A zinc finger TF of rice ZFP36 inhibits ROS production by binding to ascorbate peroxidase known to have specificity to hydrogen peroxide.	Anti-ZFP36	Huang et al. (2018) [150]
Developmental	Maize GIF in leaves and stems promotes meristematic determinacy and shoot architecture. ChIP-seq has found several GIF1 targets including mostly some transcriptional regulators like UB3, ZMPLATZ5, ZMARR7, bHLH and MYB family members.	Anti-GFP	Zhang et al. (2018) [114]
Developmental	FRUITFULL (FUL), a TF that directly repressed APETALA2 expression promotes meristem arrest and maintains the sequential expression of meristem maintenance factor WUSCHEL.	Anti-GFP	Balanzà et al. (2018) [151]
Abiotic factor	bZIP10 found to be active in Zinc regulation in *Brachypodium* which relate to oxidative stress and a motif homologous to Arabidopsis was found TGDCGACA.	Anti-GFP	Martin et al. (2018) [152]
Abiotic factor	Growth-Regulating Factor 4 (GRF4) TF co-interacts with growth inhibition regulator DELLA to regulate carbon, nitrogen metabolism and growth.	Anti-FLAG	Li et al. (2018) [153]
Abiotic factor	Rice OsTF1L mapped drought related stress and lignin biosynthesis genes.	Anti-MYC and anti-RNA Pol II	Bang et al. (2019) [154]
Epigenetics	Genome-wide ADCP1 is linked with chromosome enrichment site (pericentrome) and co-localization with H3K9me2.	Anti-GFP	Zhao et al. (2019) [155]
General	GmBZL3 is a brassinesteroids signaling molecule cross talking with many pathways like disease-related, immunity response pathways and hormone signaling.	GmBZL3 antibody	Song et al. (2019) [156]
Developmental	A Leucine zipper domain TF FD plays a crucial role in floral transition.	Anti-GFP	Collani et al. (2019) [157]
General	Found new Oryza VIP1 response element (OVRE) cis-element in abiotic and biotic responses.	Anti-FLAG	Liu et al. (2019) [39]

7. Genome-Wide Identification of Transcription Factor Co-Regulated Genes by RNA-seq and ChIP-seq

Shamimuzzaman and Vodkin (2013) [38] were interested to understand early seedling developmental stages in soya bean. They classified the stages into 7 (from pre-emerging hypocotyls to fully grown cotyledons above the ground) and performed RNA-Seq on 7 libraries generated from the different stages. The RNA-seq generated 78,773 mapped reads using ultrafast bowtie, allowing three mismatches. Reads normalization was followed using RPKM and DESeq package to identify the DEGs between developmental stage 3 and 6 at p-value < 0.05. Two TFs, NAC and YABBY, which showed promising expression levels throughout the different stages, were chosen. Consequently, NAC and YABBY antibodies were used to perform ChIP-seq using pooled cotyledons from stage 4 (yellow-green cotyledons 30–35 mm) and 5 (yellow-green cotyledons; starts of primary roots) which is a physiological transition stage between yellow food reserve to a photosynthetic green stage in order to identify their genome-wide binding sites and their co-regulated genes. ChIP-seq data was first aligned using Bowtie to generate 34 million reads and 86 million reads for NAC and YABBY, respectively, at p-value < 0.05, and subsequently MACS was used to call significant peaks, which were 8246 and 18,064 peaks, respectively, for the two TF at p-value = 1.0×10^{-5}**.** Gene location were identified from the soya bean gene annotation using a custom-made Python program. In the promoter, there were 1526 and 974 peaks for NAC and YABBY, respectively. These two TFs play an important role in regulating developmental processes and the sequence similarity analysis between RNA-seq, and NAC and YABBY TFs ChIP-seq data showed 72 genes to be potentially regulated by the NAC and 96 genes by the YABBY.

Opaque2 (O2) TF is involved in maize endosperm development and its mutation *o2* confers better nutrition with 70% higher lysine content (quality protein maize; QPM) than in wild type maize kernels, which were studied. However, the mutant plant exhibited some pleiotropic biological effects that lowers its agronomic quality. Li et al. (2015) [117] performed RNA-seq on both wild-type (WT) O2 and mutant *o2* endosperms. Fifty-five million reads were uniquely mapped to B73 maize genome sequence using TopHat and further normalized as fragments per kilobase of exon per million fragments mapped at p-value < 0.05 using MACS, 52,601 genes were found to be transcribed in both O2 and *o2*. Further analysis narrowed the genes to 3070 in O2 and 6613 genes in *o2*. At last, 1605 genes mRNA steady levels were affected by O2: 767 upregulated in O2 and 838 upregulated in endosperm deficient in O2 function. On the other hand, ChIP-seq assay was performed using O2 custom antibody on wild-type plants, and 15 million reads were specifically mapped using Bowtie2 aligner, while 1686 peaks were mapped using MACS at q-value < 0.05 by comparing O2 and IgG ChIP outputs based on poisson distribution. RNA-seq revealed 1605 DEGs between wild-type and mutant endosperm while ChIP-seq identified 39 genes as O2 putative target. Thirty-five of them were down-regulated in *o2* RNA-seq, while four were upregulated. But none of the DEGs found in *o2* were identified as O2 putative binding target in ChIP-seq, suggesting the potential involvement of non-coding RNA as downstream targets. Combination of the RNA-Seq and ChIP-Seq results had demonstrated the roles of O2 as a central regulator of multiple metabolic pathways related to anabolic functions during maize endosperm development.

Jasmonic acid (JA) mediates activation of plant resistance against insect attack (wounding) and necrotrophic pathogens. MYC2 is a basic helix-loop-helix (bHLH) TF which plays a significant role in orchestrating JA-mediated expression of defense genes. To understand the role of MYC2 TF in tomato, Du et al. (2017) [40] performed RNA-seq on mutant (MYC2-RNAi) and wild-type treated with or without methyl jasmonate (MeJA); wild-type wounded and no wound wild-type. Bowtie2, HISAT2, and TopHat2 were used to align sequencing output to tomato genome SL2.50. Expression levels were determined using eXpress [158] for calculating gene expression levels in all biological replicates at FDR-adjusted p-value < 0.05. DESeq2 was used for mRNA levels quantitation at p-value < 0.05. Pairwise comparisons of RNA-seq data recognized 6544 genes that were DEGs between treatments (with and without MeJA). Two thousand, five hundred and sixty-seven genes showed significant expression differences between untreated mutant and wild-type and 3058 genes showed significant expression differences between mutant treated and wild-type. Additional analyses of these 3058 genes revealed about 40% (2558 from 6544) of the JA-regulated genes also regulated by MYC2. Whereas, ChIP-seq was performed on MYC2-GFP transgenic plants either subjected to MeJA or wounding treatment. Sequence alignment using Bowtie2 and mapping using MACS at q-value < 0.05 were performed. BEDTools with default parameters were used to identify peaks within genic regions and a total of 12–18 thousand putative MYC2 binding peaks from the two biological replicates were identified. The replicates shared 7594 peaks agreement and further overlapped to identify 3389 MYC2-targeted genes. Comparison of ChIP-seq 3389 MYC2-targeted JA-responsive genes (MTJA) and pairwise comparison of RNA-seq data output identified 2258 genes are coregulated by MYC2 and JA. After comparing these two data sets, 655 genes were found to overlap for MYC2-targeted JA-responsive genes (MTJA). The study also identified a group of MYC2-targeted TFs that may have a direct role in regulating the JA-induced transcription of late defense genes. Altogether, it was proposed that MYC2 and the MYC2-targeted TFs form a hierarchical transcriptional cascade during JA-mediated plant immunity responsible for the initiation and amplification of the transcriptional output.

Gibberellic acid (GA) normally promotes plant growth by targeting the destruction of DELLA proteins [158]. Wheat GRV DELLAs mutant [159] is resistant against GA destruction, whereas the rice GRV mutant sd1 allele diminishes bioactive GA abundance [160,161], leading to the accumulation of DELLA protein SLR1. This confers semi-dwarfism and results in yield-reducing lodging. Lodging resistance by GRV increases nitrogen insensitivity associated with nitrogen-use efficiency. Growth regulating factor 4 (GRF4) semi-dominantly increases nitrogen ($NH4^+$) uptake rates and assimilation while SLR1 inhibits these processes. Li et al. (2018) [152] carried out a study through combined Omics (RNA-seq and ChIP-seq) to understand this process. Firstly, RNA sequencing was performed using BGISEQ-500 platform to produce 24 million clean reads mapped to Nipponbare reference genome with HISAT/Bowtie2 tools. The reads were normalized and FPKM was calculated using RSEM software [159]. DEGs were identified using FDR < 0.01 and absolute $\log_2$ ratio ≥ 2 [160]. Four thousand, two hundred and forty-one DEGs were identified between mutant (loss of function) and WT plants, while 4753 DEGs were accumulated between overexpression line and WT. Six hundred and forty-two genes were identified by RNA-seq to be upregulated by GRF4 in a rice overexpressing GRF4 variety and downregulated by SLR1 in sd1 mutant variety. Quantitative reverse transcription PCR (RT-qPCR) shows high abundance of root mRNAs for $NH4^+$ uptake transporters (AMT1.1 and AMT1.2). For the ChIP-seq assay, after BGISEQ-500 sequencing output were mapped to Nipponbare reference genome using SOAP aligner, MACS was used to call potential binding peaks. To define genomic location type, peak summit was used to overlap 100 bp around the top of the peak summit, which was then subjected to DREME motif analysis. Two loci density were drawn using density plot tool in R 3.0 after a likelihood ratio score was considered for each motif in each peak using the basic principles of Bayesian classifier [136]. ChIP-seq revealed potential GRF4 target-recognition sites, with a predominant upstream GGCGGC binding motif common to many nitrogen-metabolism gene promoters. Enriched ChIP-seq DNA through ChIP–PCR confirmed the RT-qPCR finding of GRF4 ammonium transporter AMT1.1 with a putative GCGG- promoter motif.

Albihlal et al. (2018) [39] studied resistance to environmental stress and reproductive fitness (seed yield) regulated by heat shock transcription factor A1b (HSFA1b) protein in Arabidopsis thaliana. To unravel the function of HSFA1b, they surveyed its ChIP-seq target and its significance on its RNA-seq transcriptome of wild type under heat stress (HS) and non-stress (NS), and in transgenic HSFA1b-overexpressing plants under NS. RNA-seq analysis workflow started from sequencing outputs from Illumina HiSeq2000. Reads were mapped to Arabidopsis transcriptome GSNAP (allowing five mismatches). Transcript assembly and DEGs analyses were followed using Cufflinks and Cuffdiff [82] at q-value ≤ 0.05. For wild-type treatment under NS and HS, 7137 DEGs responded to HS: 721 were HSFA1b-bound genes. These bound genes were prevalent in downstream of protein-coding genes suggesting binding to genomic regions or near cis natural long non-coding (*cis*NAT) RNA genes. RNA-seq from HSFA1b-RFP overexpression lines under NS revealed 3306 protein-coding genes showing differential expression when compared with NS WT, and 72% of them were differentially expressed in HS WT. After a Pearson correlation between NS 35S:HSFA1b and both NS (r = 0.92) and HS WT (r = 0.88), heat shock proteins expression levels in 35S:HSFA1b NS plants were found to be intermediate to WT NS and HS plants. Further analyses found a total of 952 HSFA1b-target DEGs and at least 85 of them were developmentally associated and found bound mainly under NS. In addition, 480 natural antisense non-coding RNA (cisNAT) genes bound by HSFA1b were identified, suggesting an additional mode of indirect regulation. On the ChIP-seq analysis, normalized peaks were called with MACS tool and *k*-means clustering analysis of ChIP-seq signals on HSFA1b bound genes and density maps were generated with seqMINER [161]. GO analysis of target features was performed with a singular enrichment analysis (SEA) tool in the AgriGO database [162]. HSFA1b bound region sequences motif were de novo identified by using MEME with *p*-value < 0.0001 and passed through Cistrome atlas database [163]. ChIP-seq identified 1083 and 709 HSFA1b-bound regions under NS and HS, respectively, consisting of 1207 HSFA1b target genes. *K*-means cluster analysis of binding regions identified three groups: specific to NS (group I), common to NS and HS (group II), and unique to HS (group III). After a deep analysis of binding regions in gene annotation features, HSFA1b was found to be preferentially targeted within and downstream of genes in group I (54%) while it was 30% for group II and III genes. NP:HSFA1b ChIP-seq data under HS and NS conditions intersection with the DEGs from 35S:HSFA1b compared with NS WT plants revealed 1821 genes in WT HS plants and in 35S:HSFA1b NS plants that were not bound by HSFA1b. These were designated to regulate HSFA1b indirectly, of which 281 genes were associated with plant development. It was also found that the HSFA1b does not only target the heat shock elements, but also the MADS box, LEAFY, and G-Box promoter motifs. Thus, this suggested that HSFA1b transduces environmental cues to many stress tolerance and developmental genes to enable continuous growth and developmental adjustment by plants in a varying environment under diverse environmental factors.

JA is a plant hormone involved in different plant biological processes such as plant growth, seed germination, response to water stress, wounding, and pathogen attack [40,164]. TF basic region/leucine zipper (bZIP) belongs to a diverse superfamily of TFs divided into 13 groups in Arabidopsis. VIP1, a member of bZIP group I TF, is a bridge between nuclear importin α and VirE which facilitates transport of *Agrobacterium* T-DNA strand into plant nucleus [165]. In addition to its role in Agrobacterium-mediated transformation, it functions in *Botrytis* attack, salt stress, and ABA responses [166,167]. Liu et al. (2019) [39] studied bZIP TF activity using multi-Omics strategy in rice. RNA sequencing output was generated using HiSeq3000 and mapped to *Oryza sativa* reference genome (RGAP v. 7.0), performed on OsbZIP81.1ox and WT ZH11. Gene expression levels and identification of DEGs were carried out using RPKM and edgeR, respectively. Five thousand, one hundred and forty-three DEGs (in OsbZIP81.1ox versus ZH11) and 5002 DEGsOsbZIP81.2ox versus ZH11) were identified. ChIP-seq analysis on OsbZIP81.1 under normal condition yielded 43 million reads after SOAP2 alignment and unique mapping with MACS. These reads were subjected to motif analysis using MEME-ChIP. They carefully analyzed the combined ChIP-seq and RNA-seq data of rice OsZIP81.1 and found 7 genes that were enriched in JA signaling pathway from 1332 genes that were identified. For binding motif discovery by ChIP-seq, they found 15 probable motifs which were referred to as Oryza VIP1 response element (OVRE) GCTG, which are closely related to the Arabidopsis VRE.

8. Third Generation Sequencing

The progress in NGS development has enabled researchers to study and understand the complex world of microorganisms, plants, and animals from broader and deeper perspectives. In the third-generation sequencing technology, platforms were designed to address the limitation in obtaining an effective read coverage, especially in the short read lengths, which are poorly suited for particular biological problems, including assembly and determination of complex genomic regions, gene isoform, and DNA methylation detection [168]. This is due to inherent limitations of the short-read technologies such as GC bias and problems associated with mapping to repetitive regions, differentiating paralogous sequences, and phasing alleles [169]. Long-read/third-generation sequencing technologies revolutionised genomics research as they enable genomes and transcriptomes to be analysed at an unprecedented resolution. It allows direct native DNA and full-length transcript sequencing without requiring sequence assembly. Oxford nanopore and pacific biosciences offer long sequence read technologies commercially. Both use single-molecule sequencing, but with contrasting detection methods based on nanopores and optical detection, respectively. Both provide exceptionally long read lengths, up to greater than 20 kb. These platforms allow sequencing/assembly of repetitive elements, direct variant phasing, and determination of epigenetic modifications [169].

The pacific biosciences RS platform was first released in 2010. It uses hairpin adapters ligated on either end of a DNA molecule to be sequenced, generating capped templates referred to as single-molecule real-time (SMRTbells) [169]. SMRT sequencing is a sequencing-by-synthesis technology based on real-time imaging of fluorescently tagged nucleotides that are incorporated as complementary strand is synthesized along individual DNA template [170]. DNA modifications such as methylation are detected based on the kinetic variation obtained from the light-pulse. The technology allows generation of full-length cDNA sequences without the need for assembly and characterization of transcript isoforms within targeted genes or across an entire transcriptome. It uses a DNA polymerase to drive the reaction and the sequencing reaction ends when the template and polymerase dissociate [171]. The average read length is about 3000 bp, but some may reach 20,000 bp or even longer [172].

NGS methods tend to lose information found in DNA and RNA due to the short-copied reads and the inability to retain modifications. The Oxford nanopore technologies methods that were first commercialised in 2014 can overcome these limitations through direct DNA and native poly(A) RNA sequencing strategy. It does not involve any fragmentation and amplification steps, which are potential source of bias faced in the NGS technology. The sequencing adapter contains the motor protein, an enzyme controlling the passage of the nucleic acid through the nanopore. Read length is directly proportional to the length of RNA and DNA being prepared. Nucleic acid bases determination based on changes in electrical conductivity that are generated as the RNA/DNA strand passes through a biological pore does not require chemical tagging of nucleotides. Sequence data information is produced in real-time, enabling direct data analysis and processing [173]. The long-read nanopore RNA sequencing enables precision characterisation of complete DNA and full-native RNA sequences facilitating sequence assembly and mapping. It enables unambiguous determination of transcript isoforms, giving a true reflection of gene expression with high sensitivity down to single cell level [174].

9. Conclusions and Future Prospects

The identification of DEGs is now widely available via high throughput analysis of transcriptomes. However, the quality of data generated is highly dependent on the experimental design, quality of RNA, and sequencing depth. The information generated will not always provide enough evidence of the specific role of transcription factor, the master regulator at the transcriptional level in regulating certain biological pathway or physiological process, as it only enables inferences based on co-expression of genes. In contrast, genome-wide identification of transcription factors of interest or novel transcription factors through RNA-seq can result in a more in-depth understanding of transcriptional networks associated with these transcription factors and their regulons when followed with ChIP-seq analysis. The ChIP-seq technique can provide valuable information about transcriptional regulation based on transcription factor binding to target DNA promoter motifs for coordinating transcriptional regulation in response to environmental cues, while RNA-seq alone does not provide complete information. However, a combination of these technologies opens up new prospects to better elucidate more comprehensive gene regulatory networks. This approach provides a better explanation of gene regulatory networks [117] and opportunities to explore uncharacterized genes (new genes in a treatment). The approach can give more insights in TFs and the networks they regulate, hence allowing functional study on prospective TFs found in this approach. Although integrating data assembly tools from both ChIP-seq and RNA-seq data like Partek [175] and BETA [176] could be interesting, unfortunately this approach needs to develop the full options of parameter inputs or algorithms required for each method [126]. Advances in the development of new peak calling algorithms, as briefly described in this paper, are providing more reliability and precision concerning the genes detected from high throughput sequencing data [126]. Specifically, in Salmon and Sailfish RNA-seq data analysis, pseudo-counting was introduced at the read counting stage to avoid large negative log transformed values and arithmetic error [177]. This, in combination with third generation sequencing with improved accuracy of sequence assemblies and the discovery of sequence variants, will enhance the potential of combined ChIP-seq and RNA-seq applications that can better describe the functionalities of complex plant genomes and their regulatory networks.

One way to alleviate the problem of TF-specific antibody development is to use clustered regularly interspaced short palindromic repeat associated-Cas9 (CRISPR-Cas9) for epitope tagging. CETCh-seq is designed to tag DNA-binding proteins and subsequently results in using a standard Cas9 antibody [178–180]. This will lead to the creation of individual TF maps and the maps between TF can cross-talk to give a network of TF maps, a new prospect offered by this technology. Likewise, ChIP-seq DNA motifs can be verified using CRISPR single-guide RNA (sgRNA) [181] to target the DNA motif [182]. This will pave a way for precision DNA motif discovery with CRISPR double checking.

Finally, the prime goal for this kind of multi-Omics approach is to comprehend big chunk of data generated from genomic research, which can be a source of confusion and wrong inferences. Combining RNA-seq and ChIP-seq output is like following a reductionist approach from millions of reads of RNA-seq DEGs and thousands of ChIP-seq mapped reads to a few functional TFs binding motif(s). Therefore, the plant community can understand how different experimental studies were designed and approached using multi-Omics technique, and importantly, the analysis part where two-way ANOVA and Python script and R software were employed to aid in clear understanding of data [38,40,153].

Author Contributions: I.I.M.: Conceptualization, writing, review, and editing, S.L.K.: Conceptualization, writing, and editing. S.N.A.A.: Writing, review, and editing, U.M.: Review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Education, Malaysia, Trans Disciplinary Research Grant Scheme (TRGS) (TRGS/1/2016/UPM/01/6/1).

Acknowledgments: The authors would like to thank Universiti Putra Malaysia and Ministry of Education, Malaysia, for providing Trans Disciplinary Research Grant Scheme.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

NGS	Next Generation Sequencing
RNA-seq	RNA Sequencing
RIN	RNA Integrity Number
rRNA	ribosomal RNA
mRNA	micro RNA
siRNA	small interfering RNA
piRNA	piwi-interacting RNA
cDNA	complementary DNA
dUTPs	deoxy-UTPs
UDG	Uracil-N-Glycosylase
SBS	Sequencing by synthesis
BAM	Binary of SAM
IGV	Integrative Genomic Viewer
RPKM	Reads per kilobase of transcript per million mapped reads
FPKM	Fragments per kilobase of transcript per million mapped reads
TPM	Transcripts per million

TMM	Trimmed mean of M-values
DEG	Differentially Expressed Gene
ChIP-seq	Chromatin Immunoprecipitation and Sequencing
UV	Ultraviolet
ChIP qPCR	ChIP quantitative realtime PCR
ENCODE	Encyclopedia of DNA Elements
BWA	Burrow Wheelers alignment
GFP	Green fluorescence protein
YFP	Yellow fluorescence protein
SET	Single-ends tags
PET	Paired-ends tags
JA	Jasmonic acid
CRISPR	Clustered regularly interspaced short palindromic repeats

References

1. Yang, I.; Kim, S. Analysis of whole transcriptome sequencing data: Workflow and software. *Genom. Inform.* **2015**, *13*, 119–125. [CrossRef] [PubMed]
2. Kukurba, K.; Montgomery, S. RNA sequencing and analysis. *Cold Spring Harb Protoc.* **2015**, *11*, 951–969. [CrossRef] [PubMed]
3. Anamika, K.; Verma, S.; Jere, A.; Desai, A. Transcriptomic Profiling Using Next Generation Sequencing-Advances, Advantages, and Challenges. In *Next Generation Sequencing-Advances, Applications and Challenges*; IntechOpen: Rijeka, Croatia, 2015; pp. 7355–7365.
4. Ozsolak, F.; Milos, P.M. RNA sequencing: Advances, challenges and opportunities. *Nat. Rev. Genet.* **2011**, *12*, 87–98. [CrossRef] [PubMed]
5. Bahari, M.N.A.; Sakeh, N.M.; Abdullah, S.N.A.; Ramli, R.R.; Kadkhodaei, S. Transciptome profiling at early infection of *Elaeis guineensis* by Ganoderma boninense provides novel insights on fungal transition from biotrophic to necrotrophic phase. *BMC Plant Biol.* **2018**, *18*, 377. [CrossRef] [PubMed]
6. Hoeijmakers, W.A.M.; Bártfai, R.; Stunnenberg, H.G. Transcriptome Analysis Using RNA-Seq. In *Malaria*; Humana Press: Totowa, NJ, USA, 2012; pp. 221–239.
7. Agarwal, A.; Koppstein, D.; Rozowsky, J.; Sboner, A.; Habegger, L.; Hillier, L.D.W.; Sasidharan, R.; Reinke, V.; Waterston, R.H.; Gerstein, M. Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genom.* **2010**, *11*, 383. [CrossRef] [PubMed]
8. Kröger, C.; Colgan, A.; Srikumar, S.; Händler, K.; Sivasankaran, S.K.; Hammarlöf, D.L.; Canals, R.; Grissom, J.E.; Conway, T.; Hokamp, K.; et al. An infection-relevant transcriptomic compendium for Salmonella enterica Serovar Typhimurium. *Cell Host Microbe* **2013**, *14*, 683–695. [CrossRef]
9. Sharma, C.; Hoffmann, S.; Darfeuille, F.; Nature, J.R. The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature* **2010**, *464*, 250. [CrossRef]
10. Saliba, A.; Santos, S.; Vogel, J. New RNA-seq approaches for the study of bacterial pathogens. *Curr. Opin. Microbiol.* **2017**, *35*, 78–87. [CrossRef]
11. Liu, G.; Cheng, H.; You, W.; Song, E.L.; Liu, X.M.; Wan, F.C. Transcriptome profiling of muscle by RNA-Seq reveals significant differences in digital gene expression profiling between Angus and Luxi cattle. *Anim. Prod. Sci.* **2015**, *55*, 1172–1178. [CrossRef]
12. Zhang, Y.; Li, D.; Han, R.; Wang, Y.; Li, G.; Liu, X.; Tian, Y.; Kang, X.; Li, Z. Transcriptome analysis of the pectoral muscles of local chickens and commercial broilers using Ribo-Zero ribonucleic acid sequencing. *PLoS ONE* **2017**, *12*, e0184115. [CrossRef]
13. Ghosh, M.; Sodhi, S.S.; Song, K.D.; Kim, J.H.; Mongre, R.K.; Sharma, N.; Singh, N.K.; Kim, S.W.; Lee, H.K.; Jeong, D.K. Evaluation of body growth and immunity-related differentially expressed genes through deep RNA sequencing in the piglets of Jeju native pig and Berkshire. *Anim. Genet.* **2015**, *46*, 255–264. [CrossRef] [PubMed]
14. Huang, W.; Guo, Y.; Du, W.; Zhang, X.; Li, A.; Miao, X. Global transcriptome analysis identifies differentially expressed genes related to lipid metabolism in Wagyu and Holstein cattle. *Sci. Rep.* **2017**, *7*, 5278. [CrossRef] [PubMed]

15. Tirosh, I.; Izar, B.; Prakadan, S.M.; Wadsworth, M.H.; Treacy, D.; Trombetta, J.J.; Rotem, A.; Rodman, C.; Lian, C.; Murphy, G.; et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **2016**, *352*, 189–196. [CrossRef] [PubMed]
16. Guo, Y.; Su, Z.Y.; Zhang, C.; Gaspar, J.M.; Wang, R.; Hart, R.P.; Verzi, M.P.; Kong, A.N.T. Mechanisms of colitis-accelerated colon carcinogenesis and its prevention with the combination of aspirin and curcumin: Transcriptomic analysis using RNA-seq. *Biochem. Pharmacol.* **2017**, *135*, 22–34. [CrossRef] [PubMed]
17. Pereira, A.C.; Jason, D.G.; Joshua, F.K.; Rina, L.D.; Todd, G.R.; Masahiro, O.; John, H.M.; Bruce, S.M. Age and Alzheimer's disease gene expression profiles reversed by the glutamate modulator riluzole. *Mol. Psychiatry* **2017**, *22*, 296. [CrossRef]
18. Siede, D.; Rapti, K.; Gorska, A.; Katus, H.; Altmüller, J.; Boeckel, J.; Meder, B.; Maack, C.; Völkers, M.; Müller, O.; et al. Identification of circular RNAs with host gene-independent expression in human model systems for cardiac differentiation and disease. *J. Mol. Cell. Cardiol.* **2017**, *109*, 48–56. [CrossRef]
19. Mironova, V.V.; Weinholdt, C.; Grosse, I. RNA-seq data analysis for studying abiotic stress in horticultural plants. In *Abiotic Stress Biology in Horticultural Plants*; Springer: Tokyo, Japan, 2015; pp. 197–220. ISBN 9784431552512.
20. Schlüter, U.; Denton, A.K.; Brä utigam, A. Understanding metabolite transport and metabolism in C 4 plants through RNA-seq. *Curr. Opin. Plant Biol.* **2016**, *31*, 83–90. [CrossRef]
21. Ma, W.; Chen, C.; Liu, Y.; Zeng, M.; Meyers, B.C.; Li, J.; Xia, R. Coupling of microRNA-directed phased small interfering RNA generation from long noncoding genes with alternative splicing and alternative polyadenylation in small RNA-mediated gene silencing. *New Phytol.* **2018**, *217*, 1535–1550. [CrossRef]
22. Li, Y.; Huang, J.; Song, X.; Zhang, Z.; Jiang, Y.; Zhu, Y.; Zhao, H.; Ni, D. An RNA-Seq transcriptome analysis revealing novel insights into aluminum tolerance and accumulation in tea plant. *Planta* **2017**, *246*, 91–103. [CrossRef]
23. Tian, B.; Wang, S.; Todd, T.C.; Johnson, C.D.; Tang, G.; Trick, H.N. Genome-wide identification of soybean microRNA responsive to soybean cyst nematodes infection by deep sequencing. *BMC Genom.* **2017**, *18*, 165–183. [CrossRef]
24. Chotewutmontri, P.; Stiffler, N.; Watkins, K.P.; Barkan, A. Ribosome profiling in Maize. In *Methods in Molecular Biology*; Humana Press Inc.: New York, NY, USA, 2018; Volume 1676, pp. 165–183.
25. Zhang, Y.; Gu, L.; Hou, Y.; Wang, L.; Deng, X.; Hang, R.; Chen, D.; Zhang, X.; Zhang, Y.; Liu, C.; et al. Integrative genome-wide analysis reveals HLP1, a novel RNA-binding protein, regulates plant flowering by targeting alternative polyadenylation. *Cell Res.* **2015**, *25*, 864–876. [CrossRef] [PubMed]
26. Liu, W.; Duttke, S.; Hetzel, J.; Groth, M.; Feng, S. RNA-directed DNA methylation involves co-transcriptional small-RNA-guided slicing of polymerase V transcripts in Arabidopsis. *Nat. Plants* **2018**, *4*, 181. [CrossRef] [PubMed]
27. Hellman, L.; Fried, M. Electrophoretic Mobility Shift Assay (EMSA) for Detecting Protein-Nucleic Acid Interactions. *Nat. Protoc.* **2007**, *2*, 1849–1861. [CrossRef] [PubMed]
28. Azzeme, A.M.; Abdullah, S.N.A.; Aziz, M.A.; Wahab, P.E.M. Oil palm drought inducible DREB1 induced expression of DRE/CRT-and non-DRE/CRT-containing genes in lowland transgenic tomato under cold and PEG. *Plant Physiol. Biochem.* **2017**, *112*, 129–151. [CrossRef] [PubMed]
29. Wood, K.V. Marker proteins for gene expression. *Curr. Opin. Biotechnol.* **1995**, *6*, 50–58. [CrossRef]
30. Feng, S.Y.; Ota, K.; Yamada, Y.; Sawabu, N.; Ito, T. A yeast one-hybrid system to detect methylation-dependent DNA-protein interactions. *Biochem. Biophys. Res. Commun.* **2004**, *313*, 922–925. [CrossRef]
31. Ebrahimi, M.; Abdullah, S.N.A.; Aziz, M.A.; Namasivayam, P. A novel CBF that regulates abiotic stress response and the ripening process in oil palm (*Elaeis guineensis*) fruits. *Tree Genet. Genomes* **2015**, *11*, 56. [CrossRef]
32. Aparicio, O.; Geisberg, J.V.; Sekinger, E.; Yang, A.; Moqtaderi, Z.; Struhl, K. Chromatin Immunoprecipitation for Determining the Association of Proteins with Specific Genomic Sequences In Vivo. In *Current Protocols in Molecular Biology*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2005. [CrossRef]
33. Gilmour, D.S.; Lis, J.T. Detecting protein-DNA interactions in vivo: Distribution of RNA polymerase on specific bacterial genes (UV cross-linking/gene regulation/leucine operon/attenuation). *Proc. Natl. Acad. Sci. USA* **1984**, *81*, 4275–4279. [CrossRef]
34. Song, L.; Koga, Y.; Ecker, J.R. Profiling of Transcription Factor Binding Events by Chromatin Immunoprecipitation Sequencing (ChIP-seq). *Curr. Protoc. Plant Biol.* **2016**, *1*, 293–306. [CrossRef]

35. Haring, M.; Offermann, S.; Danker, T.; Horst, I.; Peterhansel, C.; Stam, M. Chromatin immunoprecipitation: Optimization, quantitative analysis and data normalization. *Plant Methods* **2007**, *3*, 11. [CrossRef]
36. Adli, M.; Bernstein, B.E. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nat. Protoc.* **2011**, *6*, 1656–1668. [CrossRef] [PubMed]
37. Zhou, W.; Sherwood, B.; Ji, H. Computational Prediction of the Global Functional Genomic Landscape: Applications, Methods, and Challenges. *Hum. Hered.* **2016**, *81*, 88–105. [CrossRef] [PubMed]
38. Shamimuzzaman, M.; Vodkin, L. Genome-wide identification of binding sites for NAC and YABBY transcription factors and co-regulated genes during soybean seedling development by ChIP-Seq and RNA-Seq. *BMC Genom.* **2013**, *14*, 477. [CrossRef] [PubMed]
39. Liu, D.; Shi, S.; Hao, Z.; Xiong, W.; Luo, M. OsbZIP81, A Homologue of Arabidopsis VIP1, May Positively Regulate JA Levels by Directly Targetting the Genes in JA Signaling and Metabolism Pathway in Rice. *Int. J. Mol. Sci.* **2019**, *20*, 2360. [CrossRef]
40. Du, M.; Zhao, J.; Tzeng, D.T.W.; Liu, Y.; Deng, L.; Yang, T.; Zhai, Q.; Wu, F.; Huang, Z.; Zhou, M.; et al. MYC2 Orchestrates a Hierarchical Transcriptional Cascade that Regulates Jasmonate-Mediated Plant Immunity in Tomato. *Plant Cell* **2017**, *29*, 1883–1906. [CrossRef] [PubMed]
41. Buenrostro, J.D.; Giresi, P.G.; Zaba, L.C.; Chang, H.Y.; Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **2013**, *10*, 1213–1218. [CrossRef]
42. Buenrostro, J.D.; Wu, B.; Chang, H.Y.; Greenleaf, W.J. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **2015**, *109*, 21–29. [CrossRef]
43. Basturea, G.; Corney, D.C. RNA-seq Using Next Generation Sequencing A comprehensive review of RNA-seq methodologies RNA-seq Using Next Generation Sequencing. *Mater Methods* **2016**, *3*, 203.
44. Lahens, N.F.; Ricciotti, E.; Smirnova, O.; Toorens, E.; Kim, E.J.; Baruzzo, G.; Hayer, K.E.; Ganguly, T.; Schug, J.; Grant, G.R. A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression. *BMC Genom.* **2017**, *18*, 602. [CrossRef]
45. Dündar, F.; Skrabanek, L.; Zumbo, P. *Introduction to Differential Gene Expression Analysis Using RNA-Seq*; Applied Bioinformatics Core/Weill Cornell Medical College: New York, NY, USA, September 2015.
46. Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63. [CrossRef]
47. Conesa, A.; Madrigal, P.; Tarazona, S.; Gomez-Cabrero, D.; Cervera, A.; McPherson, A.; Szcześniak, M.W.; Gaffney, D.J.; Elo, L.L.; Zhang, X.; et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **2016**, *17*, 13. [CrossRef] [PubMed]
48. Bullard, J.H.; Purdom, E.; Hansen, K.D.; Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinform.* **2010**, *11*, 94. [CrossRef] [PubMed]
49. Chu, Y.; Corey, D.R. RNA sequencing: Platform selection, experimental design, and data interpretation. *Nucleic Acid Ther.* **2012**, *22*, 271–274. [CrossRef] [PubMed]
50. Endrullat, C.; Glökler, J.; Franke, P.; Frohme, M. Standardization and quality management in next-generation sequencing. *Appl. Transl. Genom.* **2016**, *10*, 2–9. [CrossRef]
51. Hrdlickova, R.; Toloue, M.; Tian, B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip. Rev. RNA* **2017**, *8*, e1364. [CrossRef]
52. Zhao, S.; Zhang, Y.; Gordon, W.; Quan, J.; Xi, H.; Du, S.; von Schack, D.; Zhang, B. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genom.* **2015**, *16*, 675. [CrossRef]
53. Van Nieuwerburgh, F.; Soetaert, S.; Podshivalova, K.; Ay-Lin Wang, E.; Schaffer, L.; Deforce, D.; Salomon, D.R.; Head, S.R.; Ordoukhanian, P. Quantitative Bias in Illumina TruSeq and a Novel Post Amplification Barcoding Strategy for Multiplexed DNA and Small RNA Deep Sequencing. *PLoS ONE* **2011**, *6*, e26969. [CrossRef]
54. Martin, J.A.; Wang, Z. Next-generation transcriptome assembly. *Nat. Rev. Genet.* **2011**, *12*, 671–682. [CrossRef]
55. Li, H.; Ruan, J.; Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **2008**, *18*, 1851–1858. [CrossRef]
56. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
57. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]

58. Kim, D.; Paggi, J.M.; Park, C.; Bennett, C.; Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **2019**, *37*, 907–915. [CrossRef] [PubMed]
59. Wang, K.; Singh, D.; Zeng, Z.; Coleman, S.J.; Huang, Y.; Savich, G.L.; He, X.; Mieczkowski, P.; Grimm, S.A.; Perou, C.M.; et al. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **2010**, *38*, e178. [CrossRef] [PubMed]
60. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [CrossRef] [PubMed]
61. Thorvaldsdóttir, H.; Robinson, J.T.; Mesirov, J.P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **2013**, *14*, 178–192. [CrossRef] [PubMed]
62. Trapnell, C.; Williams, B.A.; Pertea, G.; Mortazavi, A.; Kwan, G.; Van Baren, M.J.; Salzberg, S.L.; Wold, B.J.; Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **2010**, *28*, 511–515. [CrossRef] [PubMed]
63. Guttman, M.; Garber, M.; Levin, J.Z.; Donaghey, J.; Robinson, J.; Adiconis, X.; Fan, L.; Koziol, M.J.; Gnirke, A.; Nusbaum, C.; et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **2010**, *28*, 503–510. [CrossRef]
64. Katz, Y.; Wang, E.T.; Airoldi, E.M.; Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **2010**, *7*, 1009–1015. [CrossRef]
65. Pajoro, A.; Severing, E.; Angenent, G.C.; Immink, R.G.H. Histone H3 lysine 36 methylation affects temperature-induced alternative splicing and flowering in plants. *Genome Biol.* **2017**, *18*, 102. [CrossRef]
66. Piya, S.; Liu, J.; Burch-Smith, T.; Baum, T.J.; Hewezi, T. The roles of Arabidopsis Growth-Regulating Factors 1 and 3 in growth-stress antagonism. *J. Exp. Bot.* **2019**. [CrossRef]
67. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [CrossRef] [PubMed]
68. Xie, Y.; Wu, G.; Tang, J.; Luo, R.; Patterson, J.; Liu, S.; Huang, W.; He, G.; Gu, S.; Li, S.; et al. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatic* **2014**, *30*, 1660–1666. [CrossRef] [PubMed]
69. Robertson, G.; Schein, J.; Chiu, R.; Corbett, R.; Field, M.; Jackman, S.D.; Mungall, K.; Lee, S.; Okada, H.M.; Qian, J.Q.; et al. De novo assembly and analysis of RNA-seq data. *Nat. Methods* **2010**, *7*, 909. [CrossRef] [PubMed]
70. Garber, M.; Grabherr, M.G.; Guttman, M.; Trapnell, C. Nature Methods Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* **2011**, *8*, 469. [CrossRef]
71. Lu, B.; Zeng, Z.; Shi, T. Comparative study of *de novo* assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. *Sci. China Life Sci.* **2013**, *56*, 143–155. [CrossRef]
72. Van Verk, M.C.; Hickman, R.; Pieterse, C.M.J.; Van Wees, S.C.M. RNA-Seq: Revelation of the messengers. *Trends Plant Sci.* **2013**, *18*, 175–179. [CrossRef]
73. Wagner, G.P.; Kin, K.; Lynch, V.J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **2012**, *131*, 281–285. [CrossRef]
74. Anders, S.; Pyl, P.; Huber, W. HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics* **2015**, *31*, 166–169. [CrossRef]
75. Zyprych-Walczak, J.; Szabelska, A.; Handschuh, L.; Górczak, K.; Klamecka, K.; Figlerowicz, M.; Siatkowski, I. The Impact of Normalization Methods on RNA-Seq Data Analysis. *BioMed Res. Int.* **2015**, *2015*, 621690. [CrossRef]
76. Finotello, F.; Camillo, B. Di Measuring differential gene expression with RNA-seq: Challenges and strategies for data analysis. *Brief. Funct. Genom.* **2014**, *14*, 130–142. [CrossRef]
77. Dillies, M.; Rau, A.; Aubert, J.; Hennequet-Antier, C.; Jeanmougin, M.; Servant, N.; Keime, C.; Marot, G.; Castel, D.; Estelle, J.; et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **2012**, *14*, 671–683. [CrossRef] [PubMed]
78. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef] [PubMed]
79. Anders, S.; Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **2010**, *11*, R106. [CrossRef] [PubMed]

80. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef] [PubMed]
81. Trapnell, C.; Hendrickson, D.; Sauvageau, M.; Goff, L.; Rinn, J.; Pachter, L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **2013**, *31*, 46. [CrossRef] [PubMed]
82. Trapnell, C.; Roberts, A.; Goff, L.; Pertea, G.; Kim, D. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **2012**, *7*, 562. [CrossRef]
83. Hardcastle, T.J.; Kelly, K.A. BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinform.* **2010**, *11*, 422. [CrossRef]
84. Tarazona, S.; Furió-Tarí, P.; Turra, D.; Pietro, A.D.; Nueda, M.J.; Ferrer, A.; Conesa, A. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucl. Acids Res.* **2015**, *43*, e140. [CrossRef]
85. Chen, Y.; McCarthy, D.; Robinson, M.; Smyth, G. edgeR: Differential Expression Analysis of Digital Gene Expression Data User's Guide. *Bioconductor User's Guide*. Available online: http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf (accessed on 17 September 2008).
86. Costa-Silva, J.; Domingues, D.; Lopes, F.M. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS ONE* **2017**, *12*, e0190152. [CrossRef]
87. Law, C.W.; Chen, Y.; Shi, W.; Smyth, G.K. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **2014**, *15*, R29. [CrossRef]
88. Ashburner, M.; Ball, C.; Blake, J.; Botstein, D.; Butler, H. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef] [PubMed]
89. Dennis, G.; Sherman, B.T.; Hosack, D.A.; Yang, J.; Gao, W.; Lane, H.C.; Lempicki, R.A. DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol.* **2003**, *4*, R60. [CrossRef]
90. Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucl. Acids Res.* **2012**, *49*, D109–D114. [CrossRef] [PubMed]
91. Schwacke, R.; Ponce-Soto, G.Y.; Krause, K.; Bolger, A.M.; Arsova, B.; Hallab, A.; Gruden, K.; Stitt, M.; Bolger, M.E.; Usadel, B. MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. *Mol. Plant* **2019**, *12*, 879–892. [CrossRef] [PubMed]
92. Szklarczyk, D.; Morris, J.; Cook, H.; Kuhn, M.; Wyder, S.; Simonovic, M.; Santos, A.; Doncheva, N.T.; Roth, A.; Bork, P. The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible. *Nuceic Acids Res.* **2017**, *45*, gkw937. [CrossRef] [PubMed]
93. Strickler, S.R.; Bombarely, A.; Mueller, L.A. Designing a transcriptome next-generation sequencing project for a nonmodel plant species. *Am. J. Bot.* **2012**, *99*, 257–266. [CrossRef] [PubMed]
94. Landt, S.G.; Marinov, G.K.; Kundaje, A.; Kheradpour, P.; Pauli, F.; Batzoglou, S.; Bernstein, B.E.; Bickel, P.; Brown, J.B.; Cayting, P.; et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **2012**, *22*, 1813–1831. [CrossRef]
95. Quievryn, G.; Zhitkovich, A. Loss of DNA–protein crosslinks from formaldehyde-exposed cells occurs through spontaneous hydrolysis and an active repair process linked to proteosome function. *Carcinogenesis* **2000**, *21*, 1573–1580. [CrossRef]
96. McGhee, J.D.; Von Hippel, P.H. Formaldehyde as a Probe of DNA Structure. I. Reaction with Exocyclic Amino Groups of DNA Bases. *Biochemistry* **1975**, *14*, 1281–1296. [CrossRef]
97. Hoffman, E.A.; Frey, B.L.; Smith, L.M.; Auble, D.T. Formaldehyde crosslinking: A tool for the study of chromatin complexes. *J. Biol. Chem.* **2015**, *290*, 26404–26411. [CrossRef]
98. Orlando, V. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem. Sci.* **2000**, *25*, 99–104. [CrossRef]
99. Saleh, A.; Alvarez-Venegas, R.; Avramova, Z. An efficient chromatin immunoprecipitation (ChIP) protocol for studying histone modifications in Arabidopsis plants. *Nat. Protoc.* **2008**, *3*, 1018–1025. [CrossRef] [PubMed]
100. Kaufmann, K.; Muiño, J.M.; Østerås, M.; Farinelli, L.; Krajewski, P.; Angenent, G.C. Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). *Nat. Protoc.* **2010**, *5*, 457–472. [CrossRef] [PubMed]
101. Wiśniewski, J.R.; Duś, K.; Mann, M. Proteomic workflow for analysis of archival formalin-fixed and paraffin-embedded clinical samples to a depth of 10000 proteins. *Proteom. Clin. Appl.* **2013**, *7*, 225–233. [CrossRef]

102. Johnson, D.S.; Mortazavi, A.; Myers, R.M.; Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **2007**, *316*, 1497–1502. [CrossRef]
103. Hills, P.N.; Van Staden, J.; Scott, P. An improved DNA extraction procedure for plant tissues with a high phenolic content. *S. Afr. J. Bot.* **2005**, *68*, 549–550. [CrossRef]
104. Yamaguchi, N.; Winter, C.M.; Wu, M.-F.; Kwon, C.S.; William, D.A.; Wagner, D. PROTOCOL: Chromatin Immunoprecipitation from Arabidopsis Tissues. *Arab. Book Am. Soc. Plant Biol.* **2014**, *12*, e0170. [CrossRef]
105. Zhong, J.; Ye, Z.; Lenz, S.W.; Clark, C.R.; Bharucha, A.; Farrugia, G.; Robertson, K.D.; Zhang, Z.; Ordog, T.; Lee, J.H. Purification of nanogram-range immunoprecipitated DNA in ChIP-seq application. *BMC Genom.* **2017**, *18*, 985. [CrossRef]
106. De Folter, S.; Urbanus, S.L.; van Zuijlen, L.G.; Kaufmann, K.; Angenent, G.C. Tagging of MADS domain proteins for chromatin immunoprecipitation. *BMC Plant Biol.* **2007**, *7*, 47. [CrossRef]
107. De la Fuente, L.; Conesa, A.; Lloret, A.; Badenes, M.L.; Ríos, G. Genome-wide changes in histone H3 lysine 27 trimethylation associated with bud dormancy release in peach. *Tree Genet. Genomes* **2015**, *11*, 45. [CrossRef]
108. Hussey, S.G.; Mizrachi, E.; Groover, A.; Berger, D.K.; Myburg, A.A. Genome-wide mapping of histone H3 lysine 4 trimethylation in Eucalyptus grandis developing xylem. *BMC Plant Biol.* **2015**, *15*, 117. [CrossRef] [PubMed]
109. Brind'Amour, J.; Liu, S.; Hudson, M.; Chen, C.; Karimi, M.M.; Lorincz, M.C. An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. *Nat. Commun.* **2015**, *6*, 6033. [CrossRef] [PubMed]
110. Ko, D.K.; Rohozinski, D.; Song, Q.; Taylor, S.H.; Juenger, T.E.; Harmon, F.G.; Chen, Z.J. Temporal Shift of Circadian-Mediated Gene Expression and Carbon Fixation Contributes to Biomass Heterosis in Maize Hybrids. *PLoS Genet.* **2016**, *12*, e1006197. [CrossRef] [PubMed]
111. Posé, D.; Verhage, L.; Ott, F.; Yant, L.; Mathieu, J.; Angenent, G.C.; Immink, R.G.H.; Schmid, M. Temperature-dependent regulation of flowering by antagonistic FLM variants. *Nature* **2013**, *503*, 414–417. [CrossRef]
112. Fan, M.; Bai, M.Y.; Kim, J.G.; Wang, T.; Oh, E.; Chen, L.; Park, C.H.; Son, S.H.; Kim, S.K.; Mudgett, M.B.; et al. The bHLH transcription factor HBI1 mediates the trade-off between growth and pathogen-associated molecular pattern–triggered immunity in Arabidopsis. *Plant Cell* **2014**, *26*, 828–841. [CrossRef]
113. Lu, L.; Chen, X.; Qian, S.; Zhong, X. The plant-specific histone residue Phe41 is important for genome-wide H3.1 distribution. *Nat. Commun.* **2018**, *9*, 630. [CrossRef]
114. Zhang, D.; Sun, W.; Singh, R.; Zheng, Y.; Cao, Z.; Li, M.; Lunde, C.; Hake, S.; Zhang, Z. GRF-interacting factor1 Regulates Shoot Architecture and Meristem Determinacy in Maize. *Plant Cell* **2018**, *30*, 360–374. [CrossRef]
115. Chung, D.; Park, D.; Myers, K.; Grass, J.; Kiley, P.; Landick, R.; Keleş, S. dPeak: High Resolution Identification of Transcription Factor Binding Sites from PET and SET ChIP-Seq Data. *PLoS Comput. Biol.* **2013**, *9*, 9–11. [CrossRef]
116. Bailey, T.; Krajewski, P.; Ladunga, I.; Lefebvre, C.; Li, Q.; Liu, T.; Madrigal, P.; Taslim, C.; Zhang, J. Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. *PLoS Comput. Biol.* **2013**, *9*, 5–12. [CrossRef]
117. Li, C.; Qiao, Z.; Qi, W.; Wang, Q.; Yuan, Y.; Yang, X.; Tang, Y.; Mei, B.; Lv, Y.; Zhao, H.; et al. Genome-Wide Characterization of cis -Acting DNA Targets Reveals the Transcriptional Regulatory Framework of Opaque2 in Maize. *Plant Cell* **2015**, *27*, 532–545. [CrossRef]
118. Yant, L.; Ott, F.; Keller, H.; Weigel, D.; Schmid, M. 20th International of Conference on Arabidopsis Research, Scotland, United Kingdom. In Proceedings of the Design and Analysis of ChIP-Seq Experiments in Plants: A Systematic Comparison of ChIP-Seq and ChIP-chip for APETALA2 (AP2), FD, and SCHLAFMÜTZE (SMZ), Edinburgh, UK, 30 June–4 July 2009.
119. Nakato, R.; Shirahige, K. Recent advances in ChIP-seq analysis: From quality management to whole-genome annotation. *Brief. Bioinform.* **2017**, *18*, 279–290. [CrossRef] [PubMed]
120. Pepke, S.; Wold, B.; Mortazavi, A. Computation for ChIP-seq and RNA-seq studies. *Nat. Methods* **2009**, *6*, S22. [CrossRef] [PubMed]
121. Furey, T.S. ChIP–seq and beyond: New and improved methodologies to detect and characterize protein–DNA interactions. *Nat. Rev. Genet.* **2012**, *13*, 840–852. [CrossRef] [PubMed]
122. Baugh, L.R.; DeModena, J.; Sternberg, P.W. RNA Pol II accumulates at promoters of growth genes during developmental arrest. *Science* **2009**, *324*, 92–94. [CrossRef] [PubMed]

123. Rozowsky, J.; Euskirchen, G.; Auerbach, R.K.; Zhang, Z.D.; Gibson, T.; Bjornson, R.; Carriero, N.; Snyder, M.; Gerstein, M.B. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* **2009**, *27*, 66–75. [CrossRef] [PubMed]
124. Robinson, M.D.; Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **2010**, *11*, R25. [CrossRef]
125. Muiño, J.M.; Kaufmann, K.; van Ham, R.C.; Angenent, G.C.; Krajewski, P. ChIP-seq Analysis in R (CSAR): An R package for the statistical detection of protein-bound genomic regions. *Plant Methods* **2011**, *7*, 11. [CrossRef]
126. Taleb, H.A.; AL-Dherasi, A.; Al-Mosaib, S.; Alnoud, M.; Vilaphong, S. Peak Calling Algorithms and Their Applications for Next-Generation Sequencing Technologies. *Indian J. Nat. Sci.* **2019**, *9*, 16659–16670.
127. Johannes, F.; Wardenaar, R.; Colomé-Tatché, M.; Mousson, F.; de Graaf, P.; Mokry, M.; Guryev, V.; Timmers, H.T.M.; Cuppen, E.; Jansen, R.C. Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq. *Bioinformatics* **2010**, *26*, 1000–1006. [CrossRef]
128. Burrows, M.; Wheeler, D. *A Block-Sorting Lossless Data Compression Algorithm*; Technical Report for Digital Equipment Corporation: Maynard, MA, USA, May 1994.
129. Ferragina, P.; Manzini, G. An Experimental Study of an Opportunistic Index. In Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms, Washington, DC, USA, 7–9 January 2001; pp. 269–278.
130. Li, R.; Yu, C.; Li, Y.; Lam, T.W.; Yiu, S.M.; Kristiansen, K.; Wang, J. SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* **2009**, *25*, 1966–1967. [CrossRef]
131. Kim, D.; Langmead, B.; Salzberg, S.L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **2015**, *12*, 357–360. [CrossRef] [PubMed]
132. Chen, K.; Xi, Y.; Pan, X.; Li, Z.; Kaestner, K.; Tyler, J.; Dent, S.; He, X.; Li, W. DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res.* **2013**, *23*, 341–351. [CrossRef] [PubMed]
133. Latrasse, D.; Jégu, T.; Li, H.; de Zelicourt, A.; Raynaud, C.; Legras, S.; Gust, A.; Samajova, O.; Veluchamy, A.; Rayapuram, N.; et al. MAPK-triggered chromatin reprogramming by histone deacetylase in plant innate immunity. *Genome Biol.* **2017**, *18*, 131. [CrossRef] [PubMed]
134. Ricardi, M.M.; González, R.M.; Zhong, S.; Domínguez, P.G.; Duffy, T.; Turjanski, P.G.; Salgado Salter, J.D.; Alleva, K.; Carrari, F.; Giovannoni, J.J.; et al. Genome-wide data (ChIP-seq) enabled identification of cell wall-related and aquaporin genes as targets of tomato ASR1, a drought stress-responsive transcription factor. *BMC Plant Biol.* **2014**, *14*, 29. [CrossRef] [PubMed]
135. Zhang, Y.; Liu, T.; Meyer, C.A.; Eeckhoute, J.; Johnson, D.S.; Bernstein, B.E.; Nussbaum, C.; Myers, R.M.; Brown, M.; Li, W.; et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **2008**, *9*, R137. [CrossRef] [PubMed]
136. Lu, Z.; Yu, H.; Xiong, G.; Wang, J.; Jiao, Y.; Liu, G.; Jing, Y.; Meng, X.; Hu, X.; Qian, Q.; et al. Genome-Wide Binding Analysis of the Transcription Activator IDEAL PLANT ARCHITECTURE$_1$ Reveals a Complex Network Regulating Rice Plant Architecture. *Plant Cell* **2013**, *25*, 3743–3759. [CrossRef]
137. Mateos, J.L.; Madrigal, P.; Tsuda, K.; Rawat, V.; Richter, R.; Romera-Branchat, M.; Fornara, F.; Schneeberger, K.; Krajewski, P.; Coupland, G. Combinatorial activities of SHORT VEGETATIVE PHASE and FLOWERING LOCUS C define distinct modes of flowering regulation in Arabidopsis. *Genome Biol.* **2015**, *16*, 31. [CrossRef]
138. Zhang, F.; Wang, L.; Qi, B.; Zhao, B.; Ko, E.E.; Riggan, N.D.; Chin, K.; Qiao, H. EIN2 mediates direct regulation of histone acetylation in the ethylene response. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 201707937. [CrossRef]
139. Machanick, P.; Bailey, T.L. MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics* **2011**, *27*, 1696–1697. [CrossRef]
140. Bailey, T.L.; Boden, M.; Buske, F.A.; Frith, M.; Grant, C.E.; Clementi, L.; Ren, J.; Li, W.W.; Noble, W.S. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, W202–W208. [CrossRef]
141. Bailey, T.L. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **2011**, *27*, 1653–1659. [CrossRef] [PubMed]
142. Thomas-Chollier, M.; Darbo, E.; Herrmann, C.; Defrance, M.; Thieffry, D.; van Helden, J. A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nat. Protoc.* **2012**, *7*, 1551–1568. [CrossRef] [PubMed]

143. Zang, C.; Schones, D.E.; Zeng, C.; Cui, K.; Zhao, K.; Peng, W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **2009**, *25*, 1952–1958. [CrossRef] [PubMed]
144. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef] [PubMed]
145. Arenhart, R.A.; Bai, Y.; Valter De Oliveira, L.F.; Bucker Neto, L.; Schunemann, M.; Maraschin, F.D.S.; Mariath, J.; Silverio, A.; Sachetto-Martins, G.; Margis, R.; et al. New insights into aluminum tolerance in rice: The ASR5 protein binds the STAR1 promoter and other aluminum-responsive genes. *Mol. Plant* **2014**, *7*, 709–721. [CrossRef]
146. Zhang, M.; Xie, S.; Dong, X.; Zhao, X.; Zeng, B.; Chen, J.; Li, H.; Yang, W.; Zhao, H.; Wang, G.; et al. Genome-wide high resolution parental-specific DNA and histone methylation maps uncover patterns of imprinting regulation in maize. *Genome Res.* **2014**, *24*, 167–176. [CrossRef]
147. Lei, K.J.; Lin, Y.M.; Ren, J.; Bai, L.; Miao, Y.C.; An, G.Y.; Song, C.P. Modulation of the phosphate-deficient responses by MicroRNA156 and its targeted SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 3 in arabidopsis. *Plant Cell Physiol.* **2016**, *57*, 192–203. [CrossRef]
148. Pedmale, U.V.; Huang, S.S.C.; Zander, M.; Cole, B.J.; Hetzel, J.; Ljung, K.; Reis, P.A.B.; Sridevi, P.; Nito, K.; Nery, J.R.; et al. Cryptochromes Interact Directly with PIFs to Control Plant Growth in Limiting Blue Light. *Cell* **2016**, *164*, 233–245. [CrossRef]
149. Bi, X.; Cheng, Y.-J.; Hu, B.; Ma, X.; Wu, R.; Wang, J.-W.; Liu, C. Nonrandom domain organization of theArabidopsisgenome at the nuclear periphery. *Genome Res.* **2017**, *27*, 1162–1173. [CrossRef]
150. Huang, L.; Jia, J.; Zhao, X.; Zhang, M.Y.; Huang, X.; Ji, E.; Ni, L.; Jiang, M. The ascorbate peroxidase APX1 is a direct target of a zinc finger transcription factor ZFP36 and a late embryogenesis abundant protein OsLEA5 interacts with ZFP36 to co-regulate OsAPX1 in seed germination in rice. *Biochem. Biophys. Res. Commun.* **2018**, *495*, 339–345. [CrossRef]
151. Balanzà, V.; Martínez-Fernández, I.; Sato, S.; Yanofsky, M.F.; Kaufmann, K.; Angenent, G.C.; Bemer, M.; Ferrándiz, C. Genetic control of meristem arrest and life span in Arabidopsis by a FRUITFULL-APETALA2 pathway. *Nat. Commun.* **2018**, *9*, 565. [CrossRef] [PubMed]
152. Martin, R.C.; Vining, K.; Dombrowski, J.E. Genome-wide (ChIP-seq) identification of target genes regulated by BdbZIP10 during paraquat-induced oxidative stress. *BMC Plant Biol.* **2018**, *18*, 58. [CrossRef] [PubMed]
153. Li, S.; Tian, Y.; Wu, K.; Ye, Y.; Yu, J.; Zhang, J.; Liu, Q.; Hu, M.; Li, H.; Tong, Y.; et al. Modulating plant growth–metabolism coordination for sustainable agriculture. *Nature* **2018**, *560*, 595–600. [CrossRef] [PubMed]
154. Bang, S.W.; Lee, D.K.; Jung, H.; Chung, P.J.; Kim, Y.S.; Choi, Y.D.; Suh, J.W.; Kim, J.K. Overexpression of *OsTF1L*, a rice HD-Zip transcription factor, promotes lignin biosynthesis and stomatal closure that improves drought tolerance. *Plant Biotechnol. J.* **2019**, *17*, 118–131. [CrossRef]
155. Zhao, S.; Cheng, L.; Gao, Y.; Zhang, B.; Zheng, X.; Wang, L.; Li, P.; Sun, Q.; Li, H. Plant HP1 protein ADCP1 links multivalent H3K9 methylation readout to heterochromatin formation. *Cell Res.* **2019**, *29*, 54–66. [CrossRef]
156. Song, L.; Chen, W.; Wang, B.; Yao, Q.M.; Valliyodan, B.; Bai, M.Y.; Zhao, M.Z.; Ye, H.; Wang, Z.Y.; Nguyen, H.T. GmBZL3 acts as a major BR signaling regulator through crosstalk with multiple pathways in Glycine max. *BMC Plant Biol.* **2019**, *19*, 86. [CrossRef]
157. Collani, S.; Neumann, M.; Yant, L.; Schmid, M. FT Modulates Genome-Wide DNA-Binding of the bZIP Transcription Factor FD. *Plant Physiol.* **2019**, *180*, 367–380. [CrossRef]
158. Roberts, A.; Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* **2013**, *10*, 71–73. [CrossRef]
159. Li, B.; Dewey, C.N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [CrossRef]
160. Benjamini, Y.; Drai, D.; Elmer, G.; Kafkafi, N.; Golani, I. Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* **2001**, *125*, 279–284. [CrossRef]
161. Ye, T.; Krebs, A.R.; Choukrallah, M.-A.; Keime, C.; Plewniak, F.; Davidson, I.; Tora, L. seqMINER: An integrated ChIP-seq data interpretation platform. *Nucleic Acids Res.* **2011**, *39*, e35. [CrossRef] [PubMed]
162. Du, Z.; Zhou, X.; Ling, Y.; Zhang, Z.; Su, Z. agriGO: A GO analysis toolkit for the agricultural community. *Nucleic Acids Res.* **2010**, *38*, W64–W70. [CrossRef] [PubMed]

163. O'Malley, R.C.; Huang, S.C.; Song, L.; Lewsey, M.G.; Bartlett, A.; Nery, J.R.; Galli, M.; Gallavotti, A.; Ecker, J.R. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **2016**, *165*, 1280–1292. [CrossRef] [PubMed]
164. Chen, Y.; Chen, Y.; Shi, Z.; Jin, Y.; Sun, H.; Xie, F.; Zhang, L. Biosynthesis and Signal Transduction of ABA, JA, and BRs in Response to Drought Stress of Kentucky Bluegrass. *Int. J. Mol. Sci.* **2019**, *20*, 1289. [CrossRef] [PubMed]
165. Hu, J.; Wang, Y.; Fang, Y.; Zeng, L.; Xu, J.; Yu, H.; Shi, Z.; Pan, J.; Zhang, D.; Kang, S.; et al. A Rare Allele of GS2 Enhances Grain Size and Grain Yield in Rice. *Mol. Plant* **2015**, *8*, 1455–1465. [CrossRef] [PubMed]
166. Lapham, R.; Lee, L.Y.; Tsugama, D.; Lee, S.; Mengiste, T.; Gelvin, S.B. VIP1 and its homologs are not required for agrobacterium-mediated transformation, but play a role in botrytis and salt stress responses. *Front. Plant Sci.* **2018**, *9*, 749. [CrossRef]
167. Tsugama, D.; Liu, S.; Takano, T. A bZIP Protein, VIP1, Is a Regulator of Osmosensory Signaling in Arabidopsis 1[W]. *Plant Physiol.* **2012**, *159*, 144–155. [CrossRef]
168. Rhoads, A.; Au, K.F. PacBio Sequencing and Its Applications. *Genom. Proteom. Bioinform.* **2015**, *13*, 278–289. [CrossRef]
169. Ardui, S.; Ameur, A.; Vermeesch, J.R.; Hestand, M.S. Single molecule real-time (SMRT) sequencing comes of age: Applications and utilities for medical diagnostics. *Nucleic Acids Res.* **2018**, *46*, 2159–2168. [CrossRef]
170. Schloss, P.D.; Jenior, M.L.; Koumpouras, C.C.; Westcott, S.L.; Highlander, S.K. Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. *PeerJ* **2016**, *4*, e1869. [CrossRef]
171. Gueidan, C.; Elix, J.A.; McCarthy, P.M.; Roux, C.; Mallen-Cooper, M.; Kantvilas, G. PacBio amplicon sequencing for metabarcoding of mixed DNA samples from lichen herbarium specimens. *MycoKeys* **2019**, *53*, 73–91. [CrossRef] [PubMed]
172. Roberts, R.J.; Carneiro, M.O.; Schatz, M.C. The advantages of SMRT sequencing. *Genome Biol.* **2013**, *14*, 405. [CrossRef] [PubMed]
173. Lu, H.; Giordano, F.; Ning, Z. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genom. Proteom. Bioinform.* **2016**, *14*, 265–279. [CrossRef] [PubMed]
174. Ameur, A.; Kloosterman, W.P.; Hestand, M.S. Single-Molecule Sequencing: Towards Clinical Applications. *Trends Biotechnol.* **2019**, *37*, 72–85. [CrossRef]
175. Laboratories, K. Partek: How to Integrate ChIP-Seq and RNA-Seq Data. Available online: https://www.partek.com/how-to-integrate-chip-seq-and-rna-seq-data/ (accessed on 9 December 2019).
176. Wang, S.; Sun, H.; Ma, J.; Zang, C.; Wang, C.; Wang, J.; Tang, Q.; Meyer, C.A.; Zhang, Y.; Liu, X.S. Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat. Protoc.* **2013**, *8*, 2502–2515. [CrossRef]
177. Zhang, C.; Zhang, B.; Lin, L.-L.; Zhao, S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genom.* **2017**, *18*, 583. [CrossRef]
178. O'Geen, H.; Henry, I.M.; Bhakta, M.S.; Meckler, J.F.; Segal, D.J. A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. *Nucleic Acids Res.* **2015**, *43*, 3389–3404. [CrossRef]
179. Partridge, E.C.; Watkins, T.A.; Mendenhall, E.M. Every transcription factor deserves its map: Scaling up epitope tagging of proteins to bypass antibody problems. *BioEssays* **2016**, *38*, 801–811. [CrossRef]
180. Savic, D.; Partridge, E.C.; Newberry, K.M.; Smith, S.B.; Meadows, S.K.; Roberts, B.S.; Mackiewicz, M.; Mendenhall, E.M.; Myers, R.M. CETCh-seq: CRISPR epitope tagging ChIP-seq of DNA-binding proteins. *Genome Res.* **2015**, *25*, 1581–1589. [CrossRef]
181. Moradpour, M.; Abdulah, S.N.A. CRISPR/dCas9 platforms in plants: Strategies and applications beyond genome editing. *Plant Biotechnol. J.* **2019**. [CrossRef]
182. Li, C.; Chen, C.; Chen, H.; Wang, S.; Chen, X.; Cui, Y. Verification of DNA motifs in Arabidopsis using CRISPR/Cas9-mediated mutagenesis. *Plant Biotechnol. J.* **2018**, *16*, 1446–1451. [CrossRef] [PubMed]

International Journal of
Molecular Sciences

Article

The Terrestrial Carnivorous Plant *Utricularia reniformis* Sheds Light on Environmental and Life-Form Genome Plasticity

Saura R. Silva [1], Ana Paula Moraes [2], Helen A. Penha [1], Maria H. M. Julião [1], Douglas S. Domingues [3], Todd P. Michael [4], Vitor F. O. Miranda [5,*] and Alessandro M. Varani [1,*]

[1] Departamento de Tecnologia, Faculdade de Ciências Agrárias e Veterinárias, UNESP—Universidade Estadual Paulista, Jaboticabal 14884-900, Brazil; saura.silva@gmail.com (S.R.S.); helen.penha@gmail.com (H.A.P.); mhmjuliao@gmail.com (M.H.M.J.)
[2] Centro de Ciências Naturais e Humanas, Universidade Federal do ABC, São Bernardo do Campo 09606-070, Brazil; apaulademoraes@gmail.com
[3] Departamento de Botânica, Instituto de Biociências, UNESP—Universidade Estadual Paulista, Rio Claro 13506-900, Brazil; douglas.domingues@unesp.br
[4] J. Craig Venter Institute, La Jolla, CA 92037, USA; tmichael@jcvi.org
[5] Departamento de Biologia Aplicada à Agropecuária, Faculdade de Ciências Agrárias e Veterinárias, UNESP—Universidade Estadual Paulista, Jaboticabal 14884-900, Brazil
* Correspondence: vitor.miranda@unesp.br (V.F.O.M.); alessandro.varani@unesp.br (A.M.V.)

Received: 23 October 2019; Accepted: 15 December 2019; Published: 18 December 2019

Abstract: *Utricularia* belongs to Lentibulariaceae, a widespread family of carnivorous plants that possess ultra-small and highly dynamic nuclear genomes. It has been shown that the Lentibulariaceae genomes have been shaped by transposable elements expansion and loss, and multiple rounds of whole-genome duplications (WGD), making the family a platform for evolutionary and comparative genomics studies. To explore the evolution of *Utricularia*, we estimated the chromosome number and genome size, as well as sequenced the terrestrial bladderwort *Utricularia reniformis* ($2n$ = 40, 1C = 317.1-Mpb). Here, we report a high quality 304 Mb draft genome, with a scaffold NG50 of 466-Kb, a BUSCO completeness of 87.8%, and 42,582 predicted genes. Compared to the smaller and aquatic *U. gibba* genome (101 Mb) that has a 32% repetitive sequence, the *U. reniformis* genome is highly repetitive (56%). The structural differences between the two genomes are the result of distinct fractionation and rearrangements after WGD, and massive proliferation of LTR-retrotransposons. Moreover, GO enrichment analyses suggest an ongoing gene birth–death–innovation process occurring among the tandem duplicated genes, shaping the evolution of carnivory-associated functions. We also identified unique patterns of developmentally related genes that support the terrestrial life-form and body plan of *U. reniformis*. Collectively, our results provided additional insights into the evolution of the plastic and specialized Lentibulariaceae genomes.

Keywords: evolution; genome fractionation; ABC transporters; transcription factors; transposable elements; whole-genome duplication

1. Introduction

The carnivorous plants from the Lentibulariaceae Rich. family exhibit different life-forms according to their habitats (terrestrial, aquatic, lithophytes, epiphytes, and rheophytes), showing an enormous diversity and worldwide distribution [1]. The family is composed of three genera: *Genlisea* A. St.-Hil. (corkscrew plants), *Pinguicula* L. (butterworts), and *Utricularia* L. (bladderworts). The sequencing of four Lentibulariaceae genomes: *Genlisea aurea* A. St.-Hil. [2], *G. nigrocaulis* Steyerm, *G. hispidula* Stapf [3], and *Utricularia gibba* L. [4,5], along with the estimation of genome size [6–8] and karyological analyses (e.g., [3,9–12]), revealed a wide distribution of genome size and chromosome numbers. In *Utricularia*, genome size ranges from ultra-small (~79 Mb) in *U. purpurea* Walter to large (~706 Mb) in *U. caerulea* L., with a variable number of chromosomes (e.g., 2*n* = 18, 28, 30, 36, 38, 40, 44, 48, and 56), supporting the occurrence of polyploidy and aneuploidy/dysploidy (for review see [8,12]).

Lentibulariaceae genome shrinkage has been attributed to a decrease in the number and length of introns and intergenic regions, associated with transposable elements (TEs) silencing. This phenomenon was primarily observed in *U. gibba* (101 Mb) and *G. aurea* (63–131 Mb according to [8,12]) genomes [2,13,14]. One of the leading hypotheses to explain the genome shrinkage is based on the high respiration rates caused by the carnivorous traps [15], which can increase reactive oxygen species (ROS) formation. Moreover, the *U. gibba* and the 86 Mb *G. nigrocaulis* genomes have suffered high rates of gene deletion in comparison to other eudicots, and in the former, a positive selection of tandemly repeated genes implicated in the adaptation to the carnivorous habit is observed [3,5,16].

In contrast, genome size expansion is generally attributed to TEs proliferation, whole-genome duplication (WGD) (represented by retained polyploid duplicates), and small-scale duplication (mainly represented by proximal and tandem copies) [17,18], which are the standard processes for plant genome evolution [19,20]. Together, TEs and WGDs are considered to be the key players in generating genomic novelties and genomic diversification [21–23]. Even among the minimal Lentibulariaceae genomes, the importance of these events it is clear, since the changes in size between the large, 1.5-Gb genome of *G. hispidula* and that of *G. nigrocaulis* (86 Mb) are consequences of WGDs, along with the LTR-retrotransposons expansion in the former, and the double-strand break repair-mediated deletions in the latter [3]. Additionally, the LTR-retrotransposons silencing and significant fractionation, which returned several genes to a single copy after multiple rounds of WGDs molded the *U. gibba* genome [4]. Therefore, the dynamic and specialized Lentibulariaceae genomes can provide a natural platform for the study of genome duplication, fractionation, and TEs expansion and silencing, and their impact on the evolution of the angiosperms [4,5,8].

Utricularia reniformis A.St.-Hil. is endemic to the Brazilian Atlantic Forest, growing as a terrestrial species in wet grasslands or as an epiphyte in moist habitats [1,24]. Some but limited information is available on the *U. reniformis* nuclear genome, such as its high levels of polymorphism [25], its genome size of 292 Mb (which is an intermediate size in comparison to other *Utricularia*) and a GC content of 38% [8]. We have also previously reported the sequence analyses of *U. reniformis* chloroplast (cpDNA) and mitochondrial (mtDNA) genomes [26,27]. In order to develop insights into the forces that have shaped the genomes of the Lentibulariaceae species, here we estimated the genome size and chromosome number, and deep-sequenced the terrestrial bladderwort *U. reniformis* genome and transcriptome. We also compared the *U. reniformis* draft genome against the aquatic species *U. gibba* and other angiosperms, revealing that the *U. reniformis* has a distinct genome structure, mostly due to lineage-specific WGDs and a TE expansion, reflecting a diversified repertoire of plant developmental and carnivory-associated genes and their underlying terrestrial adaptation.

2. Results

2.1. The Utricularia reniformis Genome

We first determined the chromosome number of *U. reniformis* as 2*n* = 40 (Figure 1A,B). The small chromosomes (~1.65 µm) varied between metacentric to submetacentric, with one pair holding a distended satellite, probably representing the ribosomal DNA 45S sites (see dots in Figure 1A and the chromosome pair 10 in Figure 1B). The genome size, estimated by the flow cytometry of the nuclei from leaves, was found to be 2C = 0.652 pg (SD = 1.48%), which represents a haploid genome size of 1C = 317.1 Mb (or 0.324 pg) (Table S1).

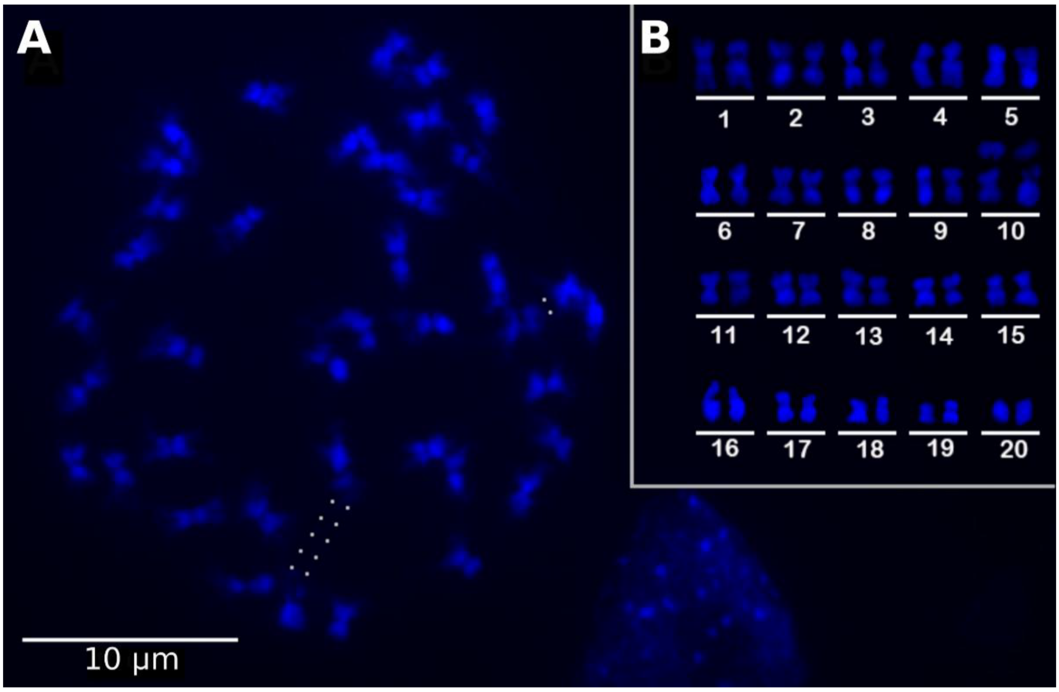


Figure 1. *Utricularia reniformis* metaphases with 2*n* = 40. (**A**) Metaphase stained with DAPI. (**B**) Karyogram using the same metaphase presented in (**A**). Dots in (**A**) represent the distended and unstained region. The bar is equivalent to 10 µm.

We generated 350 gigabytes of genomic and transcriptomic data for *U. reniformis* (Table 1). The k-mer frequency analysis demonstrated that *U. reniformis* exhibits a high heterozygosity rate of 1.8% (Figure 2 and Figure S1). In addition, the k-mer frequency plot was consistent with a tetraploid genome with the haploid at 149 Mb. Moreover, the predicted genome unique content and repeated content lengths are ~88 Mb (29%) and ~220 Mb (71%), respectively.

Table 1. DNA and RNA read produced for the *Utricularia reniformis* genome assembly and annotation.

	Technology	Library	Read Length	Raw Reads	Trimmed Reads
DNAseq	Illumina HiScan and MiSeq	Paired-end (~350 bp) Paired-end (~450 bp) Mate-paired (3–9 Kb)	100 bp 300 bp 100 bp	285,403,944 49,185,074 177,044,066	187,288,003 35,963,241 60,011,867
RNAseq	Ion Proton	Single-end - Leaves - Stolon - Utricules (preys)	~200 bp	46,622,745 41,894,450 112,414,083	40,853,284 34,716,966 97,821,388

The assembled draft genome spanned 304 Mb (96% of the flow cytometry-estimated genome size), with an average coverage of 145× (based on the number and length of aligned reads, divided by the assembled genome size), had an NG50 of 466-Kb and a BUSCO completeness of 87.8%, showing a duplication rate of ~26%, which might correspond with a partial genome duplication (Table S2) and 42,582 gene model predictions. The estimated number of error-free bases was 93% (283 of 304 Mb), which also indicated that multiple repeated genomic regions (e.g., non-autonomous LTR elements, centromeres, and telomeres) remained partial or non-assembled. However, this finding also supported that a large amount of the heterozygous content present in *U. reniformis* genome was represented in the assembled draft genome (Table 2 and Figure 2).

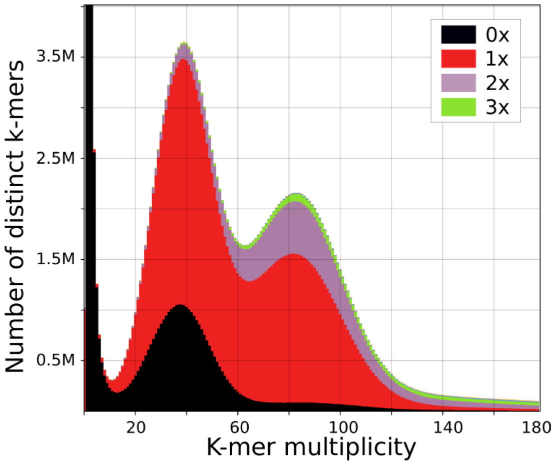


Figure 2. Genome assembly analysis: Read k-mer frequency versus *Utricularia reniformis* assembly copy number stacked histograms generated by the KAT comp tool. Read content in black is absent from the assembly, red occurs once, purple twice, and green occurs three times, indicating a scenario of both heterozygous and polyploidy genome. The heterozygous content is represented by the first peak at x = 40 and the homozygous content in the second peak at x = 80. The hidden content (the black peak) represents the heterozygous content that is lost when the assembly bubbles are collapsed. The genome assembly contains most (but not all) of the heterozygous content and introduces duplications on both heterozygous and homozygous content.

Table 2. *Utricularia reniformis* genome assembly status.

	U. reniformis Genome	
Total size of scaffolds (bp)	304,550,249	
Number of scaffolds	1830	
Number of contigs	5452	
Useful amount of scaffold sequences (≥25-Kb)	297,419,257	
% of assembled genome that is useful	93.8%	
Longest scaffold (bp)	1,862,935	
Longest contig (bp)	926,419	
Scaffolds longer than 1-Kb	1830 (100%)	
Scaffolds longer than 100-Kb	688 (37.6%)	
Scaffolds longer than 1 Mb	47 (2.6%)	
NG50 scaffold length (bp)	466,988	
LG50 scaffold count	196	
N50 contig length (bp)	161,226	
Percentage of assembly in scaffolded contigs	91	
Gaps number	3677	
Unknown bases (Ns) (bp)	5,790,542	
Average gap size (bp)/Longest gap (bp)	1575	10,802

Moreover, several cpDNA- and mtDNA-derived regions were also identified, thus, providing support for the lateral transfer between the organelles and the nuclear genome, as previously reported [26,27]. We also detected simple sequence repeat (SSRs) markers already developed for *U. reniformis* [25], in our draft genome. Among all markers, for two SSR loci, we found corresponding regions in *U. gibba* and *U. reniformis* (Table S3). In both cases, we recovered two scaffolds in *U. reniformis* and only one syntenic genomic region in *U. gibba* (Figure S2). For all other markers, we were not able to detect loci shared in both genomes. In addition to the *Arabidopsis*-type telomeric repeats identified at the scaffolds ends, we also found *Chlamydomonas*- and *Genlisea*-type telomeric repeats in their vicinity, as previously observed for *U. gibba* [5]. Furthermore, *U. reniformis* exhibited a considerable amount of TE-derived sequences, when compared to *U. gibba* (56% vs. 32%).

2.2. Structural Comparative Analysis

In order to infer *U. reniformis* genome structure and WGD history, we compared the draft genome to *U. gibba* and other angiosperms. In spite of the low assembly contiguity, we were able to detect at least one WGD round since the core eudicot γ whole-genome triplication (WGT), according to the Ks values for estimates of individual genome duplication events (Figure S3A). Moreover, blockwise relationships between *U. reniformis* and *U. gibba* were about 4:2 (Figure S3B), supporting that *U. reniformis* is a tetraploid, as compared to *U. gibba*. This finding also suggested that the two *Utricularia* genomes responded distinctly after the WGD events.

The *U. reniformis* and *U. gibba* structural genome differences become more obvious by comparative pairwise alignment. We observed 77% (SD:4%) average nucleotide identity (ANI) between the *U. reniformis* and *U. gibba* assemblies by reciprocal best blast (RBB). Additionally, we observed low global collinearity, which indicated that the *U. reniformis* genome presents a mosaic structure in comparison to *U. gibba* (Figure 3A,B). Approximately 114 Mb of the *U. reniformis* genome exhibited partial matches (average length of 84-Kb) to *U. gibba* (Table S4). The unaligned blocks corresponded to the segments exclusive to *U. reniformis*, which were mostly related to the TE-related regions (Figure 3B).

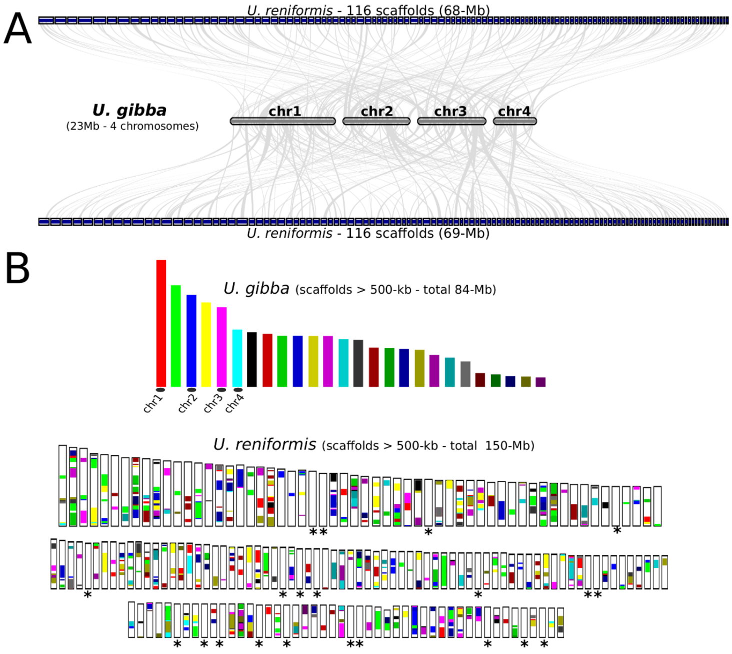

A
U. reniformis - 116 scaffolds (68-Mb)
U. gibba
(23Mb - 4 chromosomes)
chr1
chr2
chr3
chr4
U. reniformis - 116 scaffolds (69-Mb)
B
U. gibba (scaffolds > 500-kb - total 84-Mb)
chr1
chr2
chr3
chr4
U. reniformis (scaffolds > 500-kb - total 150-Mb)

Figure 3. *Utricularia reniformis* vs. *U. gibba* genome comparisons. (**A**) Macrosynteny karyotype visualization of *U. gibba* chromosome 1 (CM007989.1), 2 (CM007990.1), 3 (CM007991.1), and 4 (CM007992.1), against the corresponding *U. reniformis* scaffolds showing matches (generated with the MCscan tool with minspan = 10 option). (**B**) Dual synteny bar plot, created by the '*dual-bar plotter*' from the MCScanX. The upper panel represents *U. gibba* chromosomes and scaffolds; the lower panel represents the *U. reniformis* scaffolds. The syntenic blocks are shown in colors. Most of the white blocks present in each *U. reniformis* scaffolds corresponds to exclusive regions and transposable elements. The asterisks correspond to the *U. reniformis* scaffolds showing no matches to *U. gibba* chromosomes and scaffolds.

Additionally, we mapped less than 37% of the *U. reniformis* trimmed paired-end reads consistently with the right distance and orientation onto the *U. gibba* assembly, which also supported the considerable structural differences between these species (Table S5). As a result, it was not possible to employ the reference-assisted assembly procedure to anchor the *U. reniformis* scaffolds into the *U. gibba* fully-assembled chromosomes to reconstruct potential pseudo-molecules.

Moreover, in comparison to other angiosperms, *U. reniformis* also showed a lower percentage of macrosyntenic blocks. For instance, *U. reniformis* retained at least 47 Mb (average length of 76-Kb and ANI of 82%), 35 Mb (average length of 46-Kb and ANI of 76%), and 53 Mb (average range of 63-Kbp and ANI of 77%) of shared blocks to *A. thaliana*, *V. vinifera*, and *S. lycopersicum*. In contrast, its closest relative, *U. gibba,* retained 71 Mb (average length of 44-Kb and ANI of 80%), 57 Mb (average length of 26-Kb and ANI of 75%), and 58 Mb (average length of 35-Kb and ANI of 77%) blocks, respectively (Table S4).

We further evaluated the microsyntenic level and verified that *U. reniformis* and *U. gibba* present distinct patterns of retention and alternative deletion of duplicated genes between the polyploid subgenomes, which is consistent with the distinct post-speciation evolutionary paths (Figure 4A–C). This feature is also observed among the *Utricularia* species and *A. thaliana*, yet with fewer microsynteny blocks (Figure 4D). When both *Utricularia* genomes are compared to the phylogenetically distant eudicotyledons, such as *V. vinifera* and *S. lycopersicum,* a significant fractionation is prominent (Figure 4E,F).

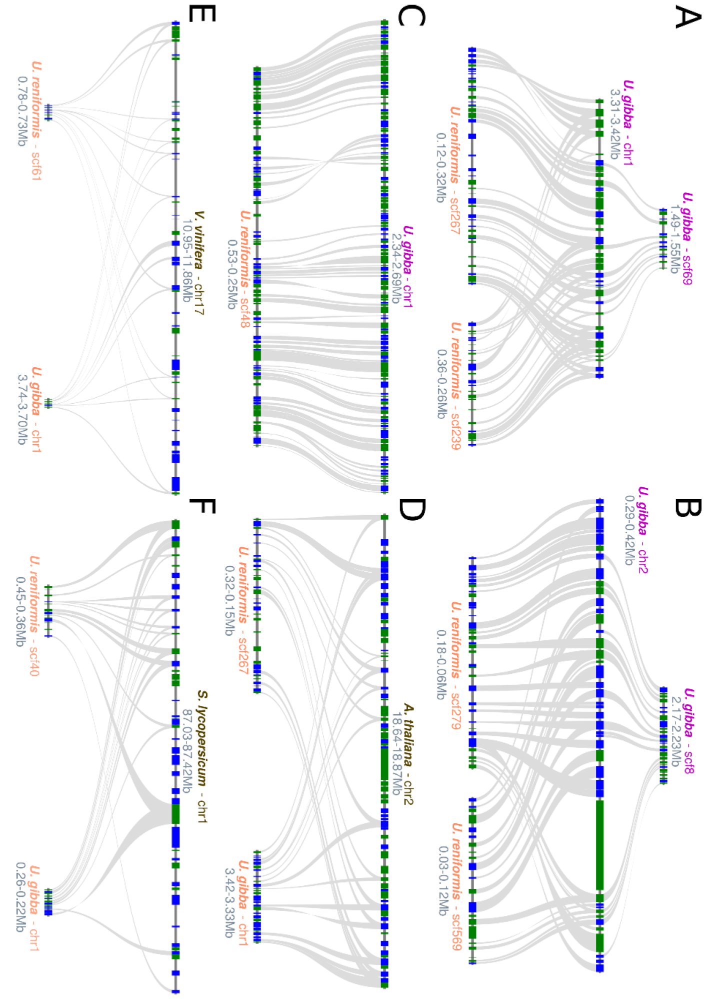


Figure 4. Microsynteny analysis of *Utricularia reniformis* against *U. gibba* and other eudicotyledons: (**A**,**B**) Polyploid subgenomes of *U. reniformis* vs. *U. gibba* microsynteny showing an alternative deletion of the duplicated genes. (**C**) Highly conserved regions of *U. reniformis* vs. *U. gibba* microsynteny. (**D**) *Arabidopsis thaliana* vs. *U. reniformis*, with *U. gibba* showing moderately conserved regions of microsynteny. (**E**) *Vitis vinifera* vs. *U. reniformis*, with *U. gibba* showing a considerable fractionation. (**F**) emphLycopersicon esculentum vs. *U. reniformis*, with *U. gibba* showing a moderate fractionation.

However, in general, the two *Utricularia* species were far more similar in gene order to each other than either were to *A. thaliana*, *V. vinifera*, and *S. lycopersicum*. The phylogenetic and ANI analyses based on 336 core eukaryotic genes corroborated our microsynteny analysis, supporting the same common ancestor for *U. reniformis* and *U. gibba*, and also for the *Utricularia* and *Genlisea* clades (Figure 5).

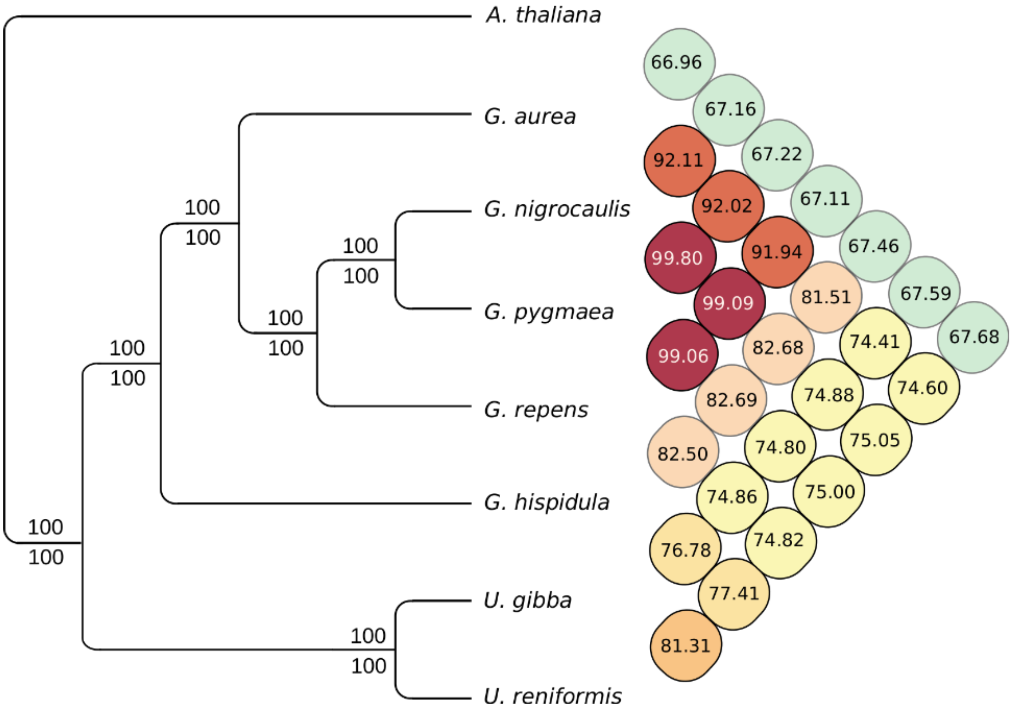


Figure 5. Overall similarity heatmap and maximum likelihood and Bayesian inference phylogenetic tree of 336 core-eukaryotic genes shared among seven carnivorous plant genomes (*Utricularia reniformis*, *U. gibba*, *Genlisea aurea*, *G. nigrocaulis*, *G. pygmaea*, *G. repens*, and *G. hispidula*) and *Arabidopsis thaliana* as an out-group. Numbers above and below the lines indicate the maximum likelihood bootstrap values and the Bayesian posterior probabilities for each clade. The *G. pygmaea* and *G. repens* genomes used in this analysis are yet to be published. However, the concatenated core nucleotide gene fasta sequence used to construct this dataset is freely available at http://doi.org/10.5281/zenodo.3268745.

2.3. *Utricularia reniformis* Comparative Annotation

In general, *U. reniformis* and *U. gibba* show different arrays of gene duplication that might directly reflect support that they are retaining genes in a biased manner. This was exemplified by the more significant number of tandem, proximal, and dispersed gene pairs identified in *U. reniformis*, when compared to *U. gibba* (Table 3). In contrast, *U. gibba* displayed more singletons and segmental duplicates. However, the more significant number of segmental copies identified in *U. gibba* might be related to the assembled genome using long-reads, which permitted a better identification of large syntenic blocks, in comparison to the *U. reniformis* genome based on short-reads. Moreover, the considerable number of singletons in *U. gibba* might also represent retained duplicates that have reverted to a single copy after the WGD, which supported the genome size reduction observed in this species.

It is noteworthy that the gene space in *U. reniformis* represents ~26% of the genome, which also demonstrated that the intergenic regions are replete with TEs-related sequences constituting more than half of the genome size. Conversely, TEs covered ~32% of *U. gibba* genome, while the gene space represented ~51%.

Table 3. *Utricularia reniformis* and *U. gibba* comparative and structural annotation.

	U. reniformis	***U. gibba***
Total number of identified genes	42,582	25,509
- Singletons	3083	5120
- Dispersed duplicates	26,546	8831
- Proximal duplicates	1679	683
- Tandem duplicates	2994	999
- Segmental duplicates	8280	9876
Annotation status		
- Annotated genes †	35,899	21,283
- Genes with GOs	27,751	17,760
- Unknown and Hypothetical genes	6683	4348
Total gene length (bp)	79,419,967	51,346,373
Total exon length (bp)	44,971,966	28,891,859
Total intron length (bp)	29,871,202	16,875,684
Longest gene (bp)	84,570	61,898
Longest exon (bp)	7309	6002
Longest intron (bp)	78,774	53,378
Longest CDS (bp)	15,201	15,801
Mean gene length (bp)	1872	2016
Mean exons per gene	5	5
% of genome covered by genes	26	51
% of genome covered by CDS	15	38
% of genome covered by TEs-like regions	56	32

† high-confidence genes models (showing a potential product, GO term, and hits to UniProt viridiplantae clade).

2.4. *Utricularia reniformis Shows a Massive Expansion of LTR from the Gypsy Superfamily*

Utricularia reniformis presents a prominent LTR-retrotransposons expansion, representing up to 145 Mb (Figure 6 and Table S6). In general, a separate array of expansion and contraction of distinct evolutionary lineages from the *Copia* and *Gypsy* superfamilies stand as the main differences between *U. reniformis* and *U. gibba*. The most noticeable increase of the *Gypsy* superfamily was observed within the *Ogre* evolutionary lineage, accounting up to 72 Mb of *U. reniformis* genome. It is noteworthy that the LTR-LARDs, *Angela*, *CRM*, are more prevalent in *U. gibba*, including the *Athila* evolutionary lineage, which was exclusively found in *U. gibba.* In contrast, from the *Copia* superfamily, the evolutionary lineages *Alesia, Bianca, Ikeros*, and *SIRE* were found solely in *U. reniformis.* As observed in other plant genomes, the Class II elements were less prominent in both species, and except for the *Helitron* super-family which we were not able to identify in *U. gibba*, mostly known super-families and evolutionary lineages commonly found in plant genomes were present.

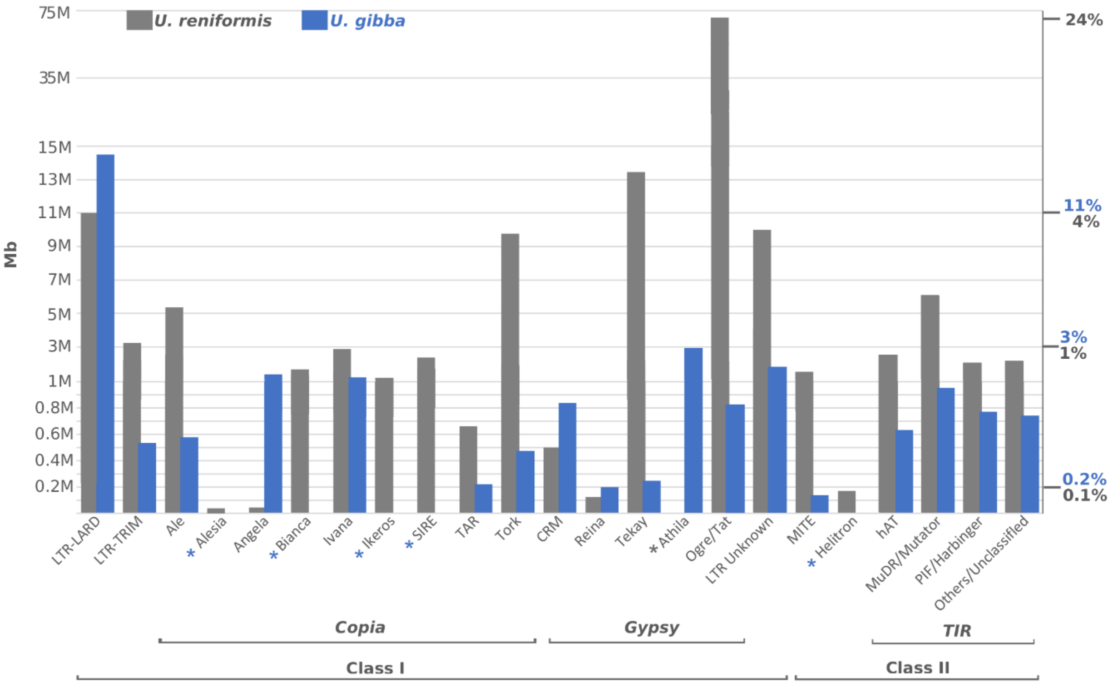


Figure 6. Transposable elements discovery in *Utricularia reniformis* and *U. gibba* genomes: Length occupied (bp) of each super-family and evolutionary lineage. A detailed distribution of each element identified is presented in Table S6. The *y*-axis on the left indicated the total length occupied in the megabases of each evolutionary lineage, whereas the *y*-axis on the right meant the % of the corresponding genome. Blue and Grey asterisk correspond to the absent evolutionary lineages in *U. gibba* and *U. reniformis*, respectively.

2.5. Distinct Functional Enrichment Patterns Among Tandem and Dispersed Duplicate Genes

Self–self-genome comparisons among the singletons and duplicated genes (dispersed, proximal + tandem, and segmentals) revealed exciting patterns of enriched GO terms and KEGG enzymes ($p < 0.05$, Fisher's exact test, Bonferroni corrected) between *U. reniformis* and *U. gibba* (Tables S7–S9). Both species retained a conserved arrangement of cellular components related to organelles and nucleus among the singletons, such as, chloroplast (GO:0009507), thylakoid (GO:0009579), mitochondrion (GO:0005739), nucleoplasm (GO:0005654), nuclear envelope (GO:0005635), endoplasmic reticulum (GO:0005783), and biological processes related to photosynthesis (GO:0015979) and DNA metabolic process (GO:0006259). Additionally, KEGG enzymes with associated functions that act on NADH or NADPH (EC:1.6.99.5), NADH dehydrogenase (EC:1.6.99.3), NADH—ubiquinone reductase (EC:7.1.1.2) were also augmented among the singletons. Together, these findings support a conserved organellar functional activity including the control of energy homeostasis functions among the singletons.

Among the dispersed duplicates, the nuclease activity (GO:0004518) was enriched in both species. However, the hydrolase activity (GO:0016787), transcription regulator activity (GO:0140110), DNA-binding transcription factor activity (GO:0003700), and tropism (GO:0009606) were significantly overrepresented in *U. reniformis*, while carbohydrate metabolic process (GO:0005975), transport (GO:0006810), kinase activity (GO:0016301), and lipid metabolic process (GO:0006629) were enriched in *U. gibba*.

It was previously observed in many angiosperm genomes that secondary metabolic function and transcriptional function are enriched among tandems and segmental duplicates, respectively [5,28–30]. Our results corroborate this observation for secondary metabolic function in the tandem repeats of both *Utricularia*, and the transcriptional functions only in *U. gibba* segmental duplicates. Interestingly, and opposite to what has been previously observed, the transcriptional features were augmented in *U. reniformis* dispersed duplicates genes. Since the dispersed copies could arise from TEs movements, and the *U. reniformis* genome had a high level of LTR-elements, this result might suggest a central role of LTR transposition in molding these genomic differences.

We also checked the potential role of tandem and proximal duplicates with the evolution of novel functions, such as those related to carnivory and the ROS metabolism, which revealed an enrichment of many GO terms and KEGG enzymes in both species (Figure 7). For instance, catalytic activity (GO:0003824), biotic stimulus (GO:0009607), transferase activity (GO:0016740), and the cellular components—lysosome (GO:0005764), lytic vacuole (GO:0000323), and the cell wall (GO:0005618) were overrepresented. The enrichment of the cell wall (GO:0005618) term, could be related to the trap movement during the prey capture, which demands profound and dynamic cell-wall changes [5], and the lysosome and lytic vacuole enrichment might be related to the digestion of preys. However, we also identified distinct GO and KEGG enrichment arrangement among species, strongly suggesting a functional divergence, subfunctionalization, or neofunctionalization among the tandem and proximal duplicates (Figure 7). Comparatively, *U. reniformis* showed an enrichment of response to chemicals (GO:0042221), a response to external stimulus (GO:0009605), and signaling (GO:0023052), whereas, as observed in the dispersed duplicates, the lipid metabolic process (GO:0006629) was also augmented in *U. gibba* tandem duplicates. Regarding the well-known carnivory [4,5] and ROS enzymatic functions, *U. gibba* presented chitinase (EC:3.2.1.14), cysteine protease (EC:3.4.22), and peroxidases (EC:1.11.1.7) enriched in tandem duplicates regions, as previously described [5], whereas *U. reniformis* exhibited glutathione transferases (EC:3.4.16), carboxylesterase (EC:2.5.1.18), acylglycerol lipase (EC:3.1.1.23), pectinesterase (EC:3.1.1.11), and enzymes that acted on the paired donors, with an incorporation or reduction of molecular oxygen (EC:1.14.13). Therefore, these findings indicated that both species presented distinct enrichment patterns of GO terms and KEGG enzymes, supporting an ongoing gene birth–death–innovation process that included metabolism and carnivory-associated functions, among the tandem and proximal duplicated genes.

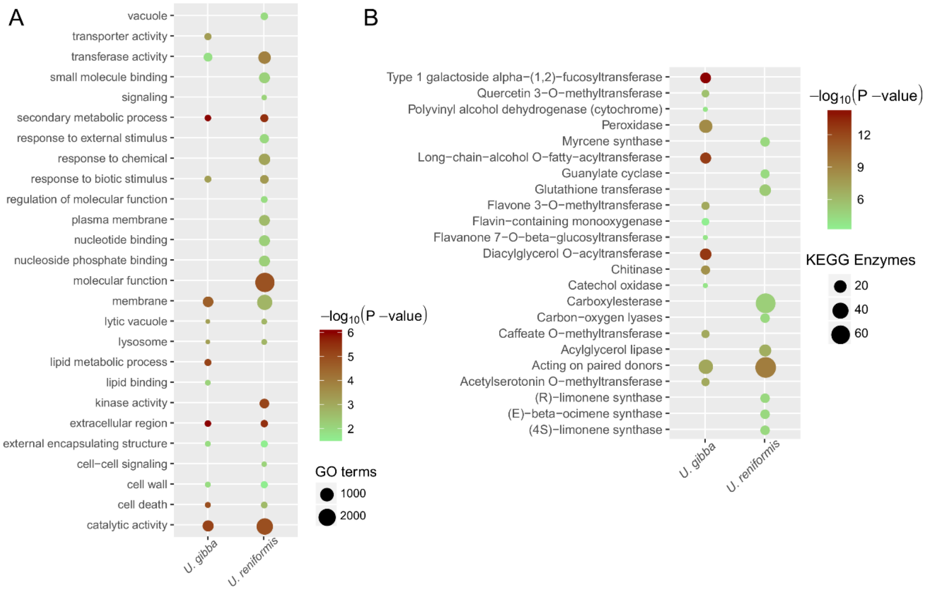


Figure 7. Functional enrichment comparative analysis of tandem duplicates among *Utricularia reniformis* and *U. gibba*: The enriched GO terms and KEGG enzymes with the corrected *p*-value < 0.05 are presented. The color of the circle represents the statistical significance of the enriched GO terms (**A**) and the KEGG Enzymes (**B**). The size of the circles represents the number of occurrences of a GO term (**A**) and the KEGG Enzyme (**B**).

Interestingly, among the segmental duplicates, similar patterns of functions related to the primary metabolism, and the plant developmental process were enriched in both species, for example the amide biosynthetic process (GO:0043604), anatomical structure development (GO:0048856), reproduction (GO:0000003), nitrogen compound metabolic process (GO:0006807), peptide metabolic process (GO:0006518), and the primary metabolic process (GO:0044238).

2.6. Utricularia reniformis Displays Unique Patterns of Carnivory-Associated, Land Adaptation, and Developmentally Related Genes

Previous studies, including literature review and comparative annotation approaches, accurately identified GO terms and KEGG enzymes strictly related to the carnivory-associated functions among the sequenced carnivorous plants available in public databases [31,32]. Based on this data, we investigated our functional annotation results and observed that *U. reniformis* had an extensive gene repertoire of carnivory-associated and ROS-related functions (Table 4). In addition, we also identified an extensive repertoire of plant development, reproduction, and life-form adaption GO terms in *U. reniformis* (Table 5), which might be related to the different life-forms in comparison to *U. gibba*. Moreover, distinct patterns of ABC transporters were observed between *U. reniformis* and *U. gibba* (Figure 8A,B). Terrestrial plants exhibited, on average, a repertoire of 128 ABC transporters, while the aquatic plants had fewer copies [33]. At least 132 and 86 ABC transporters were identified in *U. reniformis* and *U. gibba*, respectively (Figure 8A and Table S10), suggesting that *U. reniformis* had more copies due to its terrestrial adaptation. This was especially apparent from the expansion of the ABC transporter B (Figure 8A and Figure S4) and ABC transporter C families, which are commonly related to developmental processes and functions necessary for life on dry land, and the ABC G family (Figure 8A and Figure S5), which are related to response to biotic stress [33]. Furthermore, plant developmental-related transcription factor (TFs) containing the wuschel-like homeobox (WOX), homeobox-leucine zipper (HD-Zip), agamous-like (MADS-box), TCP, WRKY, RADIALIS, DIVARICATA, DICHOTOMA, and the scarecrow-like gene families present distinct patterns of expansion and contraction among both species (Figure 8B and Tables S11–S14), and thus, provide support for the two distinct life-forms between these species.

Table 4. Carnivory-associated and reactive oxygen species (ROS) detoxification GO terms distribution. *Utricularia reniformis* (7200 unique genes: Singletons—3.5%, Dispersed—65.5%, Tandem and Proximal—10.5%, and Segmental—20.5%) and *U. gibba* (3918 unique genes: Singletons—10%, Dispersed—48%, Tandem and Proximal—6%, and Segmental—36%).

Gene Ontology Term	GO Code	*U. renif* n# Genes	*U. gibba* n# Genes
amidase activity	GO:0004040	13	10
actin filament	GO:0005884	20	11
alpha-galactosidase activity	GO:0004557	15	13
alternative oxidase activity	GO:0009916	6	4
ammonium transmembrane transport	GO:0008519; 0072488	18	16
aspartic-type endopeptidase activity	GO:0004190	289	94
ATP:ADP antiporter activity	GO:0005471	9	9
ATPase activity	GO:0016887	1831	792
beta-galactosidase activity	GO:0004565	720	125
catalase activity	GO:0004096	17	11
cellulase activity	GO:0008810	62	34
chitinase activity	GO:0004568	21	13
cinnamyl-alcohol dehydrogenase activity	GO:0045551	34	15
cyclic-nucleotide phosphodiesterase activity	GO:0004112	1	1
cysteine-type peptidase activity	GO:0008234	384	185
nuclease activity	GO:0004518	1955	1094
fructose-bisphosphate aldolase activity	GO:0004332	12	10
glutathione transferase activity	GO:0004364	54	21
glutathione peroxidase activity	GO:0004602	14	9
hydrolase activity, acting on ester bonds	GO:0016788	2767	1259
heat shock protein activity	GO:0042026; 0006986; 0034620	129	72

Table 4. *Cont.*

Gene Ontology Term	GO Code	*U. renif* n# Genes	*U. gibba* n# Genes
lipase activity	GO:0016298	231	140
lipid transport	GO:0006869	170	113
myosin heavy chain kinase activity	GO:0016905	36	23
peroxidase activity	GO:0004601	192	140
peptidase activity	GO:0008233	2650	1269
phosphatase activity	GO:0016791	287	186
phospholipase activity	GO:0004620	80	43
polygalacturonase activity	GO:0004650	109	50
polygalacturonase inhibitor activity	GO:0090353	5	1
protein homodimerization activity	GO:0042803	809	525
ribonuclease activity	GO:0004540	390	154
serine-type carboxypeptidase activity	GO:0004185	82	38
superoxide dismutase activity	GO:0004784	17	10
symplast	GO:0055044	1060	679
urease activity	GO:0009039	2	1
water channel activity	GO:0015250	46	26

Table 5. Distribution of GO terms related to the biological process related to plant development, reproduction, and life-form adaptation. (*U. reniformis*; 11,468 unique genes: Singletons—3%, Dispersed—62%, Tandem and Proximal—10%, and Segmental—23%. *U. gibba*; 7071 unique genes: Singletons—11%, Dispersed—41%, Tandem and Proximal—4%, and Segmental—43%).

Gene Ontology Term	GO Code	*U. renif* n# Genes	*U. gibba* n# Genes
developmental process involved in reproduction	GO:0003006	1965	1237
reproduction	GO:0000003	2546	1600
reproductive process	GO:0022414	2223	1388
multicellular organism development	GO:0007275	4749	2773
embryo development	GO:0009790	744	465
post-embryonic development	GO:0009791	2447	1578
flower development	GO:0009908	915	601
developmental maturation	GO:0021700	264	171
developmental process	GO:0032502	4126	2704
reproductive structure development	GO:0048608	1685	1065
anatomical structure development	GO:0048856	4046	2639
cellular developmental process	GO:0048869	1352	880
reproductive shoot system development	GO:0090567	655	437
pollination	GO:0009856	343	210
tropism	GO:0009606	215	129
circadian rhythm	GO:0007623	261	161
response to stress	GO:0006950	4513	2869
response to radiation and light stimulus	GO:0009314, GO:0009416	1012	711
response to external stimulus	GO:0009605	2108	1352
response to biotic stimulus	GO:0009607	1526	981
response to abiotic stimulus	GO:0009628	3161	2041
response to endogenous stimulus	GO:0009719	2665	1693
response to chemical	GO:0042221	4590	2984
response to stimulus	GO:0050896	6579	4277

2.7. Comparative Annotation Among Other Angiosperms Genomes Reveals Species-Specific Genes Strictly Related to the Environment and Life-Form Adaptations

The comparative orthologous gene cluster analysis revealed that approximately 60% of *U. reniformis* and *U. gibba* genes are shared, indicating that they were inherited from the last common ancestor. Interestingly, *U. reniformis* presents an almost duplicated core gene set from that observed in *U. gibba*, *V. vinifera*, *A. thaliana*, and *S. lycopersicum*, and thus, supports a species-specific WGD event

(Figure 8C and Table S15). In general, a significant fraction of the species-specific genes (genes for which no orthologs could be found in any of the other species compared) and single-copy gene clusters from *U. reniformis* and *U. gibba* encode uncharacterized or hypothetical proteins (63% for *U. reniformis* and 50% for *U. gibba*). Among the species-specific genes, we highlighted ABC transporters, and several developmental-related TF gene families (Figure 8A,B), supporting the role of the species-specific genes with plant development, reproduction, prey strategies, and life-from adaptations. GO enrichment analysis ($p < 0.05$, Fisher's exact test, Bonferroni corrected) among the species-specific genes revealed an augmented functions related to the reproductive process (GO:0022414), hydrolase activity (GO:0016788), nuclease activity (GO:0004518), catalytic activity (GO:0003824), protein binding (GO:0005515), and organelles (GO:0043226) in both species. However, *U. reniformis* present a unique enrichment of kinase activity (GO:0016301), transferase activity (GO:0016740), carbohydrate metabolic process (GO:0005975), lytic vacuole (GO:0000323), and lysosome (GO:0005764), whereas, *U. gibba* exhibited cellular nitrogen compound metabolic process (GO:0034641), peptide metabolic process (GO:0006518), and signal transduction (GO:0007165). Taken together, these results indicate that the species-specific gene set might play an essential role in the diversification and adaptation to different life-forms.

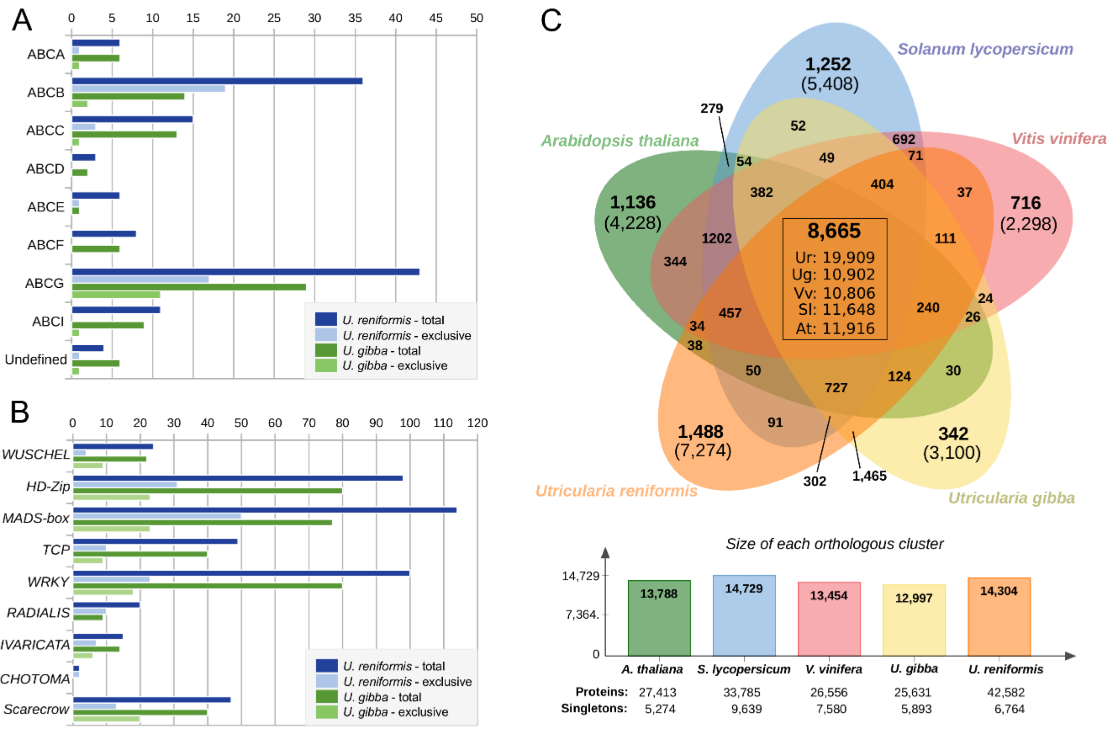


Figure 8. (**A**) Distribution of ABC transporters and (**B**) developmental-related transcription factors among *Utricularia reniformis* and *U. gibba*. Exclusive genes are species-specific genes for which no orthologs could be found in any of the other species compared. (**C**) Venn diagram showing the distribution of shared orthologous gene families among *U. reniformis*, *U. gibba*, *Arabidopsis thaliana*, *Vitis vinifera*, and *Lycopersicon esculentum* (using an MCL inflation factor of 1.5). The values present inside the box correspond to the number of core genes of each species. The number present inside the parentheses represent the total number of species-specific genes from each species cluster.

3. Discussion

3.1. The Role of WGD in *Utricularia reniformis* Genome Evolution

Polyploid occurred multiple times during angiosperms evolution, and this is not an exception for *Utricularia* lineages. For instance, multi-way microsynteny analysis revealed three sequential

WGD events in *U. gibba*, in addition to the γ WGT event shared by all core eudicots [4,5,34]. Additionally, out-crossing events might be broad among the *Utricularia* clade, and a comparative analyses revealed that the most recent *U. gibba* WGD was derived from an allopolyploidization event [5]. Allopolyploidization is recognized as a driving evolutionary innovation and is often associated with speciation when resulting in a duplicated genome structure [35–37]. Therefore, the current hypothesis is that not only was the *U. gibba* genome shaped by WGD events, but other *Utricularia* genomes and specifically the *U. reniformis* genome were shaped by WGD-induced variation and adaptation.

After a polyploidy event, each lineage undergoes distinct patterns of gene retention and fractionation, which generally leads to the genome returning to the diploid state [38]. Our results for *U. reniformis* and *U. gibba* support that each species has distinct patterns of deletions, duplicated genes, and rearrangements, suggesting that after speciation and WGD, they took distinct evolutionary paths that were consistent with their life-histories. This is exemplified by the significant number of tandem, proximal, and dispersed gene pairs identified in *U. reniformis*, when compared to *U. gibba*. Indeed, gene duplication is recognized as one of the major sources of evolutionary innovation [39]. For instance, tandem and proximal duplicated genes are often associated with the emergence of functional novelty, which commonly originated from isolated events in which an individual gene gets duplicated by unequal crossing-over between similar alleles [39]. Together with a large number of dispersed duplicates identified, the tandem duplicated genes might also arise by translocations mediated by TEs [39], generating single-gene transposition-duplication, and therefore, indicating for an essential role of TEs molding the genome structure and evolution. Additionally, WGD analysis confirmed that *U. reniformis* experienced at least one WGD round since the core eudicot WGT. However, it should be taken into account that the low assembly contiguity based on short-read technology hindered the identification of large syntenic blocks, and thus, complicated the evaluation of the complete history of *U. reniformis* WGD events.

3.2. LTR-Retrotransposons are Key Agents Governing Utricularia Genome Size Changes

One of the most common repetitive content of *U. reniformis* is related to TEs-related sequences, corresponding to at least 56% of the genome. The TEs drive the genome evolution, mainly by rearrangements, gene promoters repression, enhancers disruption, epigenetic regulation, and also by inflating the genome size by its copy-and-paste mechanism [40,41]. Apart from their deleterious effects, TEs also provide genome variation that can lead to the plant adapting more rapidly to new environmental conditions, and thus, leading to speciation [42,43]. The increased number of TEs in *U. reniformis* vs. *U. gibba* suggests that TEs might be playing a role in the genome innovation we are observing, but it could also just reflect a difference in the ability of the two genomes to purge mobile TEs.

The most prominent TE expansion in *U. reniformis* is related to LTR-retrotransposons elements from the *Gypsy* superfamily, which was also observed in the Solanaceae and Brassicaceae families [44,45]. In particular, elements from the *Tat* lineage were the most prominent, which are commonly broader in other plants, constituting up to 40% of the genome of some species [46].

Except for LTR-LARDs, all other TE groups are expanded in *U. reniformis*. The LTR-LARD and *CRM* are generally located in complex and highly repeated genomic loci, such as heterochromatic and pericentromeric regions [47,48]. These regions are often partially assembled using short-read technology, and we verified that most LTR-LARDs identified in *U. reniformis* are incompletely assembled, suggesting that they were partially determined and underestimated. Therefore, LTR-LARD and *CRM* elements might potentially represent a considerable fraction of the genome that remains unresolved in the *U. reniformis* assembly, and the results presented might be biased due to the assembly contiguity limits.

3.3. The Genomic Landmarks for Terrestrial Adaptation

Adaptation to the environment and life-form plasticity are hallmarks of the selective pressures that govern genome evolution in the *Utricularia* lineage. A common reproductive strategy for several

Utricularia is to produce clones by stolon fragmentation, which is a usual asexual reproduction of aquatic species from the section *Utricularia* (the section in which *U. gibba* is nested), with some species being sterile and rarely flowering [49,50]. In contrast, *U. reniformis* presents a specialized sexual reproductive strategy that is entirely dependent on pollinators, and is specifically dependent on self-pollination [51]. Moreover, *U. gibba* seems to have a more severe degree of Fuzzy Arberian Morphology, such as no clear delimitation of distinct vegetative organs [52]. In contrast, *U. reniformis* presents a more traditional vegetative organ delimitation (as stems and leaves), similar to other angiosperms.

Consequently, the vegetative plasticity found in different *Utricularia* species might be essential strategies used to colonize the distinct habitats. Considering the different environments (water, soil, rock surface, etc.), the various trap designs [53–55] might be shaped as adaptation for different prey faunas [56–59]. Ultimately this might impact the absorption and assimilation of nutrients. At the genomic scale, both species display distinct patterns of development-related genes and TFs, such as WOX, HD-Zip, MADS-box, WRKY, TCP, RADIALIS, DIVARICATA, DICHOTOMA, and scarecrow-like gene families. These TFs play specialized roles in plant growth and development by regulating cell division and differentiation, in particular, the trap and floral meristem and trichome development, and might also develop an important role in the regulation of flowering time and the ontogeny of reproductive organs [16,60–64]. For instance, the WOX1 (3 vs. 5 species-specific genes) from the WOX family, is possibly associated with the trap development [16] SOC1 (11 vs. 6 species-specific genes) from the MADS-box family, which might develop a role with phosphorus scavenging from the trapped prey [4], and ATHB51 (5 vs. 6 species-specific genes) related to the leaf morphogenesis. These findings are suggestive to a different array of target genes, and the processes that these TFs control, and in addition to the distinct number of developmental genes identified in both species, might apply in different body plan adaptations and prey strategies.

Moreover, both species display a distinct repertoire of ABC transporters that are considered essential to terrestrial life-form adaption [33]. Some ABC transporters members are exclusively found in *U. reniformis*. For instance, the ABCB member 9 is related to plant hormone transport, regulation of growth, and in development [65]. Additionally, the mitochondrial ABCB family member 25 was directly involved with the molybdenum cofactor biosynthesis, and nitrogen assimilation from the soil [66]. Interestingly, aquatic plants mainly rely on ammonium as their primary nitrogen source, and the molybdenum cofactor biosynthesis deficiency is commonly associated with the use of ammonium as an alternative nitrogen source in habitats with increased relative humidity [66]. This suggests that the absence of ABCB members 25 in *U. gibba* might be directly related to their aquatic adaptation. Finally, the ABCG member 36, exclusively found in *U. reniformis*, is related to the export of related defence compounds and callose deposition at the site of pathogen contact to restrict pathogen development [67,68], supporting that *U. reniformis* needs extra protection for cellular damage, which is a common characteristic of terrestrial plants.

It is known that carnivorous plants experience different and variable levels of nutritional stress [69], and the emergence of the carnivory-associated function in each lineage was a result of multiple evolutionary paths [31]. Interestingly, *U. reniformis* presents a distinct and augmented set of well-known carnivory-associated functions. Among them, alpha-galactosidase, amidase, actin filament, aspartic-type endopeptidase, beta-galactosidase, cellulase, chitinase and cysteine-type peptidase, myosin heavy chain kinase, and polygalacturonases might be related to carbon and energy cycles maintenance with cellulose and its associated polymers-decomposing activity, suggesting a potential role in the breakdown of prey polysaccharides. Moreover, the enrichment of ATPase activities might be correlated with the trap acidity and molecular transport functions, such as the release of digestive enzymes and the absorption of digested material from the preys [5,31,70].

Conversely, in addition to carnivory-associated functions, the ROS-related enzymes are also augmented in *U. reniformis*. For instance, catalase, superoxide dismutase, peroxidase, and glutathione transferase might contribute to cellular homeostasis and impacting the control of the high respiration rates possibly originated through the carnivory process. In addition, *U. reniformis* exhibits much more dispersed, tandem, and proximal duplicated genes. The emergence of tandem duplicated genes might contribute to the evolution of novel functions and adaptation. Our results provide support for an ongoing gene birth–death–innovation, occurring mainly among tandem and dispersed duplicates, impacting a number of different GO terms and KEGG enzymes, including carnivory-associated functions across both *Utricularia*. This process might give a fine-tuning of carnivory-associated functions in each selective environment (aquatic and terrestrial). For instance, a high diversity of bacteria, fungi, algae, and protozoa compose the ecosystem inside the *Utricularia* trap and act synergistically to convert complicated organic matter into a nutrients source for the plants [71]. The trap ecosystem and prey biota can significantly vary between the terrestrial and aquatic habitat. The aquatic *Utricularia* species usually have a prey spectra dominated by copepodids, cladocerans, ostracods, rotifers, and aquatic insect larvae, while terrestrial species can also have acari, nematodes, and other terrestrial microorganisms [56,58,69,72,73]. Moreover, the environment and the capacity to capture preys can be determined by the trap and prey size [74]. Thus, the trap size difference between the two species (*U. gibba* possess traps c. 0.7–1.5 mm and *U. reniformis* c. 1–2.5 mm long; [1]) can also be the result of selection pressure for the different adaptation to aquatic (*U. gibba*) and terrestrial and epiphytic life-forms. Therefore, both plants might present a distinct trap ecosystem and an enzymatic digestive repertoire to uptake nutrients, and our results supported this hypothesis at a genomic scale. However, this hypothesis warrants further functional studies for confirmation.

4. Materials and Methods

4.1. Plant Material, Genome Size Estimation, and Cytogenetic Analysis

Utricularia reniformis samples were collected in the fall of 2015 in the Serra do Mar Atlantic Forest, in the Municipality of Salesópolis, SP, Brazil, and were deposited in the Herbarium JABU at the São Paulo State University (voucher—V.F.O. de Miranda et al., 1725). The plants were grown in a jar culture before procedures. These samples were recorded in the Brazilian National System of Management of Genetic Heritage and Associated Traditional Knowledge (SisGen) under the access number #A68D114, in accordance with the Brazilian Access and Benefit Sharing (ABS) legal framework introduced by the Federal Law No. 13,123 of 2015. No permission for collecting was necessary, as the sample was not collected in protected areas and *U. reniformis* is not a threatened species, according to the global IUCN (The IUCN Red List of Threatened Species: http://www.iucnredlist.org) and the Brazilian List of Threatened Plant Species.

For the genome size estimation, approximately 25 mg of leaf tissue was macerated in 1 mL of cold Ebihara buffer [75] supplied with 0.025 µg mL^{-1} RNAse, using a scalpel blade to release the nuclei into suspension with the same mass of the internal reference standard *Raphanus sativus* var. Saxa (2C = 1.11 pg; [76]). The nuclei suspensions were stained by adding 25 µL of a 1 mg mL^{-1} solution of propidium iodide (PI, Sigma). The analysis was performed using the FACSCanto II cytometer (Becton Dickinson, San Jose, CA, USA). The histograms were obtained through the FACSDiva software based on 5000 events, and the statistical evaluation was performed using the Flowing Software 2.5.1 (http://www.flowingsoftware.com/). Three individuals were analyzed in triplicates. The quality control of samples was based on the coefficient of variation (CV) of each measurement (which should be below 5%) and the standard deviation (SD) among the 2C-values (which should be below 3%). Such limits ensure that the variations observed inside and among measurements are due to technical factors and do not represent intraspecific variation among individuals [77].

For the cytogenetic analysis, stolon tips were pre-treated in 8-hydroxyquinoline (0.002 M) for 24 h, at 10 °C, fixed in ethanol:acetic acid (3:1, *v/v*) for 24 h, at room temperature, and were stored at −20 °C. The fixed stolon tips were washed in distilled water and digested in a 2% (*w/v*) cellulase (Onozuka)/20% (*v/v*) pectinase (Sigma-Aldrich, St. Louis, MO, USA)/1% macerozyme (Sigma-Aldrich, St. Louis, MO, USA) solution, at 37 °C for 15 min. The meristems were squashed in a drop of 45% acetic acid, and the coverslip was later removed in liquid nitrogen. Metaphases were DAPI (1 µg·mL^{-1}) stained and photographed with an XM10 camera coupled to a BX 53 Olympus epifluorescence microscope.

4.2. Genome and Transcriptome Sequencing

One individual was sampled, producing the three different DNA libraries used in this study. DNA was extracted using the QIAGEN DNeasy Plant Maxi Kit extraction protocol (Qiagen, Hilden, Germany) to prepare libraries for Illumina TruSeq PCR-free and Nextera XT paired-end reads, with ~350 bp (2 × 100 bp) and ~450 bp (2 × 300 bp), insert sizes, and Illumina Nextera mate-pair gel free with insert sizes ranging from 3.5 to 10 Kb (2 × 100 bp). Illumina libraries were sequenced on HiScanSQ and MiSeq machines. The paired reads (2 × 100 bp and 2 × 300 bp) adapters were removed with Trimmomatic v0.27 [78], while the mate-pair reads (2 × 100 bp) internal adapters were trimmed with the NxTrim [79]. Sequences with phred value below 24 (<Q24) and spanning less than 50 bp were removed. For transcriptome sequencing, we used samples from leaves, stolon, and utricles. The RNA extraction and sequencing was performed according with the *U. reniformis* previous research [26,27]. Trimmed RNAseq reads from the leaves, stolon, and utricles were concatenated, and de novo assembled using Trinity v2.7.0 [80], and were exclusively used for gene model predictions (see below).

4.3. Genome Assembly and WGD Analysis

The trimmed reads were filtered against the cpDNA (NC_029719) and mtDNA (NC_034982) organellar genomes [26,27] with bowtie2 v2.26 [81]. The assembly was carried out with MaSuRCA v3.1.3 [82], SSPACE [83], and GapCloser [84]. REAPR v1.0.18 [85] was used to generate the corrected assembly and estimate the percentage of error-free bases (additional details are provided in Supplementary Materials). CEGMA v2.5 [86] and BUSCO v3 [87] using Eudicotyledons_odb10 from the OrthoDB database [88] were used to measure the genome completeness. We used GenomeScope v1 [89], the K-mer Analysis Toolkit (KAT) [90], and *KmerSpectrumPlot.pl* from ALLPATHS-LG [91] to estimate genome size and heterozygosity.

The predicted WGDs events that occurred in the *U. reniformis* evolutionary history were predicted on the basis of the distribution of synonymous substitutions per synonymous site (Ks) of gene pairs, using the SynMap2 tool from the Comparative Genomics Platform (CoGe) [92]. MCScanX [93] was employed to identify dispersed, proximal, and tandem (small-scale duplication) and segmental duplicates (WGDs). Tandem duplicates are defined as paralogs that are adjacent to each other, and the maximum distance to call a proximal copy was set to 10 genes. The anchor genes in collinear blocks inferred segmental duplicates. Structural and comparative analyses (pairwise alignment) were carried out with D-GENIES [94]. Macrosynteny and microsynteny analyses were carried out with VGSC2 and MCscan (Python version) (https://github.com/tanghaibao/jcvi.wiki.git).

4.4. Transposable Elements Identification, Classification, and Genome Annotation

The de novo detection and classification of TEs were carried out by the REPET v2.5 [95] with the "—struct" parameter. The identified elements from the REPET package were manually validated and characterized into super-families and evolutionary lineages with the usage of the Domain-based ANnotation of Transposable Elements (DANTE) with the Viridiplantae database v3.0 [96]. The TE masking was performed with the RepeatMasker Open-v4.0.7 (http://www.repeatmasker.org) using the parameter: "-s -cutoff 260".

The gene prediction was performed using the BRAKER v2 pipeline [97], and the TE-related genes were discarded. PASA pipeline [98] was used to produce spliced alignment assemblies based on RNA-seq data. The gene predictions and transcript alignments were combined with the EVidence Modeler software [99]. Finally, two PASA pipeline iterations were used to update the EVidence Modeler consensus predictions, adding UTR annotations and models for the alternatively spliced isoforms. Functional annotation was performed using Blast2GO v5 [100], InterProScan 5 [101], and EggNOG-mapper v1.0.3 with the EggNOG 5.0 database [102]. For the protein assignment, the UniProtKB/TrEMBL [103] viridiplantae database was used. The final predicted gene models were scanned against the DANTE and RepeatMasker predictions results, and all TE-related genes were further discarded.

For consistency in the comparative analysis, *Utricularia gibba* transcriptomes read from NCBI Sequence Read Archive (SRA): SRX2368915, SRX247091, and SRX038704 were assembled, and the *U. gibba* genome (Genbank Accession Number: NEEC01000001-NEEC01000518) was re-annotated using the same pipeline.

4.5. Phylogenetic Analysis

The phylogenetic analyses were performed using maximum likelihood (ML), and Bayesian inference (BI) approaches. For the ML, RAxML v. 8.2.10 [104] was used with default parameters, and the clade support estimates were calculated using rapid bootstrapping of 1000 pseudoreplicates. The BI was performed with Mr. Bayes v.3.2.6 [105] using MCMCMC, with random starting trees, of about 5,000,000 generations, with sampling for each 100 trees, using two runs and four chains and the trees were considered only after a stationary position was reached, with 25% of initially elaborated trees discarded. The GTR + G + I was used as the best-of-fit model according to the AIC (Akaike Information Criterion) with the jModeltest v. 2.1.10 [106].

4.6. Comparative Analysis

OrthoVenn v2 [107] was used to determine the orthologous gene families among *Utricularia reniformis*, *U. gibba*, and other angiosperms (*Arabidopsis thaliana, Vitis vinifera,* and *Solanum lycopersicum*) retrieved from the Phytozome database [108]. Functional enrichment analyses of the GO terms and the KEGG enzymes were conducted using Fisher's exact test and was corrected for multiple testing using two methods employed by the Blast2GO (GO terms and KEGG enzymes) and GOATOOLS (GO terms) [109]. The *p*-values were determined after the correction for multiple tests through False Discovery Rate (FDR) control, according to the Benjamini–Hochberg. The WOX, HD-Zip, MADS-Box Transcription Factors (TFs), and ABC transporters comparative analyses were based on previous [16,33] and manual annotation using the *A. thaliana* reviewed entries from UniProt Database as reference [110].

4.7. Availability of Supporting Data

The raw sequencing reads have been deposited into GenBank, BioProject PRJNA290588, linked directly to the SRA under the accession numbers SRR8289569, SRR8289570, and SRR8289571 for the Illumina datasets and SRR8289572 for Ion Torrent's RNA-Seq data. This Whole Genome Shotgun project has been stored at DDBJ/ENA/GenBank under the accession RWGZ00000000. The version described in this paper is version RWGZ01000000. The gene models and genome browser of *Utricularia reniformis* are available at https://genomevolution.org/coge/GenomeInfo.pl?gid=54799. The Gene Ontology annotation (Blast2GO) and all the raw data generated for this study are freely available at the Zenodo platform at http://doi.org/10.5281/zenodo.3268745.

5. Conclusions

In this work, we generated a draft genome sequence and annotation for the terrestrial carnivorous plant *U. reniformis* and compared it against its sister, aquatic species *U. gibba.* Our results demonstrated that a massive proliferation and loss of lineage-specific LTR-retrotransposons, predominantly from the *Gypsy* super-family and WGDs are the main governing agents of the genome size changes and evolution between these species. Interestingly, our results strongly suggest that both genomes responded distinctly after the WGD events. This is mainly observed by specific patterns of gene fractionation and retention after the lineage split, which might be reflected in an ongoing gene birth–death–innovation process among the duplicated genes. Additionally, each species carries a diversified set of plant ontogeny and carnivory-associated gene repertoire, supporting that the reproductive isolation and terrestrial and aquatic life-form constraints (e.g., different nutrients repertoire and availability, prey diversity, and trap microbiome ecosystem) in an evolutionary time scale are associated with speciation.

Supplementary Materials: Supplementary Materials can be found at http://www.mdpi.com/1422-0067/21/1/3/s1.

Author Contributions: Conceptualization, S.R.S., A.M.V., and V.F.O.M.; methodology, A.M.V., T.P.M., V.F.O.M., S.R.S., and A.P.M.; software, A.M.V., T.P.M., S.R.S., and V.F.O.M.; validation, A.M.V., T.P.M., S.R.S., and V.F.O.M.; formal analysis, A.M.V., S.R.S., A.P.M., H.A.P., M.H.M.J., D.S.D., T.P.M., and V.F.O.M.; investigation, S.R.S., T.P.M., V.F.O.M., and A.M.V.; resources, A.P.M., V.F.O.M., and A.M.V.; data curation, A.M.V., T.P.M., and V.F.O.M.; writing—original draft preparation, A.M.V., S.R.S., and V.F.O.M.; writing—review and editing, A.M.V., S.R.S., V.F.O.M., H.A.P., A.P.M., M.H.M.J., D.S.D., and T.P.M.; visualization, S.R.S., T.P.M., V.F.O.M., and A.M.V.; supervision, A.M.V., T.P.M., and V.F.O.M.; project administration, V.F.O.M. and A.M.V.; funding acquisition, V.F.O.M. and A.M.V. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the São Paulo Research Foundation/Brazil, FAPESP [13/25164-6] to A.M.V. V.F.O.M. was funded by FAPESP [13/05144-0]. A.P.M. was funded by National Council for the Scientific and Technological Development—CNPq [460990/2014-4] and FAPESP Young Investigator Project [11/22215-3]. S.R.S. was funded by a post-doctoral research fellowship from FAPESP [18/02285-6], and H.A.P. was funded by the PNPD fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001.

Acknowledgments: We thank the Microbiology and Immunology Department of IBB-UNESP/Botucatu, Brazil that kindly made available the usage of the equipment FACSCanto II cytometer for the Genome size estimation through flow cytometry. We thank the Laboratory of Plant Systematics for all support, specially Yani Aranguren Díaz for help with the DNA extractions.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Taylor, P. *The Genus Utricularia-A Taxonomic Monograph*; The Royal Botanic Gardens, Kew: London, UK, 1989.
2. Leushkin, E.V.; Sutormin, R.A.; Nabieva, E.R.; Penin, A.A.; Kondrashov, A.S.; Logacheva, M.D. The miniature genome of a carnivorous plant *Genlisea aurea* contains a low number of genes and short non-coding sequences. *BMC Genom.* **2013**, *14*, 476. [CrossRef] [PubMed]
3. Vu, G.T.H.; Schmutzer, T.; Bull, F.; Cao, H.X.; Fuchs, J.; Tran, T.D.; Jovtchev, G.; Pistrick, K.; Stein, N.; Pecinka, A.; et al. Comparative genome analysis reveals divergent genome size evolution in a carnivorous plant genus. *Plant Genome* **2015**, *8*, 1–14. [CrossRef]
4. Ibarra-Laclette, E.; Lyons, E.; Hernández-Guzmán, G.; Pérez-Torres, C.A.; Carretero-Paulet, L.; Chang, T.H.; Lan, T.; Welch, A.J.; Juárez, M.J.A.; Simpson, J.; et al. Architecture and evolution of a minute plant genome. *Nature* **2013**, *498*, 94–98. [CrossRef] [PubMed]
5. Lan, T.; Renner, T.; Ibarra-Laclette, E.; Farr, K.M.; Chang, T.H.; Cervantes-Pérez, S.A.; Zheng, C.; Sankoff, D.; Tang, H.; Purbojati, R.W.; et al. Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E4435–E4441. [CrossRef] [PubMed]
6. Hanson, L.; Mc Mahon, K.A.; Johnson, M.A.T.; Bennett, M.D. First nuclear DNA C-values for another 25 angiosperm families. *Ann. Bot.* **2001**, *87*, 251–258. [CrossRef]
7. Greilhuber, J.; Borsch, T.; Müller, K.; Worberg, A.; Porembski, S.; Barthlott, W. Smallest angiosperm genomes found in Lentibulariaceae, with chromosomes of bacterial size. *Plant Biol.* **2006**, *8*, 770–777. [CrossRef] [PubMed]
8. Veleba, A.; Bureš, P.; Adamec, L.; Šmarda, P.; Lipnerová, I.; Horová, L. Genome size and genomic GC content evolution in the miniature genome-sized family Lentibulariaceae. *New Phytol.* **2014**, *203*, 22–28. [CrossRef]
9. Kondo, K. A Comparison of Variability in *Utricularia cornuta* and *Utricularia juncea*. *Am. J. Bot.* **1972**, *59*, 23. [CrossRef]
10. Rahman, D.M.O.; Adamec, L.; Kondo, K. Chromosome numbers of *Utricularia bremii* and *Utricularia dimorphantha* (Lentibulariaceae). *Chromosom. Sci.* **2001**, *5*, 105–108.
11. Casper, S.J.; Stimper, R. Chromosome numbers in *Pinguicula* (Lentibulariaceae): Survey, atlas, and taxonomic conclusions. *Plant Syst. Evol.* **2009**, *277*, 21–60. [CrossRef]
12. Fleischmann, A.; Michael, T.P.; Rivadavia, F.; Sousa, A.; Wang, W.; Temsch, E.M.; Greilhuber, J.; Müller, K.F.; Heubl, G. Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea*

(Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Ann. Bot.* **2014**, *114*, 1651–1663. [CrossRef] [PubMed]

13. Ibarra-Laclette, E.; Albert, V.A.; Herrera-Estrella, A.; Herrera-Estrella, L. Is GC bias in the nuclear genome of the carnivorous plant *Utricularia* driven by ros-based mutation and biased gene conversion? *Plant Signal. Behav.* **2011**, *6*, 1631–1634. [CrossRef] [PubMed]
14. Ibarra-laclette, E.; Albert, V.A.; Pérez-torres, C.A.; Zamudio-Hernández, F.; de Ortega-estrada, M.J.; Herrera-Estrella, A.; Herrera-estrella, L. Transcriptomics and molecular evolutionary rate analysis of the bladderwort (*Utricularia*), a carnivorous plant with a minimal genome. *BMC Plant Biol.* **2011**, *11*, 101. [CrossRef] [PubMed]
15. Albert, V.A.; Jobson, R.W.; Michael, T.P.; Taylor, D.J. The carnivorous bladderwort (*Utricularia*, Lentibulariaceae): A system inflates. *J. Exp. Bot.* **2010**, *61*, 5–9. [CrossRef]
16. Carretero-Paulet, L.; Chang, T.H.; Librado, P.; Ibarra-Laclette, E.; Herrera-Estrella, L.; Rozas, J.; Albert, V.A. Genome-wide analysis of adaptive molecular evolution in the carnivorous plant *Utricularia gibba*. *Genome Biol. Evol.* **2015**, *7*, 444–456. [CrossRef]
17. Grover, C.E.; Wendel, J.F. Recent Insights into Mechanisms of Genome Size Change in Plants. *J. Bot.* **2010**, *2010*, 1–8. [CrossRef]
18. Hakes, L.; Pinney, J.W.; Lovell, S.C.; Oliver, S.G.; Robertson, D.L. All duplicates are not equal: The difference between small-scale and genome duplication. *Genome Biol.* **2007**, *8*, R209. [CrossRef]
19. Pellicer, J. Genome Size Diversity and Its Impact on the Evolution of Land Plants. *Genes* **2018**, *9*, 88. [CrossRef]
20. Schatz, M.C.; Witkowski, J.; McCombie, W.R. Current challenges in de novo plant genome sequencing and assembly. *Genome Biol.* **2012**, *13*, 243. [CrossRef]
21. Moriyama, Y.; Koshiba-Takeuchi, K. Significance of whole-genome duplications on the emergence of evolutionary novelties. *Brief. Funct. Genom.* **2018**, *17*, 329–338. [CrossRef]
22. Bennetzen, J.L. Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* **2000**, *42*, 251–269. [CrossRef] [PubMed]
23. Magallón, S.; Sánchez-Reyes, L.L.; Gómez-Acevedo, S.L. Thirty clues to the exceptional diversification of flowering plants. *Ann. Bot.* **2018**, *123*, 491–503. [CrossRef] [PubMed]
24. BFG-The Brazil Flora Group Growing knowledge: An overview of Seed Plant diversity in Brazil. *Rodriguésia* **2015**, *66*, 1085–1113. [CrossRef]
25. Clivati, D.; Gitzendanner, M.A.; Hilsdorf, A.W.S.; Araújo, W.L.; de Miranda, V.F.O. Microsatellite markers developed for *Utricularia reniformis* (Lentibulariaceae). *Am. J. Bot.* **2012**, *99*, 375–378. [CrossRef]
26. Silva, S.R.; Alvarenga, D.O.; Aranguren, Y.; Penha, H.A.; Fernandes, C.C.; Pinheiro, D.G.; Oliveira, M.T.; Michael, T.P.; Miranda, V.F.O.; Varani, A.M. The mitochondrial genome of the terrestrial carnivorous plant *Utricularia reniformis* (Lentibulariaceae): Structure, comparative analysis and evolutionary landmarks. *PLoS ONE* **2017**, *12*, e0180484. [CrossRef]
27. Silva, S.R.; Diaz, Y.C.A.; Penha, H.A.; Pinheiro, D.G.; Fernandes, C.C.; Miranda, V.F.O.; Michael, T.P.; Varani, A.M. The chloroplast genome of *Utricularia reniformis* sheds light on the evolution of the ndh gene complex of terrestrial carnivorous plants from the Lentibulariaceae family. *PLoS ONE* **2016**, *11*, e0165176. [CrossRef]
28. Cheng, F.; Wu, J.; Cai, X.; Liang, J.; Freeling, M.; Wang, X. Gene retention, fractionation and subgenome differences in polyploid plants. *Nat. Plants* **2018**, *4*, 258. [CrossRef]
29. Salojärvi, J.; Smolander, O.-P.; Nieminen, K.; Rajaraman, S.; Safronov, O.; Safdari, P.; Lamminmäki, A.; Immanen, J.; Lan, T.; Tanskanen, J.; et al. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nat. Genet.* **2017**, *49*, 904–912. [CrossRef]
30. Rendón-Anaya, M.; Ibarra-Laclette, E.; Méndez-Bravo, A.; Lan, T.; Zheng, C.; Carretero-Paulet, L.; Perez-Torres, C.A.; Chacón-López, A.; Hernandez-Guzmán, G.; Chang, T.-H.; et al. The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 17081–17089. [CrossRef]
31. Wheeler, G.L.; Carstens, B.C. Evaluating the adaptive evolutionary convergence of carnivorous plant taxa through functional genomics. *PeerJ* **2018**, *6*, e4322. [CrossRef]
32. Ravee, R.; Mohd Salleh, F.-. 'Imadi; Goh, H.H. Discovery of digestive enzymes in carnivorous plants with focus on proteases. *PeerJ* **2018**, *6*, e26940. [CrossRef] [PubMed]

33. Hwang, J.-U.; Song, W.-Y.; Hong, D.; Ko, D.; Yamaoka, Y.; Jang, S.; Yim, S.; Lee, E.; Khare, D.; Kim, K.; et al. Plant ABC Transporters enable many unique aspects of a terrestrial plant's lifestyle. *Mol. Plant* **2016**, *9*, 338–355. [CrossRef] [PubMed]
34. Jiao, Y.; Leebens-Mack, J.H.; Ayyampalayam, S.; Bowers, J.E.; McKain, M.R.; McNeal, J.; Rolf, M.; Ruzicka, D.R.; Wafula, E.K.; Wickett, N.J.; et al. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **2012**, *13*, R3. [CrossRef] [PubMed]
35. Feldman, M.; Levy, A.A. Allopolyploidy–A shaping force in the evolution of wheat genomes. *Cytogenet. Genome Res.* **2005**, *109*, 250–258. [CrossRef]
36. Alix, K.; Gérard, P.R.; Schwarzacher, T.; Heslop-Harrison, J.S.P. Polyploidy and interspecific hybridization: Partners for adaptation, speciation and evolution in plants. *Ann. Bot.* **2017**, *120*, 183–194. [CrossRef]
37. Clark, J.W.; Donoghue, P.C.J. Whole-genome duplication and plant macroevolution. *Trends Plant Sci.* **2018**, *23*, 933–945. [CrossRef]
38. Bertioli, D.J.; Jenkins, J.; Clevenger, J.; Dudchenko, O.; Gao, D.; Seijo, G.; Leal-Bertioli, S.C.M.; Ren, L.; Farmer, A.D.; Pandey, M.K.; et al. The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nat. Genet.* **2019**, *51*, 877–884. [CrossRef]
39. Panchy, N.; Lehti-Shiu, M.; Shiu, S.-H. Evolution of Gene Duplication in Plants. *Plant Physiol.* **2016**, *171*, 2294–2316. [CrossRef]
40. Fedoroff, N. Transposons and genome evolution in plants. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 7002–7007. [CrossRef]
41. Bennetzen, J.L.; Wang, H. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. *Annu. Rev. Plant Biol.* **2014**, *65*, 505–530. [CrossRef]
42. Serrato-Capuchina, A.; Matute, D.R. The role of transposable elements in speciation. *Genes* **2018**, *9*, 254. [CrossRef] [PubMed]
43. Dubin, M.J.; Mittelsten Scheid, O.; Becker, C. Transposons: A blessing curse. *Curr. Opin. Plant Biol.* **2018**, *42*, 23–29. [CrossRef] [PubMed]
44. Kim, K.M.; Caetano-Anollés, G. The proteomic complexity and rise of the primordial ancestor of diversified life. *BMC Evolut. Biol.* **2011**, *11*, 140. [CrossRef] [PubMed]
45. Willing, E.M.; Rawat, V.; Mandáková, T.; Maumus, F.; James, G.V.; Nordström, K.J.V.; Becker, C.; Warthmann, N.; Chica, C.; Szarzynska, B.; et al. Genome expansion of *Arabis alpi*. *Nat. Plants* **2015**, *1*, 1–7.
46. Macas, J.; Neumann, P. Ogre elements-A distinct group of plant Ty3/gypsy-like retrotransposons. *Gene* **2007**, *390*, 108–116. [CrossRef]
47. Neumann, P.; Navrátilová, A.; Koblí, A.; Kejnovský, E.; Eva, H.; Hobza, R.; Widmer, A.; Dole, J. Plant centromeric retrotransposons: A structural and cytogenetic perspective. *Mob. DNA* **2011**, 1–16. [CrossRef]
48. Jiang, N.; Bao, Z.; Temnykh, S.; Cheng, Z.; Jiang, J.; Wing, R.A.; McCouch, S.R.; Wessier, S.R. Dasheng: A recently amplified nonautonomous long terminal repeat element that is a major component of pericentromeric regions in rice. *Genetics* **2002**, *161*, 1293–1305.
49. Kameyama, Y.; Toyama, M.; Ohara, M. Hybrid origins and F1 dominance in the free-floating, sterile bladderwort, *Utricularia australis F. australis* (Lentibulariaceae). *Am. J. Bot.* **2005**, *92*, 469–476. [CrossRef]
50. Astuti, G.; Peruzzi, L. Are shoots of diagnostic value in Central European bladderworts (*Utricularia* L., Lentibulariaceae)? *Plant Biosyst.* **2018**, *152*, 1214–1226. [CrossRef]
51. Clivati, D.; Cordeiro, G.D.; Płachno, B.J.; de Miranda, V.F.O. Reproductive biology and pollination of *Utricularia reniformis* A.St.-Hil. (Lentibulariaceae). *Plant Biol.* **2014**, *16*, 677–682. [CrossRef]
52. Rutishauser, R.; Isler, B. Developmental genetics and morphological evolution of flowering plants, especially bladderworts (*Utricularia*): Fuzzy Arberian Morphology complements Classical Morphology. *Ann. Bot.* **2001**, *88*, 1173–1202. [CrossRef]
53. Fineran, B.A. Glandular trichomes in *Utricularia*—A review of their structure and function. *Isr. J. Bot.* **1985**, *34*, 295–330.
54. Płachno, B.J.; Adamec, L.; Kamińska, I. Relationship between trap anatomy and function in Australian carnivorous bladderworts (*Utricularia*) of the subgenus *Polypompholyx*. *Aquat. Bot.* **2015**, *120*, 290–296. [CrossRef]

55. Płachno, B.J.; Stpiczyńska, M.; Krajewski, Ł.; Świątek, P.; Adamec, L.; Miranda, V.F.O. Flower palate structure of the aquatic bladderworts *Utricularia bremii* Heer and *U. minor* L. from section *Utricularia* (Lentibulariaceae). *Protoplasma* **2017**, *254*, 2007–2015. [CrossRef]
56. Jobson, R.W.; Morris, E.C. Feeding ecology of a carnivorous bladderwort (*Utricularia uliginosa*, Lentibulariaceae). *Austral Ecol.* **2001**, *26*, 680–691. [CrossRef]
57. Reifenrath, K.; Theisen, I.; Schnitzler, J.; Porembski, S.; Barthlott, W. Trap architecture in carnivorous *Utricularia* (Lentibulariaceae). *Flora Morphol. Distrib. Funct. Ecol. Plants* **2006**, *201*, 597–605. [CrossRef]
58. Sanabria-Aranda, L.; González-Bermúdez, A.; Torres, N.N.; Guisande, C.; Manjarrés-Hernández, A.; Valoyes-Valois, V.; Díaz-Olarte, J.; Andrade-Sossa, C.; Duque, S.R. Predation by the tropical plant *Utricularia foliosa*. *Freshw. Biol.* **2006**, *51*, 1999–2008. [CrossRef]
59. Rajasekar, C.; Rajendran, A. Prey composition of *Utricularia striatula* Sm. (Lentibulariaceae): Lithophytic carnivore Southern Western Ghats, India. *Int. J. Fish. Aquat. Stud.* **2018**, *6*, 382–388.
60. Elhiti, M.; Stasolla, C. Structure and function of homodomain-leucine zipper (HD-Zip) proteins. *Plant Signal. Behav.* **2009**, *4*, 86. [CrossRef]
61. Lian, G.; Ding, Z.; Wang, Q.; Zhang, D.; Xu, J. Origins and Evolution of WUSCHEL-Related Homeobox Protein Family in Plant Kingdom. *Sci. World J.* **2014**, *2014*, 1–12. [CrossRef]
62. Messenguy, F.; Dubois, E. Role of MADS box proteins and their cofactors in combinatorial control of gene expression and cell development. *Gene* **2003**, *316*, 1–21. [CrossRef]
63. Jack, T. Molecular and genetic mechanisms of floral control. *Plant Cell* **2004**, *16*, S1–S17. [CrossRef] [PubMed]
64. Yang, J.; Ding, C.; Xu, B.; Chen, C.; Narsai, R.; Whelan, J.; Hu, Z.; Zhang, M. A Casparian strip domain-like gene, CASPL, negatively alters growth and cold tolerance. *Sci. Rep.* **2015**, *5*, 14299. [CrossRef] [PubMed]
65. Cho, M.; Cho, H.-T. The function of ABCB transporters in auxin transport. *Plant Signal. Behav.* **2013**, *8*, e22990. [CrossRef] [PubMed]
66. Teschner, J.; Lachmann, N.; Schulze, J.; Geisler, M.; Selbach, K.; Santamaria-Araujo, J.; Balk, J.; Mendel, R.R.; Bittner, F. A Novel Role for Arabidopsis Mitochondrial ABC Transporter ATM3 in Molybdenum Cofactor Biosynthesis. *Plant Cell* **2010**, *22*, 468–480. [CrossRef]
67. Stein, D.J.; Chamberlain, S.R.; Fineberg, N. An A-B-C Model of Habit Disorders: Hair-Pulling, Skin-Picking, and Other Stereotypic Conditions. *CNS Spectr.* **2006**, *11*, 824–827. [CrossRef]
68. Clay, N.K.; Adio, A.M.; Denoux, C.; Jander, G.; Ausubel, F.M. Glucosinolate metabolites required for an *Arabidopsis* innate immune response. *Science* **2009**, *323*, 95–101. [CrossRef]
69. Juniper, B.E.; Robins, R.J.; Joel, D.M. *The Carnivorous Plants*; Academic Press: London, UK; San Diego, CA, USA, 1989; ISBN 978-0-12-392170-3.
70. Brownlee, C. Carnivorous plants: Trapping, digesting and absorbing all in one. *Curr. Biol.* **2013**, *23*, R714–R716. [CrossRef]
71. Sirová, D.; Bárta, J.; Šimek, K.; Posch, T.; Pech, J.; Stone, J.; Borovec, J.; Adamec, L.; Vrba, J. Hunters or farmers? Microbiome characteristics help elucidate the diet composition in an aquatic carnivorous plant. *Microbiome* **2018**, *6*, 225. [CrossRef]
72. Mette, N.; Wilbert, N.; Barthlott, W. Food composition of Aquatic Bladderworts (*Utricularia*, Lentibulariaceae) in Various Habiats. *Beiträge Biol. Pflanz.* **2000**, *72*, 1–13.
73. Kurbatova, S.A.; Yershov, I.Y. Crustaceans and Rotifers in the Predatory Feeding of *Utricularia*. *Inland Water Biol.* **2009**, *2*, 271–275. [CrossRef]
74. Harms, S. Prey selection in three species of the carnivorous aquatic plant *Utricularia* (bladderwort). *Arch. Hydrobiol.* **1999**, *146*, 449–470. [CrossRef]
75. Ebihara, A.; Ishikawa, H.; Matsumoto, S.; Lin, S.-J.; Iwatsuki, K.; Takamiya, M.; Watano, Y.; Ito, M. Nuclear DNA, chloroplast DNA, and ploidy analysis clarified biological complexity of the *Vandenboschia radicans* complex (Hymenophyllaceae) in Japan and adjacent areas. *Am. J. Bot.* **2005**, *92*, 1535–1547. [CrossRef] [PubMed]
76. Doležel, J.; Sgorbati, S.; Lucretti, S. Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiol. Plant.* **1992**, *85*, 625–631. [CrossRef]
77. Pellicer, J.; Leitch, I.J. The application of flow cytometry for estimating genome size and ploidy level in plants. In *Molecular Plant Taxonomy. Methods in Molecular Biology (Methods and Protocols)*; Bresse, P., Ed.; Humana Press: Totowa, NJ, USA, 2014; pp. 279–307. ISBN 9781627037662.

78. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
79. O'Connell, J.; Schulz-Trieglaff, O.; Carlson, E.; Hims, M.M.; Gormley, N.A.; Cox, A.J. NxTrim: Optimized trimming of Illumina mate pair reads. *Bioinformatics* **2015**, *31*, 2035–2037. [CrossRef]
80. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [CrossRef]
81. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
82. Zimin, A.V.; Marçais, G.; Puiu, D.; Roberts, M.; Salzberg, S.L.; Yorke, J.A. The MaSuRCA genome assembler. *Bioinformatics* **2013**, *29*, 2669–2677. [CrossRef]
83. Boetzer, M.; Henkel, C.V.; Jansen, H.J.; Butler, D.; Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **2011**, *27*, 578–579. [CrossRef]
84. Luo, R.; Liu, B.; Xie, Y.; Li, Z.; Huang, W.; Yuan, J.; He, G.; Chen, Y.; Pan, Q.; Liu, Y.; et al. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* **2012**, *1*, 18. [CrossRef] [PubMed]
85. Hunt, M.; Kikuchi, T.; Sanders, M.; Newbold, C.; Berriman, M.; Otto, T.D. REAPR: A universal tool for genome assembly evaluation. *Genome Biol.* **2013**, *14*, R47. [CrossRef] [PubMed]
86. Parra, G.; Bradnam, K.; Korf, I. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **2007**, *23*, 1061–1067. [CrossRef] [PubMed]
87. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef]
88. Kriventseva, E.V.; Kuznetsov, D.; Tegenfeldt, F.; Manni, M.; Dias, R.; Simão, F.A.; Zdobnov, E.M. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **2019**, *47*, D807–D811. [CrossRef]
89. Vurture, G.W.; Sedlazeck, F.J.; Nattestad, M.; Underwood, C.J.; Fang, H.; Gurtowski, J.; Schatz, M.C. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* **2017**, *33*, 2202–2204. [CrossRef]
90. Mapleson, D.; Garcia Accinelli, G.; Kettleborough, G.; Wright, J.; Clavijo, B.J. KAT: A K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **2017**, *33*, 574–576. [CrossRef]
91. Gnerre, S.; MacCallum, I.; Przybylski, D.; Ribeiro, F.J.; Burton, J.N.; Walker, B.J.; Sharpe, T.; Hall, G.; Shea, T.P.; Sykes, S.; et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 1513–1518. [CrossRef]
92. Haug-Baltzell, A.; Stephens, S.A.; Davey, S.; Scheidegger, C.E.; Lyons, E. SynMap2 and SynMap3D: Web-based whole-genome synteny browsers. *Bioinformatics* **2017**, *33*, 2197–2198. [CrossRef]
93. Wang, Y.; Tang, H.; Debarry, J.D.; Tan, X.; Li, J.; Wang, X.; Lee, T.; Jin, H.; Marler, B.; Guo, H.; et al. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **2012**, *40*, e49. [CrossRef]
94. Cabanettes, F.; Klopp, C. D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **2018**, *6*, e4958. [CrossRef] [PubMed]
95. Flutre, T.; Duprat, E.; Feuillet, C.; Quesneville, H. Considering Transposable Element Diversification in De Novo Annotation Approaches. *PLoS ONE* **2011**, *6*, e16526. [CrossRef] [PubMed]
96. Neumann, P.; Novák, P.; Ho, N. Systematic survey of plant LTR- retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob. DNA* **2019**, *10*, 1–17. [CrossRef] [PubMed]
97. Hoff, K.; Lomsadze, A.; Borodovsky, M.; Stanke, M. Whole-Genome Annotation with BRAKER. In *Gene Prediction. Methods in Molecular Biology*; Kollmar, M., Ed.; Humana: New York, NY, USA, 2019; pp. 65–95.
98. Haas, B.J.; Delcher, A.L.; Mount, S.M.; Wortman, J.R.; Smith, R.K.S., Jr.; Hannick, L.I.; Maiti, R.; Ronning, C.M.; Rusch, D.B.; Town, C.D.; et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **2003**, *31*, 5654–5666. [CrossRef] [PubMed]

99. Haas, B.J.; Salzberg, S.L.; Zhu, W.; Pertea, M.; Allen, J.E.; Orvis, J.; White, O.; Buell, C.R.; Wortman, J.R. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **2008**, *9*, R7. [CrossRef] [PubMed]
100. Conesa, A.; Götz, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genom.* **2008**, *2008*, 619832. [CrossRef]
101. Jones, P.; Binns, D.; Chang, H.-Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **2014**, *30*, 1236–1240. [CrossRef]
102. Huerta-Cepas, J.; Forslund, K.; Coelho, L.P.; Szklarczyk, D.; Jensen, L.J.; von Mering, C.; Bork, P. Fast genome-wide functional annotation through orthology massignment by eggNOG-Mapper. *Mol. Biol. Evol.* **2017**, *34*, 2115–2122. [CrossRef]
103. Pundir, S.; Magrane, M.; Martin, M.J.; O'Donovan, C. Searching and navigating UniProt databases. *Curr. Protoc. Bioinform.* **2015**, *50*, 1–27. [CrossRef]
104. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef]
105. Ronquist, F.; Teslenko, M.; Van Der Mark, P.; Ayres, D.L.; Darling, A.; Höhna, S.; Larget, B.; Liu, L.; Suchard, M.A.; Huelsenbeck, J.P. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **2012**, *61*, 539–542. [CrossRef] [PubMed]
106. Darriba, D.; Taboada, G.L.; Doallo, R.; Posada, D. JModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* **2012**, *9*, 772. [CrossRef] [PubMed]
107. Xu, L.; Dong, Z.; Fang, L.; Luo, Y.; Wei, Z.; Guo, H.; Zhang, G.; Gu, Y.Q.; Coleman-Derr, D.; Xia, Q.; et al. OrthoVenn2: A web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* **2019**, *47*, W52–W58. [CrossRef] [PubMed]
108. Goodstein, D.M.; Shu, S.; Howson, R.; Neupane, R.; Hayes, R.D.; Fazo, J.; Mitros, T.; Dirks, W.; Hellsten, U.; Putnam, N.; et al. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.* **2012**, *40*, 1178–1186. [CrossRef]
109. Klopfenstein, D.V.; Zhang, L.; Pedersen, B.S.; Ramírez, F.; Warwick Vesztrocy, A.; Naldi, A.; Mungall, C.J.; Yunes, J.M.; Botvinnik, O.; Weigel, M.; et al. GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* **2018**, *8*, 10872. [CrossRef]
110. Consortium, T.U. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2018**, *47*, D506–D515. [CrossRef]

International Journal of
Molecular Sciences

MDPI

Article

Combined De Novo Transcriptome and Metabolome Analysis of Common Bean Response to *Fusarium oxysporum* f. sp. *phaseoli* Infection

Limin Chen [1], Quancong Wu [1,*], Weimin He [1], Tianjun He [1], Qianqian Wu [2] and Yeminzi Miao [1]

[1] Integrated Plant Protection Center, Lishui Institute of Agricultural and Forestry Sciences, 827 Liyang Stress, Lishui 323000, China
[2] School of Agricultural and Food Science, Zhejiang A&F University, Hangzhou 311300, China
* Correspondence: lsqcw@163.com; Tel.: +86-578-2028375; Fax: +86-578-2173070

Received: 24 November 2019; Accepted: 11 December 2019; Published: 12 December 2019

Abstract: Molecular changes elicited by common bean (*Phaseolus vulgaris* L.) in response to *Fusarium oxysproum* f. sp. *Phaseoli* (FOP) remain elusive. We studied the changes in root metabolism during common bean–FOP interactions using a combined de novo transcriptome and metabolome approach. Our results demonstrated alterations of transcript levels and metabolite concentrations in common bean roots 24 h post infection as compared to control. The transcriptome and metabolome responses in common bean roots revealed significant changes in structural defense i.e., cell-wall loosening and weakening characterized by hyper accumulation of cell-wall loosening and degradation related transcripts. The levels of pathogenesis related genes were significantly higher upon FOP inoculation. Interestingly, we found the involvement of glycosylphosphatidylinositol- anchored proteins (GPI-APs) in signal transduction in response to FOP infection. Our results confirmed that hormones have strong role in signaling pathways i.e., salicylic acid, jasmonate, and ethylene pathways. FOP induced energy metabolism and nitrogen mobilization in infected common bean roots as compared to control. Importantly, the flavonoid biosynthesis pathway was the most significantly enriched pathway in response to FOP infection as revealed by the combined transcriptome and metabolome analysis. Overall, the observed modulations in the transcriptome and metabolome flux as outcome of several orchestrated molecular events are determinant of host's role in common bean–FOP interactions.

Keywords: common bean; *Fusarium oxysproum*; plant–pathogen interaction; transcriptome; metabolome

1. Introduction

Fusarium wilt, caused by *Fusarium oxysporum* f. sp. *phaseoli* (FOP), is a destructive soil-borne common bean (*Phaseolus vulgaris* L.) disease. Since its first identification in USA in 1929, this disease has been detected in all bean growing areas such as Africa, East Asia, Europe, Latin America, United Sates, and China [1,2]. Sever fusarium wilt epidemics have been reported in Heilongjiang and other common bean growing areas of China where beans follow vegetables [3]. High moisture, excessive irrigation, or poorly drained fields and a lack of rotation encourage the disease and FOP can persist in the soil indeterminately because of the production of chlamydospores and to the colonization of plant residues including roots of non-susceptible crops cultivated in rotation [3,4]. The majority of studies showed that invasion begins with the hyphal network development around root hairs followed by penetration and colonization of the epidermis and subsequently into the vascular tissues of the root. The colonization in vascular tissues leads it to the stem or the whole plant causing phloem blockage, internal stem discoloration, and total plant wilt. Infected plants display stunting, complete wilting, extensive chlorosis, and necrosis on the leaves [4,5].

Host-*F. oxysporum* pathosystems have been characterized in limited crops i.e., banana (*Musa paradisiaca*) [6,7], melon (*Cucumis melo*) [8,9], chickpea (*Cicer arietinum*) [10–12], cotton (*Gossypium hirsutum*) [13], tomato (*Lycopersicon esculentum*) [14], *Arabidopsis* [15,16], *Medicago truncatula* [17]. For each plant species, the respective *Fusarium* pathogens and a variety of defense mechanisms have been observed, including wound responses and hypersensitive reactions as well as many gene expression and metabolic changes. After the infection, plants recognize *Fusarium oxysporum* (FO) attack by understanding endogenous signals originating from the cell-wall through surveillance of cellular intactness. In this regard, many genes such as subtilin-like proteases, leucine rich-repeat proteins, proline-rice glycoproteins, cellulose synthases, and syntaxins are regulated as revealed in melon and banana [6,9,18]. In addition, constitutive enzymatic responses to FO infection appear to be important with changes in glutathione S-transferases, peroxidases, and phenylalanine ammonia lyase enzyme levels and activities being significant upon pathogen attack [19]. Changes also occur in the types and levels of cell-wall proteins, proteinase inhibitors, hydrolytic enzymes, and pathogenesis-related (PR) proteins and phytoalexin biosynthetic enzymes also appear to play important roles in FO defense [19–21]. Upregulation of genes involved in shikimate phenylpropanoid-lignin and cellulose synthesis pathways is possibly the reason of resistance in many cultivars where reduced spores are attached and resistance to FO is enhanced [18]. Identification of microbial surface derived molecules i.e., pathogen/microbe-associated molecular patterns (PAMPs/MAMPs) via pattern recognition receptors (PRRs) followed by binding of PAMPs to specific PRRs activates them and sends downstream signals to trigger broad spectrum immunity [22]. In terms of chemical defense, several genes inducing chitinases, xylem proteinases, β-1,3-glucanases thaumatin-like proteins, and peroxidases have been reported to be induced in melon against FOM infection [9]. Hypersensitive response signal molecules such as salicylic acid (SA), jasmonic acid (JA), plant growth regulating hormones, antioxidants, defense metabolites (polyphenols, phenolic acids, and flavonoids), and certain organic acids have been reported to be induced as an active plant defense [9,23]. Certain changes at the metabolic level such as altered activity of genes involved in sugar metabolism (sucrose synthase, invertase, and β-amylase) are also considered a plant response to FO. Sugars offer a dual function in plants as a nutrient as well as a signal to onset of disease, hence, these reactions are important when considering plant defense response [10,24]. The redox status of the intracellular (symplastic) and extracellular (apoplastic) spaces also change with Fusarium wilt infection [25].

In common bean, only a limited number of studies have been conducted to understand the mechanism and pathways involved in response to FOP infection. A recent study using cDNA amplified fragment length polymorphism technique reported transcript-derived fragments functionally characterized as metabolism, signal transduction, protein synthesis and processing, development and cytoskeletal organization, redox reaction, defense and stress response, transport proteins, and gene expression and RNA metabolism related genes/proteins [3]. However proteomic and metabolomics scale responses of common bean against FOP infection is poorly understood.

Unbiased modern high-throughput technologies such as combination of metabolomics and transcriptomics are required to improve our understanding of the plant–fungus interactions in common bean. Such combined approaches have recently resulted in the elucidation of different pathways in plants such as reprogramming of metabolites in chickpea roots in response to FO [12], understanding system responses to brown planthopper and rice stem borer infestation in rice [26,27]. Considering the amount of work done in model plants and other plant species infected with FO, it is important to understand the metabolic changes, transcriptional regulation, or physiological responses of bioactive and signaling compounds during infection of common bean with FOP. We aimed at identifying the differentially expressed genes (DEGs) and metabolites in common bean in response to FOP. The results showed that the transcriptome and metabolome response included structural defense, pathogen recognition receptors, and other components of innate immune system. Hormones played a crucial role in signaling pathways in common bean-FOP pathosystem. The most significantly enriched pathway as per metabolome results and confirmed by transcriptome analysis was the flavonoid biosynthesis pathway. We observed a highly orchestrated response with significant modulation in various metabolic processes. The results described here thus improve our fundamental knowledge of molecular responses to the common bean–FOP interaction and potentially useful in designing strategies against wilt disease in common bean.

2. Results

Fusarium wilt progression in common bean (Liyun No. 2) infected with FOP (FO; inoculated) and the control (CK; non-inoculated) was monitored by phenotypic screening at 4, 8, 12, 18, and 24 h post inoculation (hpi). The fusarium infection symptoms i.e., disease incidence index (severity rate 1–5), root length, fresh weight, and root volume were recorded for each time point for nine plants. After 12 hpi, the symptoms started to appear, and disease could be confirmed to a scale of 3. The disease incidence continued to affect roots and shoots. At 24 hpi, the disease incidence reached to a scale of 5 for all nine plants confirming the establishment of the disease. At this time point we confirmed that all treated seedlings at 24 hpi showed significant changes in recorded characteristics (Table S1). The control seedlings in contrast showed normal growth as compared to infected seedlings, remained healthy, and showed no fusarium wilt confirming the successful inoculation of treated plants (Figure 1).

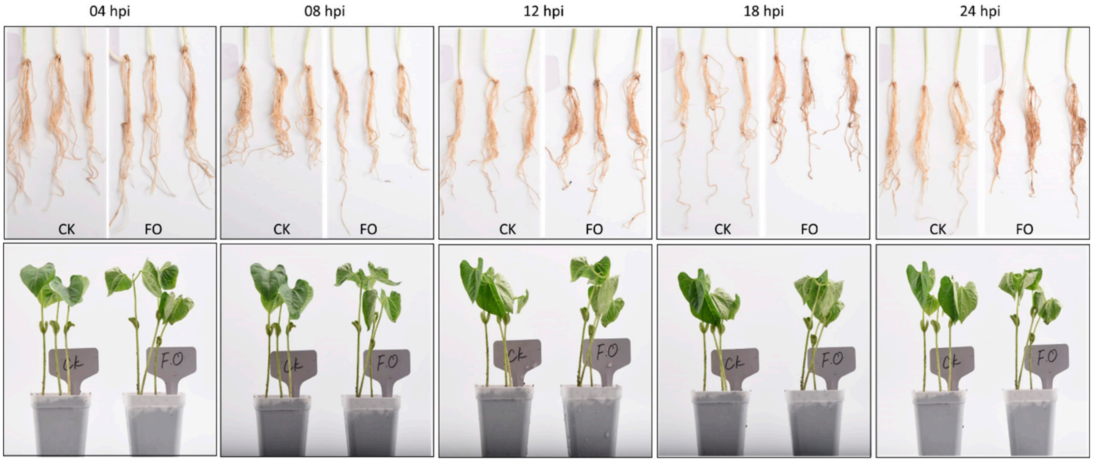


Figure 1. *Fusarium oxysproum* f. sp. *phaseoli* infected (FO) and non-infected (CK) common bean roots and seedlings at 4, 8, 12, 18, and 24 h post infection.

2.1. RNA Sequencing and Identification of Differentially Expressed Genes

The transcriptome of six samples of infected and non-infected common bean seedlings were sequenced using the Illumina HiSeq High-throughput sequencing platform. Illumina reads ranging from 43.65 to 48.71 million/sample (on average 46.25 million reads) were obtained from the six samples (Table S2). After filtering low quality reads and adapter sequences, a total of 39.57 Gb clean data was obtained. The clean data of each sample reached 5 Gb, and the Q30 base percentage was 93% or more. The clean data of the six samples of infected and non-infected common bean seedlings were de novo assembled as the reference gene set using the Trinity software and 136,238 unigenes, comprising 195,895,876 bp, were obtained, with a mean length of 1438 bp, a N50 of 2150 bp, and a N90 of 682 bp.

Functional annotation of all the unigenes was conducted, and a total of 80,409 (59.02%), 105,731 (77.61%), 71,638 (52.58%), 7,102,724 (75.4), 61,960 (45.48), 88,302 (64.81%), and 82,599 (60.63%) unigenes were annotated to the Kyoto encyclopedia of genes and genomes (KEGG), non-redundant (NR), Swiss-Prot, Trembl, euKaryotic Ortholog Groups (KOG), Gene ontology (GO), and Pfam database, respectively. Out of all unigenes, 106,777 (78.38%) were annotated in at least one database (Figure S1A). The functional information of homologous sequences in related species showed that the transcript sequences had 68.71%, 9.88%, 2.96%, 2.39%, 2.01%, 1.43%, 1.13%, 0.89%, and 10.59% similarity with *P. vulgaris*, *Quercus suber*, *Vigna angularis*, *Vigna radiate* var. *radiate*, *Cajanus cajan*, *Glycine max*, *Ricinus* *communis*, *Vigna angularis* var. *angularis*, and other genomes, respectively (Figure S1B). The GO annotation indicated 88,302 unigenes were categorized into 60 functional terms in three categories. Among them, genes associated with metabolic and cellular process in the category 'biological process'; cell, organelle, and cell part in the category 'cellular components'; and catalytic activity and transport activity in the category 'molecular function', were the most abundant (Figure S1C). The KEGG pathway database was used to analyze intracellular metabolic processes, and 80,409 unigenes were assigned to 144 KEGG pathways (Figure S1D).

For the FO-24 samples 84.51–87.26% and for the CK-24 samples 89.17–89.82% reads could be mapped to reference gene set (Table S3). Overall the Fragments Per Kilobase of Transcript per Million fragments mapped (FPKM) for FO-24 seedlings was higher for all three samples as compared to control (Figure 2a). Pearson correlations between FOP inoculated replicates ranged from 0.78 to 0.94 (Figure 2b). The CK replicates clustered together while the FOP inoculated replicates showed variability suggesting that inoculation within the treated plants differed (Figure 2c).

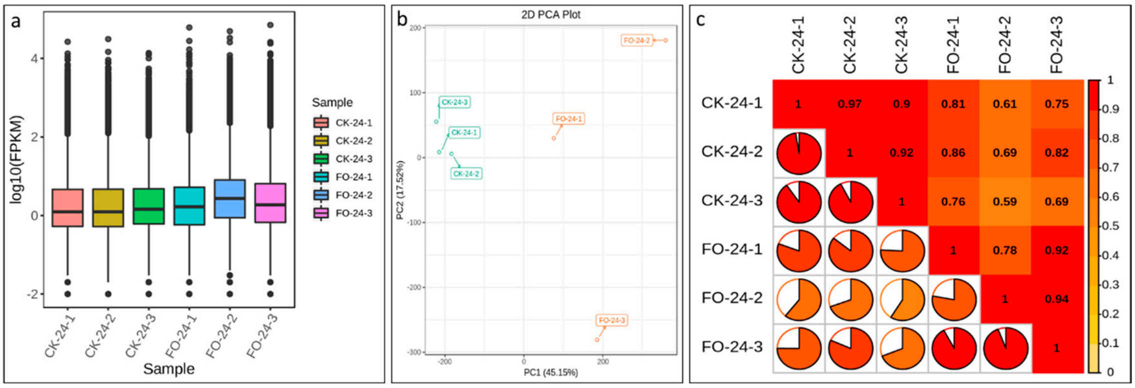


Figure 2. (**a**) Overall distribution of sample gene expression, (**b**) principle component analysis of expressed genes, and (**c**) Pearson correlations between CK-24 and FO-24 replicates.

2.2. *Differential Gene Expression Analysis Related to FOP Infection*

The screening conditions for the differentially expressed genes (DEG) were |log 2 Fold Change| ≥ 1, and False discovery rate (FDR) < 0.05. A total of 22,040 unigenes were expressed differentially with 8269 downregulated and 13,771 upregulated (Figure 3a).

Of the DEGs, 2780 downregulated genes were exclusively expressed in CK-24 while, 8231 upregulated DEGs were exclusively expressed in FO-24. These results show that the transcriptional changes are intense in FO infected common bean seedlings at 24 hpi. We further performed KEGG analysis to look at the key biological pathways involved in response to infection of FOP at 24 hpi. Spliceosome was the significantly enriched pathway with highest ratio of the number of differential genes annotated to this pathway to the number of annotated differential genes i.e., 314 out of 7204 genes in response to FOP infection 24 hpi. Other pathways such as endocytosis, RNA transport, mitogen-activated protein kinase (MAPK) signaling pathway-plant, mRNA surveillance pathway, amino sugar and nucleotide sugar metabolism pathway, and peroxisomes were significantly enriched (Figure 3b).

KEGG database (http://www.genome.jp/kegg/) was used to perform pathway mapping of the DEGs involved in common bean–FO interactions to facilitate the inspection of the plant gene networks. KEGG analysis revealed that unigenes were significantly enriched in various components involved in pathogen resistance mechanisms or signaling pathways (Figure 4).

To identify the most potential candidate genes related to resistance mechanism in common bean against FOP, we focused subsequent analysis on the 11,950 DEGs with fold change > 2 (Table S4).

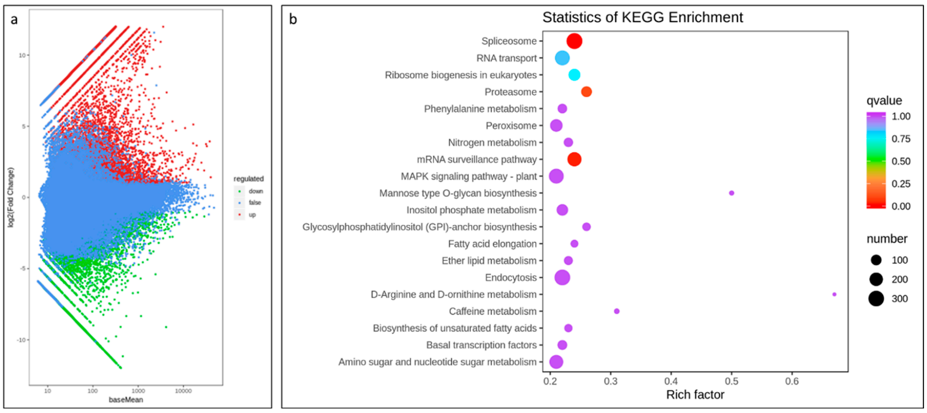


Figure 3. (**a**) Differential gene MA map. The ordinate represents the log2 fold change value; the abscissa represents the average value of gene expression in the two samples; the red dot represents the upregulation of the gene expression, and the green dot represents the downregulation of the expression. Blue indicates no significant difference in gene expression. (**b**) Kyoto encyclopedia of genes and genomes (KEGG) enrichment scatter plot. The ordinate represents the KEGG pathway. The abscissa represents the Rich factor. The larger the Rich factor, the greater the enrichment. The larger the point, the greater the number of differential genes enriched in the pathway. The redder the color of the dots, the more significant the enrichment.

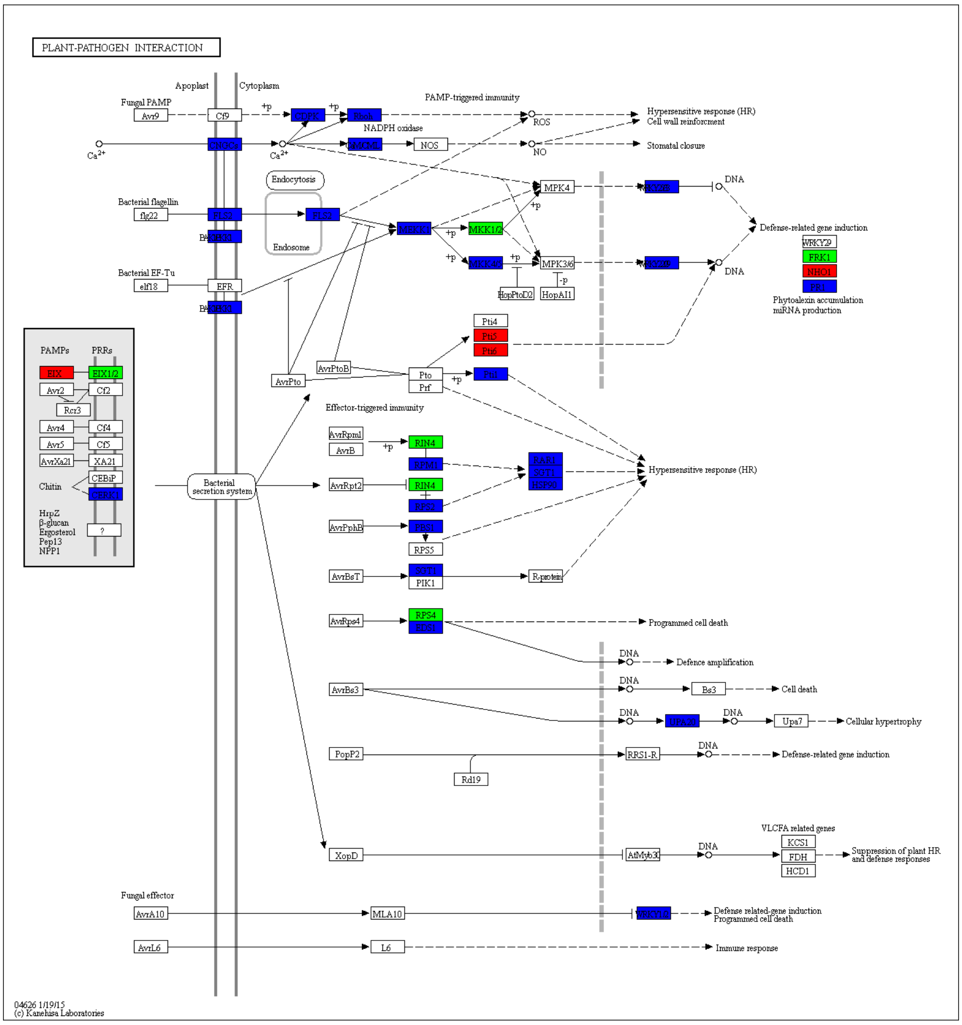


Figure 4. KEGG orthology map (ko04626, plant–pathogen interaction) of common bean-*Fusarium oxysproum*

f. sp. *Phaseoli* (FOP) pathosystem. For the treatment group, the red box labeled enzyme is associated with the upregulated gene, and the green box labeled enzyme is associated with the downregulated gene. The blue labeled enzyme is related to both upregulation and downregulation. This pathway map is associated with DEGs. The enzymes are all marked with different colors.

2.2.1. Structural Defense

Plants recognize FOP attack for effective defense. Plants perceive an arsenal of endogenous signals originating from their own cell-wall by surveillance of cellular interactions. The plant cell-wall is the first line of defense for plant cells and defines the primary strength of plants to restrict entry of pathogen to cell. We found 21 subtilisin-like protease and 46 DEGs encoding for leucine rich-repeat (LRR) proteins were of the highly expressed proteins in common bean seedlings infected by FOP 24 hpi. Several genes involved in shikimate phenylpropanoid-lignin and cellulose biosynthesis pathways are reported to strengthen the cell-wall in resistant plants in response to FO infection [18]. We observed higher expression of 3-deoxy-d-arabino-heptulosonate-7-phosphate synthase, a lower expression of coumarate-CoA ligase gene which is reported to strengthen the cell-wall, indicating cell-wall weakening in response to FOP infection 24 hpi. A polyphenol oxidase gene had a very high expression in FOP infected plants as compared to the CK. Similarly, other genes such as UDP-glucuronic acid decarboxylase and cellulose synthases were downregulated in FOP infected plants. Other genes for cell-wall reinforcement have been reported to express during FO infection in tomato [9], we also noticed a shift in expression of four proline-rich glycoprotein, four hydroxyproline-rich glycoprotein, and 18 syntaxin genes. In addition, the bean response to FOP infection 24 hpi was characterized by hyper accumulation of transcripts coding for cell-wall degradation i.e., pectate lyases, pectin methylesterase inhibitors (PMEI), pectin methylesterases (PME), and Polygalacturonases (PG) (Table S4).

2.2.2. Signaling

Plants recognize pathogen surface-derived molecules i.e., (PAMP/MAMP) via PRRs. This binding of PAMP to specific PRR activates these receptors and relays the signal downstream to convergent signaling pathways triggering broad-spectrum immunity. A total of 15 pathogenesis-related (PR) genes were differentially expressed. The fragments released in response to disruption of the first line of defense i.e., cell-wall (galacturonic acid-containing fragments) act as signals and mediate defense response by strengthening defensive barriers i.e., Chitin elicitor-binding protein (CEBiP) and chitin elicitor receptor kinase (CERK). In beans infected with FO-24, seven fungal elicitor immediate early-responsive genes showed higher expression as compared to CK-24. Further, 46 receptor kinases belonging to different gene families were expressed in FO-24 with a limited or zero expression in CK-24.

Involvement of glycosylphosphatidylinositol-anchored proteins (GPI-Aps) with extracellular ligands such as pathogen molecules as well as other ligands i.e., phytohormones, signaling polypeptides, leads to the phosphorylation of the intracellular kinase domain, which consequently activate cytoplasmic signaling components and switch on the response mechanisms. We also observed that glycosylphosphatidylinositol (GPI)-anchor biosynthesis pathway (K000563) was within significantly enriched pathways (Figure 3b). Apart from this, calmodulin (CaM) related DEGs were also induced suggesting the involvement of the CaM dependent signaling pathway.

The role of hormones in signaling pathways is well established which involves systematic acquired resistance. It has been suggested that resistance to FOP is mediated by SA, jasmonate (JA), and ethylene (ET) pathways [9,18]. In this regard 24 DEGs encoding for Ankyrin repeat containing protein genes were highly expressed in FO-24 beans. Genes responsive to ET are activated during early infection. We observed that four ET-insensitive protein 2 genes were highly responsive to FOP infection in FO-24 bean seedlings. This suggests that ET responsive genes might also be involved in latter infection stages of FOP and should be investigated further. Levels of core JA-signaling related genes i.e., four ZIM domain containing proteins, 12 TIFY, 26 ethylene-responsive transcription factor genes were highly expressed. Among these ethylene-responsive transcription factor-RAP2-7 and RAP2, AP2-like ethylene-responsive transcription factor, ethylene-responsive transcription factor ERF113, ethylene-responsive transcription factor 1, ERF15 were uniquely expressed in FO-24 bean seedlings. There were nine superoxide dismutases expressed in FO-24 in higher fold change values as compared to CK-24 supporting the notion that JA levels are quite high in infected common bean seedlings. Further, the elevated levels of the transcripts of lipoxygenases, linoleate 13/9 S-lipoxygenases, and allene oxide cyclases suggested higher JA levels. However, the role of JA pathway in susceptibility and tolerance FO is still controversial. Almost a five-fold increase in transcript abundance of a defensin-like protein (*PHAVU_005G071400g*) was observed. Homologs of this gene have been reported in *Arabidopsis* and are induced in response to FO infection. It is known that FO infection activates the transcription of auxin-related genes leading to a higher auxin biosynthesis. Three auxin influx career genes, nine auxin induced proteins, and one auxin responsive protein had higher expression in FO-24 seedlings as compared to CK-24. We observed the activation of the MAPK cascade. This pathway was also found to be significantly enriched (Figure 3b).

2.3. Validation of DEGs by qRT-PCR

We validated the expression profiles of eight common bean genes of particular interest (Figure 5). The *Actin* gene was used as an internal control to standardize the data, and the amount of target gene transcript was normalized compared to the constitutive abundance of common bean *Actin* gene [3,28]. Among the common bean DEGs analyzed, five genes were upregulated in FO-24 as compared to CK-24 i.e., *PHAVU_007G070400g*, *PHAVU_004G134300g*, *PHAVU_011G042100g*, *PHAVU_008G232600g*, and *PHAVU_007G185300g*. All these upregulated genes were characterized by similar trend in transcript accumulation at tested hpi supporting the RNA-Seq data. The other three genes i.e., *PHAVU_003G141800g*, *PHAVU_007G0495001g*, and *PHAVU_007G236300g* were downregulated in response to FOP infection in FO-24 seedlings confirming the reliability of our RNA-Seq data. The three downregulated genes are cell wall related, a serine/threonine kinase activity related gene, and an Ankyrin repeat containing gene, respectively.

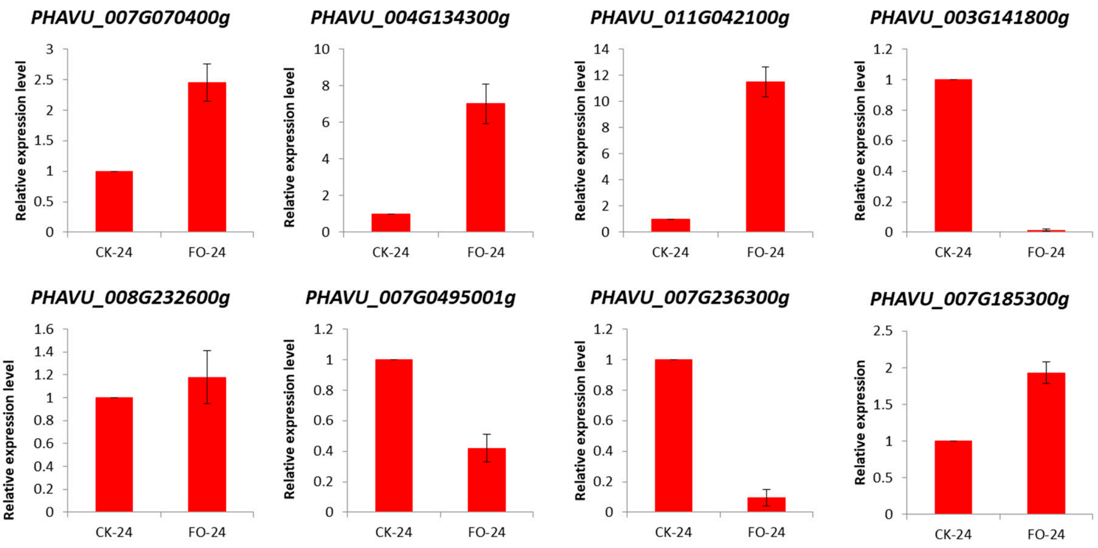


Figure 5. qRT-PCR validation of the selected common bean differentially expressed genes (DEGs) in control (CK-24) and FOP infected plants (FO-24) 24 h post infection.

2.4. Metabolite Profile

A combination of UPLC-MS/MS detection platform, self-built database, and multivariate statistical analysis was used to study the differences in metabolome between CK-24 and FO-24. It offers a platform to detect a great diversity of metabolites as previously reported in tomato [29], *Prunus mira* [30], and hulless barley [31–33]. In total, 754 metabolites were successfully detected in both sample types (Table S5).

The diverse set of detected molecules could be roughly grouped into 23 major classes, predominantly organic acids and derivatives, amino acids and their derivatives, nucleotides and their derivatives, lipids, phenylpropanoids, and flavones. Collectively, phenolics were the major components of metabolome (flavanone, flavone, flavonoid, flavonol, isoflavone, polyphenol, anthocyanins, proanthocyanidins, phenolamides, phenylpropanoids) accounting for 1/3 of the total metabolites detected (Table S5). Moreover, since most of the identified metabolites in this study have not yet been reported in the *Phaseolus vulgaris* metabolic network, our work offers prospects for new bioactive compound discovery.

We compared the quantitative metabolic profiles between CK-24 and FO-24 roots in order to identify the compounds that differentially accumulated after infection. A series of pairwise OPLS-DA were applied to maximize the discrimination between experimental samples and to focus on metabolic variations significantly contributing to the resulting classifications. The differences between the control and infected groups in the OPLS-DA suggested that significant biochemical perturbation occurred in these samples (Figure S2). All significantly differentially expressed metabolites (fold change ≥ 2 or ≤ 0.5), with the variable importance in the projection (VIP ≥ 1.0) between FO-24 and CK-24 roots are listed in Table S6 and Figure 6a. In total, 158 metabolites were differentially expressed, with 110 upregulated and 48 downregulated.

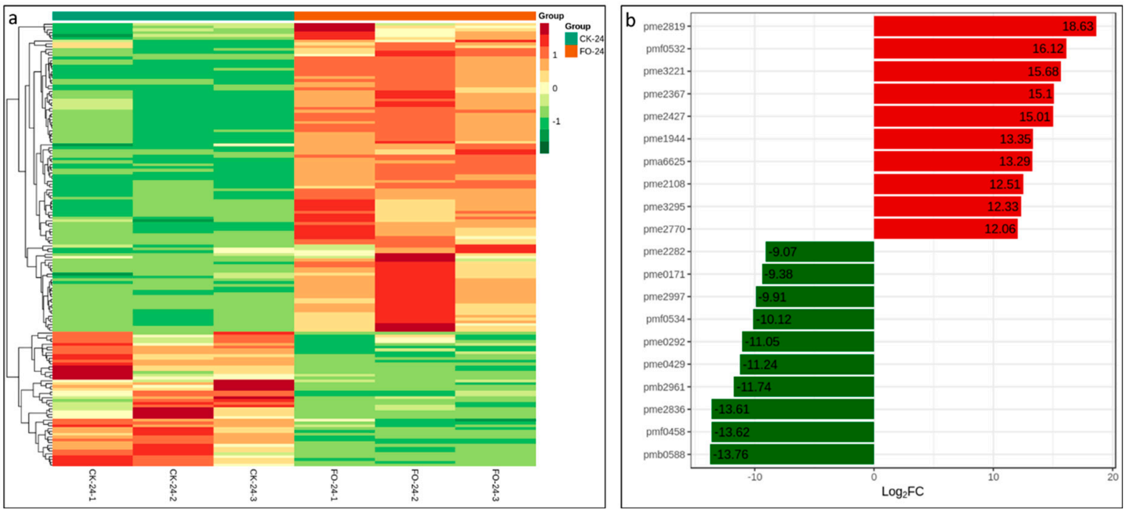


Figure 6. (**a**) Heatmap hierarchical clustering of differentially expressed metabolites. Hierarchical trees were drawn based on differentially accumulated metabolites in CK-24 and FO-24. (**b**) Top 10 differentially accumulated metabolites in CK-24 and FO-24.

We found that most of the significantly altered metabolites between FO-24 and CK-24 are phenolic compounds (Table S6). The top 10 most differentially expressed metabolites are listed in Figure 6b. Among them, the upregulated metabolites include All-trans-13,14-dihydroretinol, Phillyroside, Isoeugenol, Quinone, N-Acetyl-L-tyrosine, D-Mannitol, E-3,4,5′-Trihydroxy-3′-glucopyranosylstilbene, L-Carnitine, Prunetin, and L-Cystathionine. Similarly, the top 10 most downregulated metabolites include Luteolin 3′,7-di-O-glucoside, 6-Gingerol, 5-Hydroxytryptophol, Peonidin O-malonylhexoside, 3,4,5-Trimethoxycinnamic acid, (3,4-Dimethoxyphenyl) acetic acid, Hinokitiol, 4-Hydroxycoumarin, N-Isovaleroylglycine, and Guanosine monophosphate.

Differentially accumulated metabolites were functionally annotated using the KEGG database. It was observed that flavonoid biosynthesis, Glycerolipid metabolism, and Glycerophospholipid metabolism pathways were the most enriched (Figure 7).

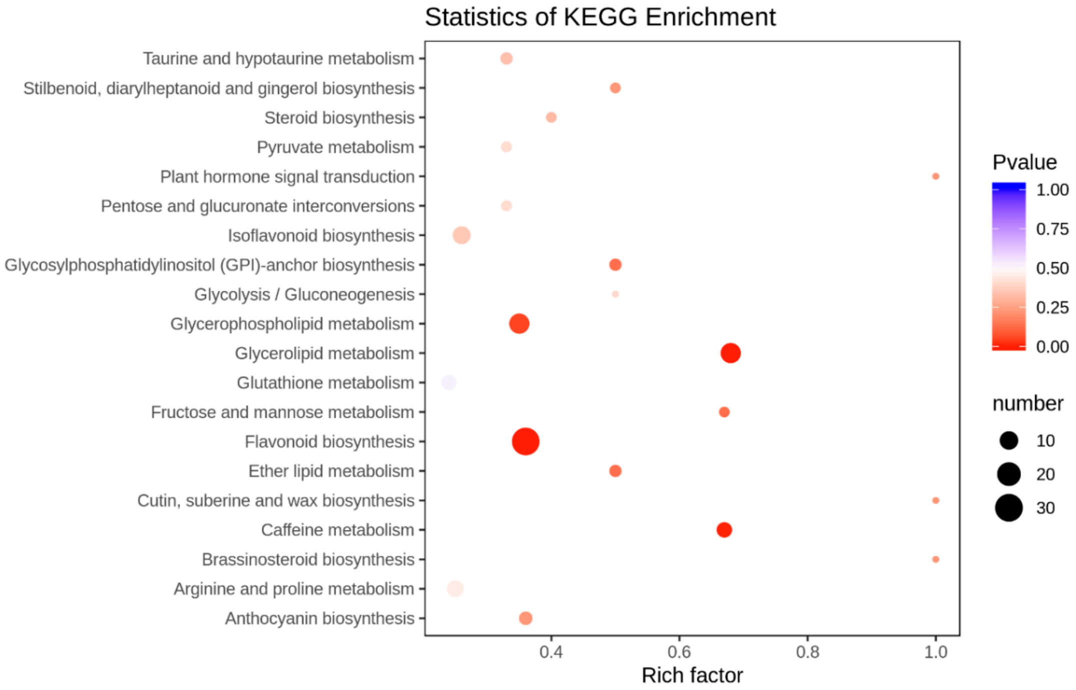


Figure 7. Differential metabolite KEGG enrichment. The larger value indicates that the degree of enrichment is greater. The closer the *p*-value is to 0, the more significant the enrichment is. The size of the points in the graph represents the number of distinct significant metabolites enriched into the corresponding pathway.

3. Discussion

3.1. Structural Defense in Response to FOP Infection in Common Bean

The present study reports the first combined de novo metabolome and RNA-seq analysis designed to describe the response of common bean infected with FOP. Previously, plant–FOP interaction has been demonstrated in many experiments in different crop plants [6–16]. In common bean it is known that colonization of FOP induces the defense responses in both a constitutive and inducible way. Pathogens are perceived by two different recognition systems that initiate pattern-triggered immunity and effector triggered immunity in order to repel pathogens via induced defense responses [22]. Common bean, like other plants, employs cell-wall as the first barrier and defines primary or basic strength to encounter FOP infection. In this regard, the lower expression of UDP-glucuronic acid decarboxylase and cellulose synthases, coumarate-CoA ligase and hyper accumulation of pectate lyases, PMEIs, PMEs, and PGs indicate that FOP has established itself at 24 hpi and cell-wall weakening in response to FOP infection has started. The higher expression of 3-deoxy-d-arabino-heptulosonate-7-phosphate synthase and polyphenol oxidase highlighted the timely recognition of FOP invasion and induction of the defense system [9,18]. These results confirmed that upon infection and establishment of FOP in common bean root tissues, the FOP secreted enzymes loosen and degrade the cell-wall i.e., pectin, cellulose, and hemicellulose [34].

3.2. Modulation of Defense Related Proteins in Common Bean

In order to trigger immunity, plants recognize the pathogen surface derived molecules PAMP/MAMP by employing PRRs which in turn activates the receptors and transcends the signals to different pathways and triggers broad spectrum immunity in plants. The PR proteins, among all defense related proteins, are induced and accumulated in response to FOP infection at 24 hpi and are considered an indispensable component of the innate immune system [35]. The 15 differentially expressed PRs in FO-24 suggest that in common bean, the innate immunity system is activated at this stage. Once, the FOP has established itself in bean tissues, the fragments released in response to disruption of the first line of defense i.e., cell-wall (galacturonic acid-containing fragments) act as signals and mediate defense response by strengthening defensive barriers i.e., CEBiP and CERK. In this regard, the high activation of fungal elicitor immediate early-responsive genes in FO-24 as compared to CK-24 confirms that such signals are received by common bean root tissues [18,22,36]. Involvement of GPI-APs with extracellular ligands such as pathogen molecules as well as other ligands i.e., phytohormones, signaling polypeptides, leads to the phosphorylation of the intracellular kinase domain, which consequently activate cytoplasmic signaling components and switch on the response mechanisms. In FO-24 bean roots, the GPI-anchor biosynthesis pathway was one of the significantly enriched pathways both in transcriptome as well metabolome analysis. This confirms that in common bean, GPI-APs is involved in signal transduction in response to FOP infection [37] (Table S3; Figure 3b).

3.3. Crucial Role of Hormones in Signaling Pathways in Common Bean-FOP Pathosystem

Role of hormones in signaling pathways is well established which involves systematic acquired resistance. It has been suggested that resistance to FOP is mediated by SA, JA, and ET pathways [9,18]. Previous reports on functional characterization of rice Ankyrin repeat containing proteins confirmed their role in defense against pathogen attack, particularly against *Magnaporthe oryzae* [38]. The contrasting expression of 24 Ankyrin repeat containing genes in FO-24 and CK-24 observed in our transcriptome clearly indicates that bean roots, under FOP infection, employ them as a defense response. Activation of the ET responsive genes at FO-24 is an important observation as previously it was known that some ET responsive proteins such as ET-insensitive protein-2 genes are involved in early infection in banana [18]. Hence their activation/expression at 24 hpi suggests that these genes might be involved in latter FOP infection stages and should be investigated further. The unique expression of JA-signaling related genes in FO-24 seedlings clearly indicates that hormone signaling pathways are involved defense responses in common bean against FOP infection (Table S4). However, the role of JA pathway in susceptibility and tolerance FO is still controversial [39,40]. The emerged response to FOP infection in our study confirms the involvement of ET/JA-dependent pathways together with the activation of TIFY, ET-responsive TFs against FOP infection. Similar response has been observed by Sebastiani et al. [9] in melon against FO infection. Our data suggests that FOP infection activates the transcription of auxin related genes exclusively in FO-24 roots which in turn increases the auxin biosynthesis and indicates direct involvement of auxin in common bean-FOP pathosystem. Together with signaling and structural defense responses, our results envisage that common bean employs a cascade of defense mechanisms including structural and signaling responses and that the auxin, ET, and JA are the main hormones involved in common bean-FOP pathosystem similar to what has been reported in tomato and banana [9,18,22].

3.4. FOP Induced Energy Metabolism and Nitrogen Mobilization

Obligate biotrophs depend on host metabolism for intake of nutrients, which is a measure of pathogenicity of the fungus within the host. Many plant defensive compounds are derived from amino acid precursors such as glucosinolates and secondary metabolites [41,42]. Our results confirm that in common bean roots infected with FOP, amino acids and their derivatives such as N-acetlyl-L-tyrosin, L-cystathionin, glutathione oxidized, glutathione reduced form, 5-aminovaleric acid, and Nα-Acetyl-L-arginine were upregulated play a crucial role in defense against FOP. In relation to this, it is important to relate the significantly enriched pathway observed in KEGG enrichment scatter plot i.e., amino sugar and nucleotide sugar metabolism (Figure 3b). This reveals that primary metabolites i.e., amino acids and sugars are playing a critical role in innate defense pathways. The activation and rapid accumulation of amino acids and sugars affect FOP susceptibility as observed in chickpea [12]. The upregulation of glutamate synthase, glutamate dehydrogenase, and aspargine synthase indicate the role of nitrogen mobilization. Significant upregulation of these genes correlated well with metabolite outcome (Tables S4 and S5).

3.5. FOP Resistance in Common Bean is Mediated by Flavonoid Biosynthesis Pathway

It has been established that plants respond to pathogens by increased activation of the phenylpropanoid pathway leading towards biosynthesis of flavonoids, isoflavonoids, and phenolics. In our metabolome, the most significantly enriched pathway was flavonoid biosynthesis pathway (Figure 7). These compounds are regulated by chalcone synthase (CHS), chalcone isomerase (CHI), isoflavone synthase (IFS), isoflavone reductase (IFR), flavanone 3-hydroxylase (F3H), dihydroflavonol 4-reductase (DFR), anthocyanidin synthase (ANS), and leucoanthocyanidin reductase (LAR) genes and manipulated by MYB transcription factors [43]. As detailed in Table S5, these genes (except LAR gene(s)) were upregulated in FO-24 infected common bean roots. Similar results were revealed in chickpea infected with FO [12]. Previously, it was also known that fungal extract can induce these genes (CHI, IFS, and IFR) in Medicago cell suspension culture [44]. Consistently the accumulation of polyphenols, anthocyanins, flavanones, flavones, flavonoids, isoflavones, phenolamides, quinones, and terpenes identified by the metabolome profiles of CK-24 vs. FO-24 infected roots in the present study correlated well with transcriptomic data. Hence, the accumulation of flavonoid biosynthesis related genes and metabolites in FO-24 infected roots suggested their potential involvement in FOP resistance.

4. Materials and Methods

4.1. Plant Growth and In Vivo Inoculations

Seeds of common bean (Liyun No. 2) were obtained from the Lishui Institute of Agricultural Sciences, China. Seeds were sown in 15 cm diameter pots. The growing material filled in pots was sterile vermiculite and clay mixed in a 3:1 volume/volume ratio. The seedlings were allowed to grow under normal conditions i.e., 25 °C/18 °C day/night temperatures with a 16-h light/8-h dark photoperiod, and 60% humidity for 5 days. Plants were separated into two groups: control (CK) and *Fusarium oxysporum* f. sp. *phaseoli* (FOP) treated plants. Five individual seedlings were monitored at 4, 8, 12, and 24 h post infection. Then, 5-day-old seedlings were used for the infection at the fully expanded trifoliate leaves. The inoculum was prepared as reported earlier [3]. The control plants were supplemented with sterile ultrapure water. All the treated and control plants were evaluated for root quantitative traits such as root length, root volume, and fresh weight. As 24 h time point provided the best contrasting phenotype between CK and FO treated plants (Table S1, Figure 1), whole root samples from three individuals (biological replicates) in CK-24 and FO-24 were used for transcriptome and metabolome analysis. All treatments were grown in the same greenhouse with a 16 h light and 8 h dark cycle.

4.2. RNA Extraction, cDNA Library Construction, and Transcriptome Sequencing

Total RNAs were extracted using Spin Column Plant total RNA Purification Kit following the manufacturer's protocol (Sangon Biotech, Shanghai, China) [45]. Purity of the extracted RNAs was assessed on 1% agarose gels as well as by NanoPhotometer spectrophotometer (IMPLEN, Los Angeles, CA, USA). For RNA quantification we used a Qubit RNA Assay Kit in Qubit 2.0 Flurometer (Life Technologies, Carlsbad, CA, USA). Further, RNA integrity was assessed by the RNA Nano 6000 Assay Kit of the Agilent Bioanalyzer 2100 system (Agilent Technologies, Santa Clara, CA, USA).

Sequencing libraries was created using NEB Next Ultra RNA Library Prep Kit following manufacturer's instructions. The index codes were added to each sample. Briefly, the mRNA was purified from 3 μg total RNA of each of three replicate using poly-T oligo-attached magnetic beads. Subsequently, the fragmentation buffer was used to break the RNA into short fragments, and the short-fragment RNA was used as a template to synthesize the first strand cDNA with random hexamers, followed by buffer, dNTPs (dUTP, dATP, dGTP, and dCTP). The double-stranded cDNA was synthesized with DNA polymerase I, and then the double-stranded cDNA was purified using AMPure XP beads. The purified double-stranded cDNA was subjected to terminal repair, A tail was added, and the sequencing linker was ligated, and then AMPure XP beads were used for fragment size selection, and finally PCR enrichment was performed to obtain a final cDNA library. Library quality was initially quantified using Qubit 2.0 using the 2100 to test the insert size of the library followed by accurately quantifying the effective concentration of the library (>2 nM) by Q-PCR. Finally, six paired-end cDNA libraries with an insert size of 300 bp were constructed for transcriptome sequencing and sequenced on Illumina HiSeq platform (Illumina Inc., San Diego, CA, USA) by Wuhan MetWare Biotechnology Co., Ltd. (www.metware.cn).

4.3. De Novo Assembly, Functional Annotation, Classification, and Metabolic Pathway Analysis

The clean reads were retrieved after trimming adapter sequences, removal of low quality (containing > 50% bases with a Phred quality score < 20) and reads with unknown nucleotides (more than 1% ambiguous residues N) using the FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). GC content distribution check was performed. To stitch clean reads, Trinity was used (Version r20140717, [46]). For hierarchical clustering, Corset was used (https://code.google.com/p/corset-project/). The longest cluster sequence was obtained by clustering with Corset hierarchy as Unigene for subsequent analysis. The assembled unigenes were then aligned with various databases such as KEGG [47], GO [48], Clusters of Orthologous Groups (COG) [49], PfAM, Swissprot [50], egNOG [51], NR [52], KOG [53] using BLAST [54] with a threshold of E-value $< 1.0 \times 10^{-5}$.

The software KOBAS2.0 [55] was employed to get the unigene KEGG orthology. The analogs of the unigene amino acid sequences were searched against the Pfam database [56] using HMMER tool [57] with a threshold of E-value $< 1.0 \times 10^{-10}$. The sequenced reads were compared with the unigene library using Bowtie [58], and the level of expression was estimated in combination with RSEM [59]. The gene expression level was determined according to the FPKM.

4.4. Differential Expression and Enrichment Analysis

The read count was normalized and EdgeR Bioconductor package [60] was used to determine the differential expression genes (DEGs) between CK-24 and FO-24 with the fold change > 2 [61] and FDR correction set at $p < 0.01$. GO enrichment analysis was performed using the topGO method based on the wallenius noncentral hypergeometric distribution with $p < 0.05$ [62]. KEGG pathway enrichment analysis of the DEGs was done using KOBAS2.0 [55]. The FDR correction was employed ($p < 0.05$) to reduce false positive prediction of enriched GO terms and KEGG pathways.

4.5. Quantitative RT-PCR Analysis

Eight DEGs, characterized by interesting expression profiles in response to FOP infection in FO-24 common bean plants were selected for qRT-PCR. First strand cDNAs were synthesized from 100 ng of total RNA using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystem, Carlsbad, CA, USA). Primers were designed using Primer3 Software (http://frodo.wi.mit.edu/primer3/; Table S7) and the specificity was checked by blasting their sequences in the NCBI database. The *Actin* constitutively expressed gene was used as the reference gene [3]. All qRT-PCR reactions were carried out on a Rotor-Gene 6000 machine (Qiagen, Shanghai, China) with the following thermal cycling profile: 50 °C for 2 min and 95 °C for 2 min, followed by 40 cycles at 95 °C for 3 s and 60 °C for 30 s. Melting curve analysis was performed to verify single product amplification with temperature ranging from 55 to 95 °C by increasing of 1 °C every step. All reactions were performed in a total volume of 10 µL containing 30 ng of cDNA, 5 µL 1 × SYBR® Select Master Mix (Applied Biosystem, Carlsbad, CA, USA), and 0.2 µL (20 µM) of each primer. For each sample, two biological replicates were analyzed in independent runs and a no-template control was included for each gene. Intra-assay variation was evaluated by performing all reactions in triplicate. The quantification cycle (Cq) was automatically determined using Rotor-Gene 6000 Series Software, version 1.7 as reported earlier [9].

4.6. Widely Targeted Metabolomics

The sample preparation, extract analysis, metabolite identification, and quantification were performed at Wuhan MetWare Biotechnology Co., Ltd. (www.metware.cn) following their standard procedures [30].

4.7. Sample Preparation

The vacuum freeze-dried root samples were crushed using a grinder (MM 400, Retsch, Haan, Germany) to a powder. A total of 100 mg powder was weighed and aliquots were extracted at 4 °C with 0.6 mL 70% aqueous methanol and vortexed six times to increase the extraction rate. After centrifuging at 10,000× *g* for 10 min, the supernatant was aspirated, and the sample was filtered through a microporous membrane (0.22 µm pore size) and stored in a sample bottle for UPLC-MS/MS analysis.

4.8. Chromatographic Mass Spectrometry Acquisition Conditions

The data acquisition instrument system included Ultra Performance Liquid Chromatography (UPLC) (Shim-pack UFLC SHIMADZU CBM30A, https://www.shimadzu.com.cn/) and tandem mass spectrometry (SHIMADZU Corp., Kyoto, Japan) (MS/MS) (Applied Biosystems 4500 QTRAP, http://www.appliedbiosystems.com.cn/). The liquid phase conditions included (1) column: waters ACQUITY UPLC HSS T3 C18 1.8 µm, 2.1 mm × 100 mm; (2) mobile phase: phase A = ultrapure water (0.04% acetic acid was added), phase B = acetonitrile (0.04% acetic acid was added); (3) elution gradient: 0.00 min B = 5% in comparison, B was linearly increased to 95% in 10.00 min, and maintained at 95% 1 min, 11.00–11.10 min, B was reduced to 5%, and was 5% balanced to 14 min; (4) flow rate 0.35 mL/min; column temperature 40 °C; injection volume 4 µL. Whereas the mass spectrometry conditions were as following: the electrospray ionization (ESI) temperature was 550 °C, the mass spectrometry voltage was 5500 V, the curtain gas (CUR) was 30 psi, and the collision-induced dissociation (CAD) parameter was set high. In the triple quadrupole (QQQ), each ion pair was scanned for detection based on optimized decolusting potential (DP) and collision energy (CE) [63].

Based on the self-built database MWDB (metware database) at Wuhan MetWare Biotechnology Co., Ltd. (www.metware.cn), the material was characterized according to the secondary spectrum information. The isotope signal was removed during the analysis, and the repeated signals including K+ ions, Na+ ions, NH4+ ions, and fragment ions which are themselves other larger molecular weight substances were removed.

Metabolite quantification was performed using multiple reaction monitoring (MRM, as shown below) in triple quadrupole mass spectrometry. In the MRM mode, the fourth-stage rod first screens the precursor ions (parent ions) of the target substance, and excludes the ions corresponding to other molecular weight substances to initially eliminate the interference; the precursor ions break through the collision chamber to induce ionization and break to form a lot of fragment ions. The triple quadrupole filter is then used to select a desired feature fragment ion to eliminate non target ion interference, which makes the quantification more accurate and repeatable. After obtaining metabolite mass spectrometry data for different samples, peak area integration was performed on the mass spectral peaks of all the substances, and the mass spectral peaks of the same metabolite in different samples were integrated [64].

4.9. Metabolomics Data Analysis

Data matrices with the intensity of metabolite features under FOP and control conditions were uploaded to the Analyst 1.6.1 software (AB SCIEX, Ontario, Canada). For statistical analysis, missing values were assumed to be below the limits of detection, and these values were imputed with a minimum compound value [63]. The relative abundance of each metabolite was log transformed before analysis to meet normality. A Dunnett's test was used to compare the abundance of each metabolite between control and FOP. False discovery rate was used for controlling multiple testing. The supervised multivariate method, partial least squares-discriminant analysis (PLS-DA), was used to maximize the metabolome difference between the control and FOP treated samples. The relative importance of each metabolite to the PLS-DA model was checked using a parameter called the variable importance in projection (VIP). Metabolites with VIP > 1.0 were considered as differential metabolites for group discrimination. Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA), and KEGG pathway analysis were performed in R software (www.r-project.org).

5. Conclusions

In the present study the whole transcriptome and metabolome of common bean infected by FOP 24 hpi were characterized. The differences in terms of DEGs between the inoculated and non-inoculated common bean showed that nitrogen metabolism and energy metabolism participated in defense response to FOP infection. Flavonoid pathway was the main defense response in common bean. Transcriptome analysis showed that the spliceosome, RNA transport, ribosome biogenesis in eukaryotes, proteasome, and phenylalanine metabolism were the top five significantly enriched pathways. Cell-wall related genes proved to be the first response to FOP attack and started a cascade of signaling leading to accumulation of cell wall degradation related transcripts. PAMP/MAMP, PRRs, and PRs were being regulated in response to FOP infection suggesting triggering of immunity in common bean. Activation of systematic acquired resistance was also observed in our study where the role of hormones in the signaling pathway was observed. These results demonstrate the common bean in response to FOP utilizes different and effective defense pathways comprising of a complex resistance network of structural, signaling, and chemical responses. Further investigations will be focused on functional validation and mapping of the DEGs, which could represent a helpful tool for developing common bean resistant varieties toward FOP.

Supplementary Materials: Supplementary files can be found at http://www.mdpi.com/1422-0067/20/24/6278/s1. Supplementary Table S1. Comparison of means for root quantitative data. Supplementary Table S2. Summary of sequencing output statistics. Supplementary Table S3. Mapping results of RNA-Seq data from inoculated and non-inoculated common bean seedlings. Supplementary Table S4. Complete list of differentially expressed genes regulated by infection of FOP at 24 hpi in control and infected roots. Supplementary Table S5. List of all detected metabolites. Supplementary Table S6. List of differentially expressed metabolites. Supplementary Table S7. List of primers used for qRT-PCR analysis. Supplementary Figure S1. Differential expression gene function annotation and enrichment analysis. Supplementary Figure S2. Orthogonal partial least squares discriminant analysis.

Author Contributions: Conceptualization, L.C., and Q.W.; methodology, L.C.; validation, W.H., T.H., and L.C.; formal analysis, Y.M.; data curation, L.C.; writing—original draft preparation, L.C., Q.W.; writing—review and editing, Q.W., Y.M.; visualization, T.H.; supervision, Q.W.; project administration, W.H.; funding acquisition, Q.W.

Funding: This research was funded by Zhejiang Natural Science Foundation, grant number No. LGN18C140004.

Conflicts of Interest: The authors declare no conflict of interest.

Availability of Data: The data for RNA-sequencing is available at National Center for Biotechnology Information (NCBI) with accession number PRJNA589754.

References

1. Harter, L. A Fusarium disease of beans. *Phytopathology* **1929**, *19*, 1–84.
2. Buruchara, R.A.; Camacho, L. Common bean reaction to *Fusarium oxysporum* f. sp. *phaseoli*, the cause of severe vascular wilt in Central Africa. *J. Phytopathol.* **2000**, *148*, 39–45. [CrossRef]
3. Xue, R.; Wu, J.; Zhu, Z.; Wang, L.; Wang, X.; Wang, S.; Blair, M.W. Differentially expressed genes in resistant and susceptible common bean (*Phaseolus vulgaris* L.) genotypes in response to *Fusarium oxysporum* f. sp. *phaseoli*. *PLoS ONE* **2015**, *10*, e0127698. [CrossRef] [PubMed]
4. Niño-Sánchez, J.; Tello, V.; Casado-del Castillo, V.; Thon, M.R.; Benito, E.P.; Díaz-Mínguez, J.M. Gene expression patterns and dynamics of the colonization of common bean (*Phaseolus vulgaris* L.) by highly virulent and weakly virulent strains of *Fusarium oxysporum*. *Front. Microbiol.* **2015**, *6*, 234. [CrossRef]
5. Batista, R.O.; Silva, J.L.O.; Nicoli, A.; Carneiro, P.C.S.; Carneiro, J.E.d.S.; Júnior, P.; Queiroz, M.V.d. Resistance to Fusarium wilt in common bean. *Crop Breed. Appl. Biotechnol.* **2016**, *16*, 226–233. [CrossRef]
6. De Ascensao, A.R.; Dubery, I.A. Panama disease: Cell wall reinforcement in banana roots in response to elicitors from *Fusarium oxysporum* f. sp. *cubense* race four. *Phytopathology* **2000**, *90*, 1173–1180. [CrossRef]

7. Zhang, L.; Cenci, A.; Rouard, M.; Zhang, D.; Wang, Y.; Tang, W.; Zheng, S.-J. Transcriptomic analysis of resistant and susceptible banana corms in response to infection by *Fusarium oxysporum* f. sp. *cubense* tropical race 4. *Sci. Rep.* **2019**, *9*, 8199. [CrossRef]
8. Zvirin, T.; Herman, R.; Brotman, Y.; Denisov, Y.; Belausov, E.; Freeman, S.; Perl-Treves, R. Differential colonization and defence responses of resistant and susceptible melon lines infected by *Fusarium oxysporum* race 1· 2. *Plant Pathol.* **2010**, *59*, 576–585. [CrossRef]
9. Silvia Sebastiani, M.; Bagnaresi, P.; Sestili, S.; Biselli, C.; Zechini, A.; Orrù, L.; Cattivelli, L.; Ficcadenti, N. ranscriptome analysis of the melon-*Fusarium oxysporum* f. sp. *melonis* race 1.2 pathosystem in susceptible and resistant plants. *Front. Plant Sci.* **2017**, *8*, 362. [CrossRef]
10. Gupta, S.; Chakraborti, D.; Rangi, R.K.; Basu, D.; Das, S. A molecular insight into the early events of Chickpea (*Cicer arietinum*) and *Fusarium oxysporum* f. sp. *ciceri* (Race 1) interaction through cDNA-AFLP analysis. *Phytopathology* **2009**, *99*, 1245–1257. [CrossRef]
11. Gupta, S.; Chakraborti, D.; Sengupta, A.; Basu, D.; Das, S. Primary metabolism of chickpea is the initial target of wound inducing early sensed *Fusarium oxysporum* f. sp. *ciceri* race I. *PLoS ONE* **2010**, *5*, e9030. [CrossRef] [PubMed]
12. Kumar, Y.; Zhang, L.; Panigrahi, P.; Dholakia, B.B.; Dewangan, V.; Chavan, S.G.; Kunjir, S.M.; Wu, X.; Li, N.; Rajmohanan, P.R. *Fusarium oxysporum* mediates systems metabolic reprogramming of chickpea roots as revealed by a combination of proteomics and metabolomics. *Plant Biotechnol. J.* **2016**, *14*, 1589–1603. [CrossRef] [PubMed]
13. Dowd, C.; Wilson, I.W.; McFadden, H. Gene expression profile changes in cotton root and hypocotyl tissues in response to infection with *Fusarium oxysporum* f. sp. *vasinfectum*. *Mol. Plant Microbe Interact.* **2004**, *17*, 654–667. [CrossRef] [PubMed]
14. Lagopodi, A.L.; Ram, A.F.; Lamers, G.E.; Punt, P.J.; Van den Hondel, C.A.; Lugtenberg, B.J.; Bloemberg, G.V. Novel aspects of tomato root colonization and Infection by *Fusarium oxysporum* f. sp. *radicis-lycopersici* revealed by confocal laser scanning microscopic analysis using the green fluorescent protein as a marker. *Mol. Plant Microbe Interact.* **2002**, *15*, 172–179. [CrossRef] [PubMed]
15. Berrocal-Lobo, M.; Molina, A. Arabidopsis defense response against *Fusarium oxysporum*. *Trends Plant Sci.* **2008**, *13*, 145–150. [CrossRef] [PubMed]
16. Kidd, B.N.; Kadoo, N.Y.; Dombrecht, B.; Tekeoglu, M.; Gardiner, D.M.; Thatcher, L.F.; Aitken, E.A.; Schenk, P.M.; Manners, J.M.; Kazan, K. Auxin signaling and transport promote susceptibility to the root-infecting fungal pathogen *Fusarium oxysporum* in Arabidopsis. *Mol. Plant Microbe Interact.* **2011**, *24*, 733–748. [CrossRef]
17. Ramírez-Suero, M.; Khanshour, A.; Martinez, Y.; Rickauer, M. A study on the susceptibility of the model legume plant *Medicago truncatula* to the soil-borne pathogen *Fusarium oxysporum*. *Eur. J. Plant Pathol.* **2010**, *126*, 517–530. [CrossRef]
18. Li, C.-Y.; Deng, G.-M.; Yang, J.; Viljoen, A.; Jin, Y.; Kuang, R.-b.; Zuo, C.-W.; Lv, Z.-C.; Yang, Q.-S.; Sheng, O. Transcriptome profiling of resistant and susceptible Cavendish banana roots following inoculation with Fusarium oxysporum f. sp. cubense tropical race 4. *BMC Genom.* **2012**, *13*, 374. [CrossRef]
19. Klessig, D.F.; Durner, J.; Shah, J.; Yang, Y. Salicylic acid-mediated signal transduction in plant disease resistance. In *Phytochemical Signals and Plant-microbe Interactions*; Romeo, J., Downum, K., Verpoorte, R., Eds.; Springer: Boston, MA, USA, 1998; Volume 32, pp. 119–137.
20. Yang, Y.; Shah, J.; Klessig, D.F. Signal perception and transduction in plant defense responses. *Genes Dev.* **1997**, *11*, 1621–1639. [CrossRef]
21. Nawaz, M.A.; Rehman, H.M.; Imtiaz, M.; Baloch, F.S.; Lee, J.D.; Yang, S.H.; Lee, S.I.; Chung, G. Systems Identification and Characterization of Cell Wall Reassembly and Degradation Related Genes in Glycine max (L.) Merill, a Bioenergy Legume. *Sci. Rep.* **2017**, *7*, 10862. [CrossRef]
22. Bigeard, J.; Colcombet, J.; Hirt, H. Signaling mechanisms in pattern-triggered immunity (PTI). *Mol. Plant* **2015**, *8*, 521–539. [CrossRef] [PubMed]
23. Hondo, D.; Hase, S.; Kanayama, Y.; Yoshikawa, N.; Takenaka, S.; Takahashi, H. The LeATL6-associated ubiquitin/proteasome system may contribute to fungal elicitor-activated defense response via the jasmonic acid-dependent signaling pathway in tomato. *Mol. Plant Microbe Interact.* **2007**, *20*, 72–81. [CrossRef] [PubMed]
24. Rolland, F.; Moore, B.; Sheen, J. Sugar sensing and signaling in plants. *Plant Cell* **2002**, *14*, S185–S205. [CrossRef] [PubMed]

25. García-Limones, C.; Hervás, A.; Navas-Cortés, J.A.; Jiménez-Díaz, R.M.; Tena, M. Induction of an antioxidant enzyme system and other oxidative stress markers associated with compatible and incompatible interactions between chickpea (*Cicer arietinum* L.) and *Fusarium oxysporum* f. sp. *ciceris*. *Physiol. Mol. Plant Pathol.* **2002**, *61*, 325–337.
26. Liu, C.; Hao, F.; Hu, J.; Zhang, W.; Wan, L.; Zhu, L.; Tang, H.; He, G. Revealing different systems responses to brown planthopper infestation for pest susceptible and resistant rice plants with the combined metabonomic and gene-expression analysis. *J. Proteome Res.* **2010**, *9*, 6774–6785. [CrossRef]
27. Liu, Q.; Wang, X.; Tzin, V.; Romeis, J.; Peng, Y.; Li, Y. Combined transcriptome and metabolome analyses to understand the dynamic responses of rice plants to attack by the rice stem borer *Chilo suppressalis* (*Lepidoptera*: *Crambidae*). *BMC Plant Biol.* **2016**, *16*, 259. [CrossRef]
28. Chen, J.-B.; Wang, S.-M.; Jing, R.-L.; Mao, X.-G. Cloning the PvP5CS gene from common bean (*Phaseolus vulgaris*) and its expression patterns under abiotic stresses. *J. Plant Physiol.* **2009**, *166*, 12–19. [CrossRef]
29. Zhu, G.; Wang, S.; Huang, Z.; Zhang, S.; Liao, Q.; Zhang, C.; Lin, T.; Qin, M.; Peng, M.; Yang, C. Rewiring of the fruit metabolome in tomato breeding. *Cell* **2018**, *172*, 249–261.e212. [CrossRef]
30. Zhang, S.; Ying, H.; Pingcuo, G.; Wang, S.; Zhao, F.; Cui, Y.; Shi, J.; Zeng, H.; Zeng, X. Identification of Potential Metabolites Mediating Bird's Selective Feeding on *Prunus mira* Flowers. *BioMed Res. Int.* **2019**, *2019*. [CrossRef]
31. Yuan, H.; Zeng, X.; Shi, J.; Xu, Q.; Wang, Y.; Jabu, D.; Sang, Z.; Nyima, T. Time-course comparative metabolite profiling under osmotic stress in tolerant and sensitive tibetan hulless barley. *BioMed Res. Int.* **2018**. [CrossRef]
32. Yuan, H.; Zeng, X.; Yang, Q.; Xu, Q.; Wang, Y.; Jabu, D.; Sang, Z.; Tashi, N. Gene coexpression network analysis combined with metabonomics reveals the resistance responses to powdery mildew in Tibetan hulless barley. *Sci. Rep.* **2018**, *8*, 14928. [CrossRef] [PubMed]
33. Wang, Y.; Zeng, X.; Xu, Q.; Mei, X.; Yuan, H.; Jiabu, D.; Sang, Z.; Nyima, T. Metabolite profiling in two contrasting Tibetan hulless barley cultivars revealed the core salt-responsive metabolome and key salt-tolerance biomarkers. *AoB Plants* **2019**, *11*, plz021. [CrossRef] [PubMed]
34. Wojtasik, W.; Kulma, A.; Dymińska, L.; Hanuza, J.; Czemplik, M.; Szopa, J. Evaluation of the significance of cell wall polymers in flax infected with a pathogenic strain of *Fusarium oxysporum*. *BMC Plant Biol.* **2016**, *16*, 75. [CrossRef] [PubMed]
35. Jain, D.; Khurana, J.P. Role of pathogenesis-related (PR) proteins in plant defense mechanism. In *Molecular Aspects of Palnt-Pathogen Interact*; Springer: Singapore, 2018. [CrossRef]
36. Andersen, E.J.; Ali, S.; Byamukama, E.; Yen, Y.; Nepal, M.P. Disease resistance mechanisms in plants. *Genes* **2018**, *9*, 339. [CrossRef] [PubMed]
37. Zhou, K. Glycosylphosphatidylinositol-anchored proteins in Arabidopsis and one of their common roles in signaling transduction. *Front. Plant Sci.* **2019**, *10*, 1022. [CrossRef]
38. Mou, S.; Liu, Z.; Guan, D.; Qiu, A.; Lai, Y.; He, S. Functional analysis and expressional characterization of rice ankyrin repeat-containing protein, *OsPIANK1* in basal defense against *Magnaporthe oryzae* attack. *PLoS ONE* **2013**, *8*, e59699. [CrossRef]
39. Anderson, J.P.; Badruzsaufari, E.; Schenk, P.M.; Manners, J.M.; Desmond, O.J.; Ehlert, C.; Maclean, D.J.; Ebert, P.R.; Kazan, K. Antagonistic interaction between abscisic acid and jasmonate-ethylene signaling pathways modulates defense gene expression and disease resistance in Arabidopsis. *Plant Cell* **2004**, *16*, 3460–3479. [CrossRef]
40. Kidd, B.N.; Edgar, C.I.; Kumar, K.K.; Aitken, E.A.; Schenk, P.M.; Manners, J.M.; Kazan, K. The mediator complex subunit PFT1 is a key regulator of jasmonate-dependent defense in Arabidopsis. *Plant Cell* **2009**, *21*, 2237–2252. [CrossRef]
41. Barth, C.; Jander, G. Arabidopsis myrosinases TGG1 and TGG2 have redundant function in glucosinolate breakdown and insect defense. *Plant J.* **2006**, *46*, 549–562. [CrossRef]
42. Fan, J.; Crooks, C.; Creissen, G.; Hill, L.; Fairhurst, S.; Doerner, P.; Lamb, C. Pseudomonas sax genes overcome aliphatic isothiocyanate–mediated non-host resistance in Arabidopsis. *Science* **2011**, *331*, 1185–1188. [CrossRef]
43. Deng, Y.; Li, C.; Li, H.; Lu, S. Identification and characterization of flavonoid biosynthetic enzyme genes in *Salvia miltiorrhiza* (*Lamiaceae*). *Molecules* **2018**, *23*, 1467. [CrossRef] [PubMed]
44. Farag, M.A.; Huhman, D.V.; Dixon, R.A.; Sumner, L.W. Metabolomics reveals novel pathways and differential mechanistic and elicitor-specific responses in phenylpropanoid and isoflavonoid biosynthesis in *Medicago truncatula* cell cultures. *Plant Physiol.* **2008**, *146*, 387–402. [CrossRef] [PubMed]

45. Lee, S.I.; Muthusamy, M.; Nawaz, M.A.; Hong, J.K.; Lim, M.-H.; Kim, J.A.; Jeong, M.-J. Genome-wide analysis of spatiotemporal gene expression patterns during floral organ development in *Brassica rapa*. *Mol. Genet. Genom.* **2019**, 1–18. [CrossRef] [PubMed]
46. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644. [CrossRef]
47. Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **2004**, *32*, D277–D280. [CrossRef]
48. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25. [CrossRef]
49. Tatusov, R.L.; Galperin, M.Y.; Natale, D.A.; Koonin, E.V. The COG Database: A tool for fenome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **2000**, *28*, 33–36. [CrossRef]
50. Apweiler, R.; Bairoch, A.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **2004**, *32*, D115–D119. [CrossRef]
51. Huerta-Cepas, J.; Szklarczyk, D.; Forslund, K.; Cook, H.; Heller, D.; Walter, M.C.; Rattei, T.; Mende, D.R.; Sunagawa, S.; Kuhn, M. eggNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **2015**, *44*, D286–D293. [CrossRef]
52. Deng, Y.; Li, J.; Wu, S.; Zhu, Y.; Chen, Y.; He, F. Integrated nr database in protein annotation system and its localization. *Comput. Eng.* **2006**, *32*, 71–72.
53. Koonin, E.V.; Fedorova, N.D.; Jackson, J.D.; Jacobs, A.R.; Krylov, D.M.; Makarova, K.S.; Mazumder, R.; Mekhedov, S.L.; Nikolskaya, A.N.; Rao, B.S. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **2004**, *5*, R7. [CrossRef] [PubMed]
54. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]
55. Xie, C.; Mao, X.; Huang, J.; Ding, Y.; Wu, J.; Dong, S.; Kong, L.; Gao, G.; Li, C.-Y.; Wei, L. KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* **2011**, *39*, W316–W322. [CrossRef] [PubMed]
56. Finn, R.D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J. Pfam: The protein families database. *Nucleic Acids Res.* **2013**, *42*, D222–D230. [CrossRef]
57. Eddy, S.R. Profile hidden markov models. *Bioinformatics* **1998**, *14*, 755–763. [CrossRef]
58. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **2009**, *10*, R25. [CrossRef]
59. Li, B.; Dewey, C.N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [CrossRef]
60. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef]
61. Anders, S.; McCarthy, D.J.; Chen, Y.; Okoniewski, M.; Smyth, G.K.; Huber, W.; Robinson, M.D. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.* **2013**, *8*, 1765. [CrossRef]
62. Young, M.D.; Wakefield, M.J.; Smyth, G.K.; Oshlack, A. Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biol.* **2010**, *11*, R14. [CrossRef]
63. Chen, W.; Gong, L.; Guo, Z.; Wang, W.; Zhang, H.; Liu, X.; Yu, S.; Xiong, L.; Luo, J. A novel integrated method for large-scale detection, identification, and quantification of widely targeted metabolites: Application in the study of rice metabolomics. *Mol. Plant* **2013**, *6*, 1769–1780. [CrossRef]
64. Fraga, C.G.; Clowers, B.H.; Moore, R.J.; Zink, E.M. Signature-discovery approach for sample matching of a nerve-agent precursor using liquid chromatography− mass spectrometry, XCMS, and chemometrics. *Anal. Chem.* **2010**, *82*, 4165–4173. [CrossRef]

International Journal of
Molecular Sciences

Article

Molecular Analysis of UV-C Induced Resveratrol Accumulation in *Polygonum cuspidatum* Leaves

Zhongyu Liu, Junxiong Xu, Xiang Wu, Yanyan Wang, Yanli Lin, Duanyang Wu, Hongjie Zhang and Jianbing Qin *

College of Life Science, Yangtze University, Jingzhou 434025, China; yzrs91@163.com (Z.L.); 2017714456@yangtzeu.edu.cn (J.X.); xufyz20016@126.com (X.W.); llqww17@126.com (Y.W.); linyanli1998@foxmail.com (Y.L.); kdysuperme@163.com (D.W.); zhk18672558067@163.com (H.Z.)

* Correspondence: yzpc19@126.com; Tel.: +186-7256-3466

Received: 11 November 2019; Accepted: 4 December 2019; Published: 7 December 2019

Abstract: Resveratrol is one of the most studied plant secondary metabolites owing to its numerous health benefits. It is accumulated in some plants following biotic and abiotic stress pressures, including UV-C irradiation. *Polygonum cuspidatum* represents the major natural source of concentrated resveratrol but the underlying mechanisms as well as the effects of UV-C irradiation on resveratrol content have not yet been documented. Herein, we found that UV-C irradiation significantly increased by 2.6-fold and 1.6-fold the resveratrol content in irradiated leaf samples followed by a dark incubation for 6 h and 12 h, respectively, compared to the untreated samples. De novo transcriptome sequencing and assembly resulted into 165,013 unigenes with 98 unigenes mapped to the resveratrol biosynthetic pathway. Differential expression analysis showed that *P. cuspidatum* strongly induced the genes directly involved in the resveratrol synthesis, including phenylalanine ammonia-lyase, cinnamic acid 4-hydroxylase, 4-coumarate-CoA ligase and stilbene synthase (*STS*) genes, while strongly decreased the chalcone synthase (*CHS*) genes after exposure to UV-C. Since CHS and STS share the same substrate, *P. cuspidatum* tends to preferentially divert the substrate to the resveratrol synthesis pathway under UV-C treatment. We identified several members of the MYB, bHLH and ERF families as potential regulators of the resveratrol biosynthesis genes.

Keywords: regulation; RNA-seq; abiotic stress; biosynthesis pathway; chalcones; stilbenes

1. Introduction

Polygonum cuspidatum Sieb. et Zucc. is a member of the buckwheat family (Polygonaceae). It is a tall and resilient herbaceous perennial with woody rhizomes [1], native to East Asia in countries such as Korea, Japan and China. Although this plant has been recognized as an invasive species in Europe and North America [2,3], *P. cuspidatum* has an extraordinary value in phytotherapy. In China, the roots of *P. cuspidatum* have been long employed in traditional medicine to combat cough, hepatitis, infection, arthralgia, tumors, bronchitis, jaundice, bleeding, amenorrhea and hypertension [4–6]. Analytic investigations of the major health-promoting molecules of *P. cuspidatum* roots have revealed the presence of several lead compounds belonging to the group of polyphenols [7,8]. Distinctly, one of the most important compounds in *P. cuspidatum* roots, which has drawn growing interest on this species, is resveratrol, a molecule with proven anti-inflammatory and antioxidant activity [9,10].

The "French paradox", a curious observation referring to the low level of coronary heart disease in France despite high intake of dietary cholesterol and saturated fat [11], has been linked to the high consumption of red wine containing resveratrol. Resveratrol (3,5,40-trihydroxy-trans-stilbene) is a naturally occurring stilbene metabolite found in grapes, berries, nuts and other plants such as *P. cuspidatum*. Over the past decades, extensive researches have been carried out on the physiological functions of resveratrol in human and it has been suggested that resveratrol is a potential remedy for a range of diseases, including heart disease, diabetes, cancer and Alzheimer's disease [12]. The compound was first discovered in *P. cuspidatum* [13], which is the most important concentrated source of free or glycosylated resveratrol. Therefore, it is expected that this species should be a model plant to study resveratrol biosynthesis and engineering in plant. Surprisingly, very limited advances in these fields have been made so far in *P. cuspidatum* [14], contrasting with the extensive knowledge generated in grape (reviewed by Hasan and Bae, [15]). In fact, the lack of genomic resources for *P*. *cuspidatum* hinders genetic discoveries. Particularly, the key genes and molecular mechanisms underlying the striking accumulation of resveratrol in this species are still unknown. On the opposite, early genome sequencing of grape [16] has triggered much research on the biosynthesis of stilbenes. The biosynthetic pathway of resveratrol has been well characterized and involves four key enzymes, including phenylalanine ammonia lyase (PAL), cinnamic acid 4-hydroxylase (C4H), 4-coumarate: CoA ligase (4CL) and stilbene synthase (STS) [17]. p-coumaroyl-CoA is a product of PAL, which is abundant in plants and used as a precursor for the synthesis of both resveratrol and chalcone. Therefore, in stilbene-synthesizing plants, STS competes with chalcone synthase (CHS) for the synthesis of resveratrol [18].

Resveratrol is basically a defense compound (phytoalexin) in plants but it is produced in very small amounts [19]. Therefore, how to induce a strong accumulation of resveratrol in plant organs in order to satisfy the increasing global demand of resveratrol has become one of the hot topics in secondary metabolite research. Manipulation of different synthetic enzymes and the identification of their regulator genes such as *MYB14*, *MYB15* and *WRKY8* [20,21] provide currently prospects to increase resveratrol production in planta. Moreover, it has been demonstrated that biotic and abiotic factors, including fungal attack, UV-C irradiation, jasmonic acid, salicylic acid, H_2O_2 and $AlCl_3$, can induce the accumulation of resveratrol in grape [22–28].

Hao et al. [14] previously identified 54 unigenes predicted to participate in the resveratrol biosynthesis pathway in *P. cuspidatum* roots. However, the mechanism leading to the high resveratrol accumulation in this species and the responses of these genes to abiotic factors such as UV-C irradiation remain open to study. Here, we investigated the effect of UV-C irradiation on leaf resveratrol content in *P. cuspidatum* and further de novo sequenced the transcriptome, offering a novel insight into the UV-C induced biosynthesis of resveratrol in plants.

2. Results

2.1. Effect of UV-C Irradiation on Leaf Resveratrol Content in Polygonum cuspidatum

Polygonum cuspidatum leaves were treated with UV-C light for 10 min and then incubated in the dark for 6 h (PC6H) and 12 h (PC12H). We quantified the leaf resveratrol accumulation in UV-C treated leaves and untreated leaves (PC). As shown in Figure 1, we observed a significant difference in the resveratrol contents of UV-C treated leaves as compared to untreated leaves ($p < 0.001$) and between the different incubation times of treated leaves ($p < 0.001$). UV-C significantly increased the leaf resveratrol contents and 6 h incubation time after UV-C treatment showed the highest accumulation of resveratrol in the leaves of *P. cuspidatum*.

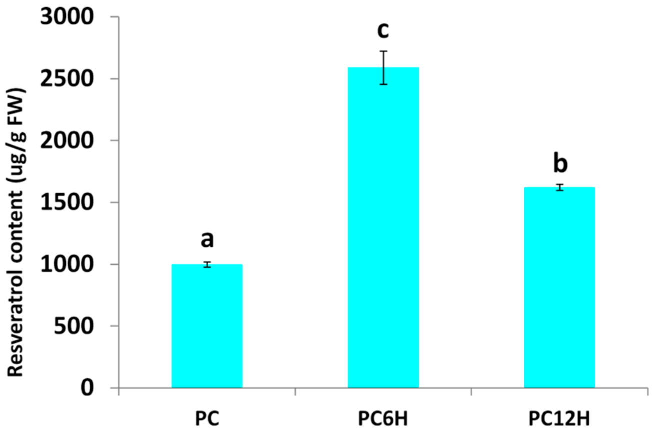


Figure 1. Effect of UV-C irradiation and incubation times on the resveratrol contents in *Polygonum*.

2.2. De Novo Transcriptome Sequencing and Unigene Assembly

In order to elucidate the molecular mechanism underlying the UV-C induced resveratrol accumulation in *Polygonum cuspidatum* leaves, we synthesized nine cDNA libraries from leaf samples of PC12H, PC6H, PC and generated de novo RNA-sequencing data.

A total of 73.63 Gb raw read data was generated. After cleaning, 98.7% of the reads were kept as clean reads with 93.5% of bases scoring Q30 and above (Table 1). The assembly was performed using the Trinity software and a total of 165,013 unigenes were obtained with N50 length about 1744 bp (Table 2). The unigene lengths ranged from 200 to 17,269 bp with a significant number of genes having length of 200–300 bp (Figure 2).

Table 1. Overview of the transcriptome sequencing dataset and quality check. *Polygonum cuspidatum* mature leaves. PC, PC6H and PC12H represent samples collected before, 6 h and 12 h after UV-C treatment, respectively. The letters above the bar represent significant difference between samples.

Type	Read Length	Total Clean Reads	Clean Reads Q20 (%)	Clean Reads Q30 (%)	Detected Genes
P12H1	150	55,184,328	97.81	93.76	122,534
P12H2	150	57,685,708	97.47	93.21	123,131
P12H3	150	60,727,642	97.83	93.85	124,005
P6H1	150	51,083,608	97.66	93.41	122,200
P6H2	150	53,076,416	97.64	93.36	120,570
P6H3	150	48,587,712	97.74	93.56	119,464
PC1	150	53,966,128	97.52	93.14	123,390
PC2	150	51,267,598	97.5	93.19	117,489
PC3	150	52,633,030	97.77	93.69	125,163

Table 2. Statistics of the unigene assembly results.

Sample	Total_Number	Min_Length	Max_length	Mean_Length	N50	N90	GC
All	165,013	200	17,269	1,102.73	1,744	489	0.4216

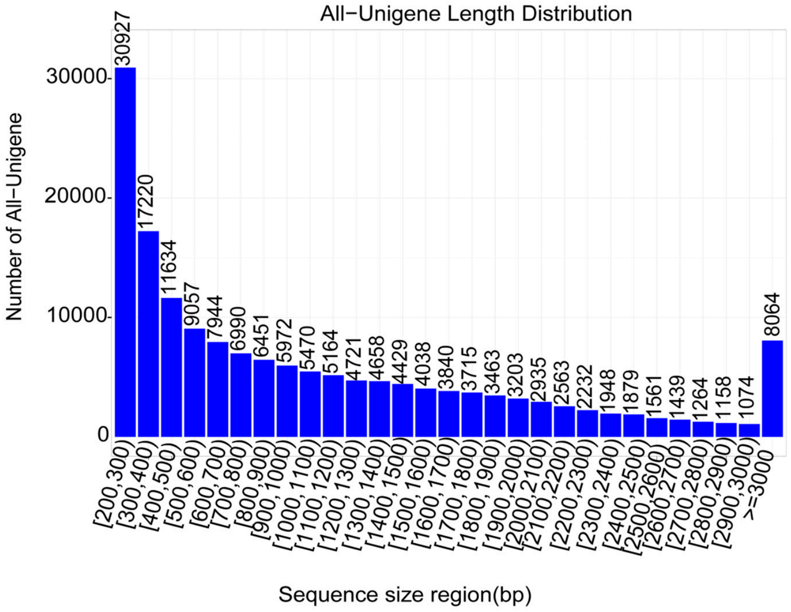


Figure 2. Unigene length distribution. The numbers above the bar represent the number of unigenes with a sequence length comprised in the classes of size displayed in the x-axis.

We performed the functional annotation of the unigenes in various databases, including NR, NT, Swiss-Prot, Kyoto Encyclopedia of Genes and Genomes (KEGG), Clusters of Orthologous Groups (COG) and Gene Ontology (GO), which resulted in 104,903 annotated unigenes (Figure 3A) with 32,448 unigenes successfully annotated based on all the five databases (Figure 3B). Overall, these unigenes contributed to various GO terms of which the most represented ones were binding, catalytic activity and transporter activity (molecular function). They were mainly located in the cell, cell part and membrane (cell component) and involved in the cellular process, metabolic process, etc. (biological process; Figure S1). We further analyzed the genes encoding for transcription factors within the annotated unigenes. As shown in Table 3, we identified 5526 TFs classified into 58 different TF families with bHLH, MYB, bZIP and ERF as the most abundant TFs. In addition, we searched for all genes belonging to the resveratrol biosynthetic pathway and identified 98 unigenes including, 26 phenylalanine ammonia-lyase (PAL), 15 cinnamic acid 4-hydroxylase (C4H), 20 4-coumarate-CoA ligase (4CL), 4 stilbene synthase (STS) and 33 chalcone synthase (CHS; Table S1).

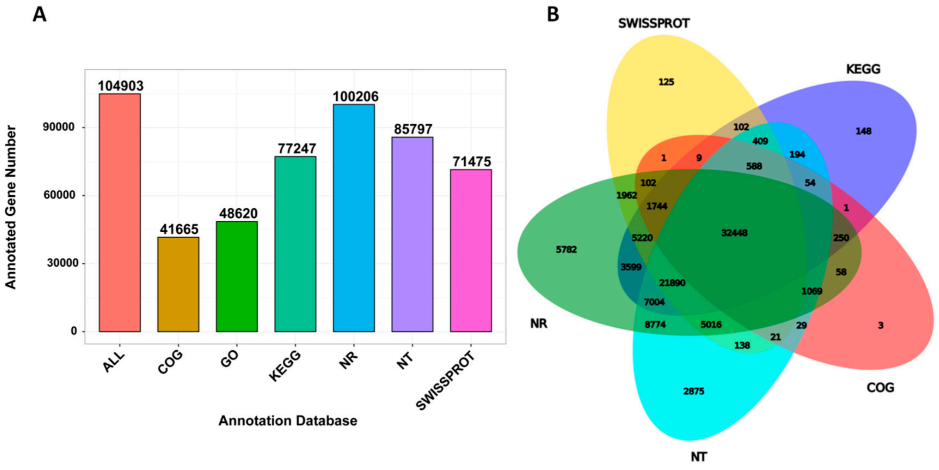


Figure 3. Unigene functional annotation in *P. cuspidatum*. (**A**) The annotated gene number in different databases and (**B**) Venn diagram showing the number of shared and unique annotated genes in the different databases.

Table 3. Statistics of genes encoding for transcription factors.

TF	Count	TF	Count	TF	Count
LFY	3	M-type	5	HB-PHD	17
SAP	23	GATA	105	E2F/DP	23
WRKY	298	DBB	44	Trihelix	126
RAV	13	ZF-HD	35	ARF	137
CPP	35	LBD	50	MIKC	98
HRT-like	3	MYB	402	SRS	15
HB-other	45	LSD	14	NF-X1	8
FAR1	67	WOX	15	Nin-like	43
ERF	341	C2H2	204	MYB_related	272
Whirly	11	CAMTA	23	VOZ	18
ARR-B	57	GRAS	109	G2-like	157
HSF	90	GRF	32	NF-YA	56
NZZ/SPL	8	EIL	36	Dof	119

Next, the clean reads data were mapped to the assembled unigene libraries and the statistics of mapping results are presented in Table S2. A total of 157,665 genes were expressed with the number of fragments per kilobase of exon per million fragments mapped (FPKM) values raging from 0.02 to 18,562.92 (Figure 4A). Principal component analysis (PCA) of the samples based on FPKM showed that all the biological replicates clustered together, suggesting a high reliability of our RNA-sequencing data (Figure 4B). Furthermore, the PCA clearly distinguished the control and the UV-C treated leaf samples, indicating that a large number of genes were differentially expressed after exposure to UV-C irradiation.

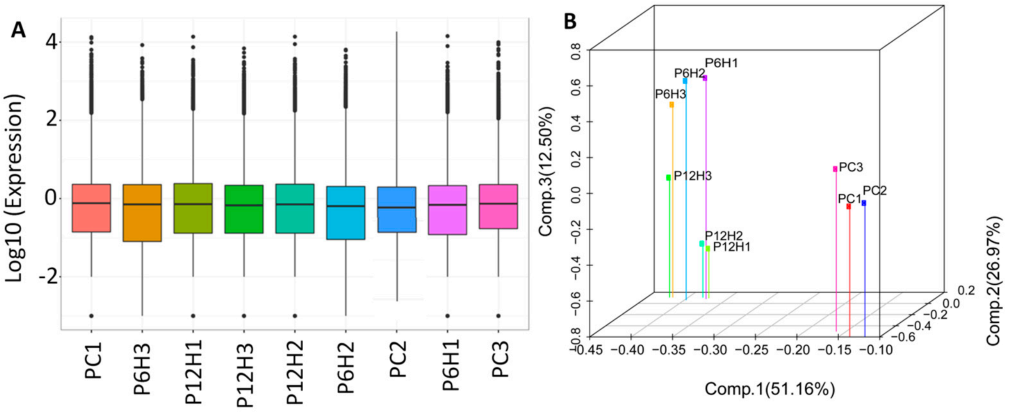


Figure 4. Overview of the transcriptome sequencing in *P. cuspidatum* leaves. (**A**) Gene expression profiles in the nine libraries. PC, PC6H and PC12H represent samples collected before, 6 h and 12 h after UV-C treatment, respectively and (**B**) 3D principal component analysis showing clustering pattern among leaf samples based on global gene expression profiles.

2.3. Differential Expressed Genes (DEG) between Control and UV-C Treated Leaf Samples

We compared the gene expression levels between control samples (PC) to UV-C treated samples PC6H and PC12H with the aim to detect the genes affected by the UV-C treatment. Concerning PC vs. PC6H, we identified 38,985 differentially expressed genes (DEGs) with 17,859 and 21,126 up- and down-regulated, respectively. A slightly lower number of DEGs (32,312) was found for PC vs. PC12H, including 14,416 and 17,896 up and down-regulated genes, respectively. Cross-comparison of the two sets of DEGs showed that in total 45,222 genes were affected by the UV-C treatment and 26,075 DEGs were constantly conserved after 6 h and 12 h post UV-C irradiation (Figure 5A). These genes represent the key genes involved in the metabolic changes in response to UV-C treatment. To get insight into the biological pathways contributed by these DEGs, we performed KEGG enrichment analysis. The results indicated that the DEGs play various roles but were mainly involved in metabolic pathways and biosynthesis of secondary metabolites. In addition, ribosome, plant hormone transduction and phenylpropanoid biosynthesis were the other enriched pathways contributed by these DEGs (Figure 5B).

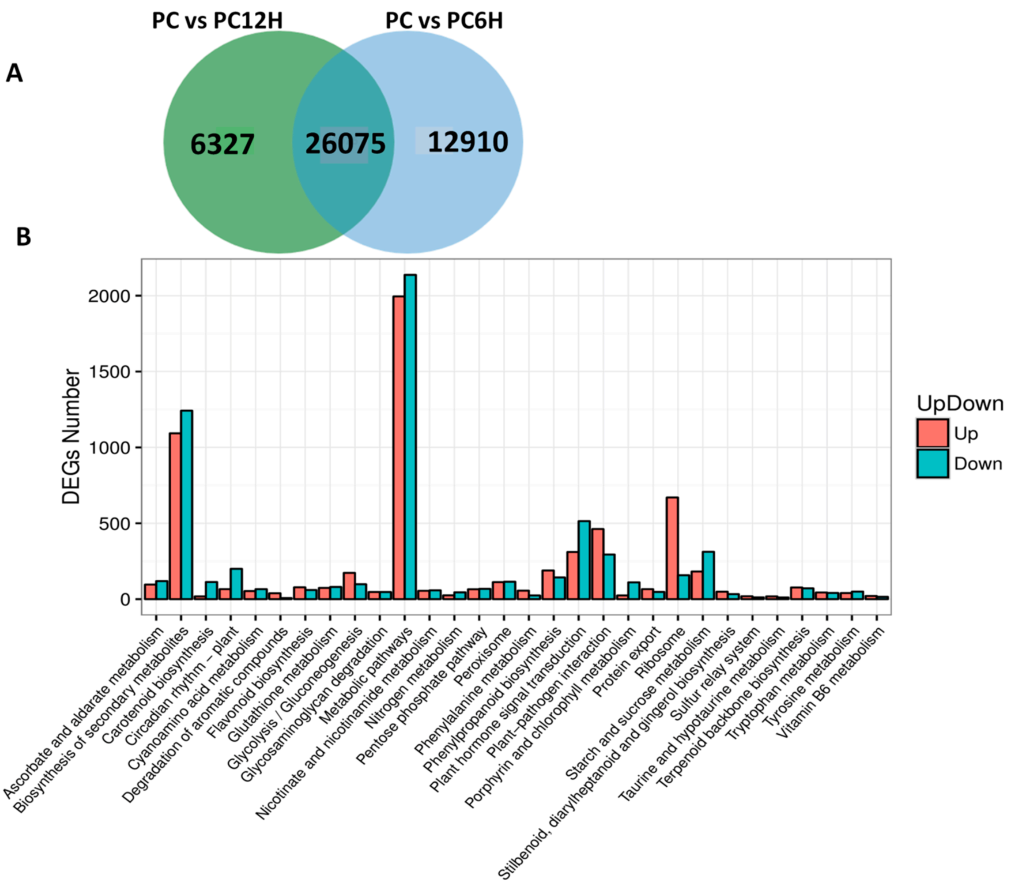


Figure 5. Differentially expressed genes (DEG) between the UV-C treated samples and the untreated control. (**A**) Venn diagram showing the unique and conserved DEGs between PC vs. PC6H and PC vs. PC12H; (**B**) Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis of all the identified DEGs. PC, PC6H and PC12H represent samples collected before, 6 h and 12 h after UV-C treatment, respectively.

Gene expression levels are modulated by regulators such as transcription factors [29]. We extended the analysis to detect the major transcription factor families involved in the response to UV-C exposure in the *Polygonum cuspidatum* leaf. Surprisingly, nearly half (2461) of the total annotated TF genes (5526) in *P. cuspidatum* was found differentially expressed and included 2132 and 1895 TFs DEGs detected at 6 h and 12 h, respectively. The majority of these genes were MYB, bHLH and ERF family members (Table S3).

2.4. Mapping of DEGs Related to the Resveratrol Biosynthetic Pathway

Deamination of phenylalanine ammonia by PAL is the first step in the resveratrol biosynthesis pathway. Then, the conversions of cinnamic acid into p-coumaric acid and subsequently into 4-coumarate-CoA are catalyzed by C4H and 4CL, respectively. The last step in the pathway consists of the conversion of one 4-coumarate-CoA and three malonyl-CoA units into resveratrol or naringenin chalcone by STS or CHS, respectively. Later, resveratrol is converted into pterostilbene by resveratrol O-methyltransferase (ROMT). We searched within the DEGs all genes encoding the key enzymes involved in the resveratrol biosynthesis pathway and successfully identified 37 related DEGs, including

36 and 34 at 6 h and 12 h post UV-C exposure, respectively. These genes included 10 PAL, eight *C4H*, nine *4CL*, two *STS*, six *CHS* and two *ROMT*.

We mapped the 37 DEGs related to the resveratrol biosynthesis along with their gene expression fold change in order to understand the effect of UV-C radiation on resveratrol biosynthesis (Figure 6). Interestingly, we observed that all the 29 genes directly involved in the resveratrol biosynthesis (*PAL, C4H, 4CL* and *STS*) were significantly up-regulated after exposure to UV-C. Meanwhile, the six *CHS* genes were found strongly down-regulated. Concerning the DEGs encoding ROMT, all of the two genes were found sharply induced by UV-C treatment in *Polygonum cuspidatum* leaf (Figure 6). Globally, it appeared that the regulation of these 37 key genes was more intense after 6 h incubation of the leaves in the dark post UV-C irradiation as compared to 12 h.

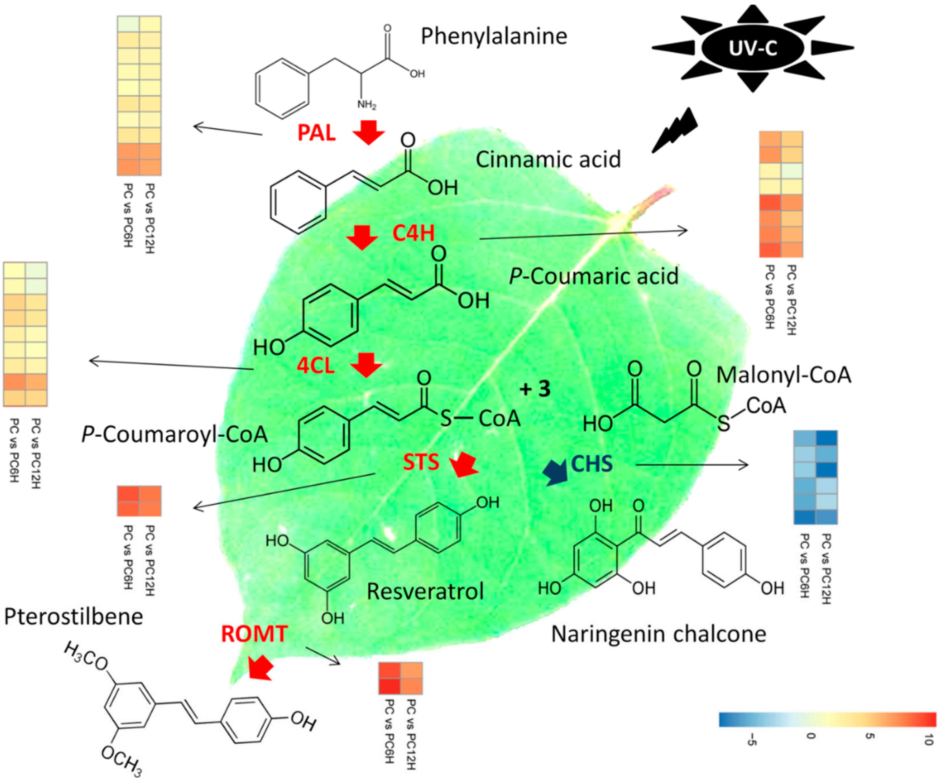


Figure 6. Proposed model of the molecular mechanism leading to the high accumulation of resveratrol after treatment with UV-C in *Polygonum cuspidatum* leaf. Deamination of phenylalanine ammonia by phenylalanine ammonia-lyase (PAL) is the first step in the resveratrol biosynthesis pathway. Then, the conversions of cinnamic acid into *p*-coumaric acid and subsequently into 4-coumarate-CoA are catalyzed by cinnamic acid 4-hydroxylase (C4H) and 4-coumarate-CoA ligase (4CL), respectively. The last step in the pathway consists of the conversion of one 4-coumarate-CoA and three malonyl-CoA units into resveratrol or naringenin chalcone by stilbene synthase (STS) or chalcone synthase (CHS), respectively. Later, resveratrol is converted into pterostilbene by resveratrol O-methyltransferase (ROMT). PAL, C4H, 4CL and STS were found strongly up-regulated in PC6H and PC12H as compared to PC, while CHS displayed the opposite trend. STS and CHS share the same substrate. *P. cuspidatum* tends to prioritize resveratrol accumulation by diverting the substrate “one 4-coumarate-CoA and three malonyl-CoA units” to the resveratrol synthesis pathway over the naringenin chalcone synthesis pathway through up-regulation of STS genes and down-regulation of CHS genes in response to UV-C exposure. High induction of ROMT also suggests that pterostilbenes may also be accumulated. PC, PC6H and PC12H represent samples collected before, 6 h and 12 h after UV-C treatment, respectively.

2.5. qRT-PCR Validation of Candidate Genes

In order to validate the results obtained through the RNA-seq, we selected 12 candidate DEGs from the enzymes involved in the resveratrol biosynthesis and some highly regulated transcription factors genes and conducted a qRT-PCR analysis (Table S4). As shown in the Figure 7, the expression levels of all the tested genes were significantly altered after treatment with UV-C and the qRT-PCR results were strongly correlated with the RNA-seq reports ($R^2 = 0.81$, Figure S2). These results demonstrate the reliability of the RNA-seq data obtained in the present study.

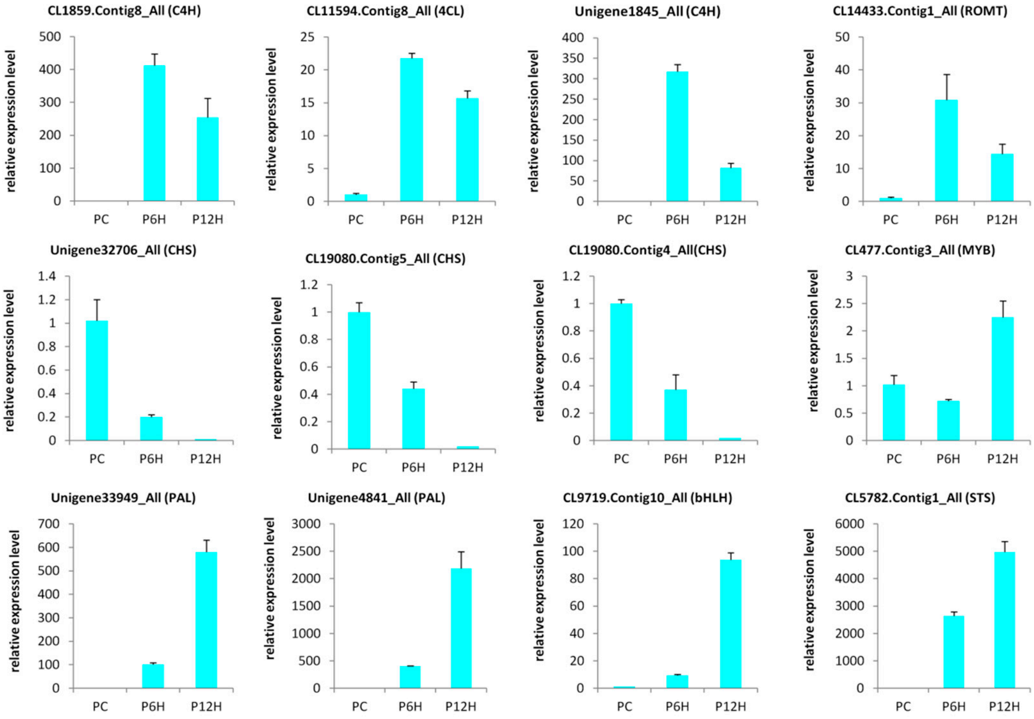


Figure 7. Quantitative real time PCR validation of selected candidate genes. The error bar represents the SD of three biological replicates. The *Actin* gene was used as the internal reference gene for normalization; PC, PC6H and PC12H represent samples collected before, 6 h and 12 h after UV-C treatment, respectively.

3. Discussion

Resveratrol exhibits diverse beneficial properties in humans, including anti-inflammatory effects, anti-tumor activities and anti-aging effects, which have drawn extensive studies on this precious molecule [12]. As a phytoalexin, resveratrol plays important functions as an antimicrobial and antioxidant compound in plant responses to environmental stresses, such as UV irradiation and fungal infection [15]. Plants have developed various alleviation mechanisms to mitigate short wavelength UV-C radiation damaging effects, including a strong accumulation of UV-absorbing phenolic and flavonoid molecules in epidermal cells to block light penetration and anti-oxidative molecules to limit photo-oxidative damage [30–32]. In grape, UV-C irradiation induces a strong accumulation of resveratrol [28,33,34]. It was reported that UV-C could significantly stimulate the synthesis of resveratrol in *Vitis vinifera* × *V. amurensis*, mainly in the form of trans-resveratrol [35]. Similarly, UV-C treatments increased resveratrol synthesis in *Gnetum parvifolium* [36]. These reports suggest that the UV-C induction of resveratrol is quite conserved in stilbenoid-synthesizing plants. Although, *P. cuspidatum* represents the highest natural source of resveratrol known to date, no information is available regarding the effect of abiotic stress such as UV-C treatment on the level of resveratrol. Herein, we studied the effect of UV-C irradiation on the resveratrol metabolism in *P. cuspidatum* mature leaves. As expected, we found that *P. cuspidatum* leaf contained a very high level of resveratrol (1000 μg/g

FW) and UV-C could significantly induce the resveratrol accumulation (Figure 1), indicating that UV-C irradiation represented a good prospect for increasing resveratrol content in *P. cuspidatum*. Different UV irradiation treatments resulted in various levels of resveratrol, therefore an optimized UV treatment protocol is important for a significant induction of resveratrol. In grape, irradiation strength ranging from 30 to 510 W up to 1 min, followed by incubation in 2–4 days, resulted in 10.8-fold higher accumulation of resveratrol than that observed in the untreated control [37,38]. Cantos et al. [39] also showed that a distance of 40 cm, irradiation time of 30 s, source power of 500 W and storage time of 3 days generated the highest resveratrol accumulation. In this study, we incubated the UV-C irradiated leaves for 6 and 12 h. We observed an increase of 2.6-fold and 1.6-fold of resveratrol after 6 h and 12 h incubation times, respectively, suggesting that the effect of the UV-C treatment diminished after 6 h. Although the induction levels were not high as compared to studies in grape [37,38], based on our results and pending more optimization of the UV-C treatment protocol, we might recommend 10 min UV-C irradiation and 6 h incubation in the dark for obtaining a relatively high resveratrol in *P. cuspidatum* leaves.

To understand the molecular mechanism underlying the strong accumulation of resveratrol in *P. cuspidatum* after exposure to UV-C, we de novo sequenced the transcriptome and assembled the unigenes (Tables 1 and 2; Figure 2). In total, 165,013 unigenes were identified in this study, a number that is quite the double of the unigenes number (86,418) detected by Hao et al. [14]. The high discrepancy between the detected unigene numbers in both studies could result from differences in the tissue types, the employed software for assembly and more importantly the sequencing platform. Differential gene expression (DEG) analysis showed that UV-C treatments affected nearly $\frac{1}{4}$ of the expressed genes and various cellular processes including hormones and secondary metabolism were altered (Figure 5). Our results were in perfect concordance with works by Yin et al. [40], who demonstrated that more than 100 functional subcategories were contributed by the DEGs between UV-treated grape berries and untreated samples and particularly, "response to stress" and "metabolic processes" were the most represented terms.

Previous studies have demonstrated that UV-C irradiation alters the activity levels of several structural genes participating in the resveratrol biosynthesis pathway [35,36,40–42]. Within the DEGs detected in this study, 37 were mapped to the resveratrol biosynthesis pathway (Figure 6). Interestingly, all the genes directly involved in the resveratrol biosynthesis (10 phenylalanine ammonia-lyase (PAL), 8 cinnamic acid 4-hydroxylase (C4H), 9 4-coumarate-CoA ligase (4CL) and 2 stilbene synthase (STS)) were significantly up-regulated in UV-C treated samples as compared to untreated samples (Figure 6). Conversely, the six chalcone synthase (CHS) DEGs were all down-regulated in samples exposed to UV-C (Figure 6). In fact, CHS and STS enzymes use the same substrate (one 4-coumarate-CoA and three malonyl-CoA units) for the production of naringenin chalcone and resveratrol, respectively. This result indicates that after exposure to UV-C, *P. cuspidatum* diverts the substrate "one 4-coumarate-CoA and three malonyl-CoA units" to the resveratrol synthesis pathway over the naringenin chalcone synthesis pathway by up-regulating STS and down-regulating CHS. The conclusions of several authors, including Xi et al. [41], Suzuki et al. [42] and Yin et al. [40] support well our findings as they demonstrated that since STS and CHS share the same substrate, there may be a competitive and/or inhibitory relationship between them in response to UV-C exposure, which may in turn play a vital role in resveratrol accumulation in grape berries.

The tyrosine ammonia-lyase (TAL) enzyme can also utilize L-tyrosine as a substrate to produce p-coumaric acid, which subsequently is converted into resveratrol by STS [43]. In this study, we did not find any *TAL* gene among the DEGs, indicating that the high accumulation of resveratrol after UV-C treatments was essentially due to the strong conversion of phenylalanine and cinnamic acid through PAL and C4H, respectively (Figure 6).

Trans-resveratrol is highly instable upon exposure to light and oxygen or under harsh PH, leading to the reduction its bioavailability and bioactivity [44]. Other natural stilbenes derived from resveratrol such as pterostilbene and pinostilbene, display higher oral bioavailability and bioactivity, therefore, efforts are ongoing to develop strategies to obtain these bioactive resveratrol derivatives in abundance. An efficient technique to achieve this goal is the manipulation of resveratrol O-methyltransferase (ROMT), which is the enzyme that converts resveratrol into pterostilbenes [27,45,46]. Here, we observed that 2 ROMT genes were significantly up-regulated in UV-C treated *P. cuspidatum* leaf samples (Figure 6), suggesting that UV-C treatments not only increase the trans-resveratrol levels but may also increase the methylated derivatives of resveratrol. Comparing 6 h and 12 h incubation times with respect to the expression levels of resveratrol biosynthesis related DEGs, we found that gene induction/repression was more vigorous at 6 h than 12 h, which correlates with the higher resveratrol content observed at this time point through resveratrol quantification.

Transcription factors (TF) such as MYB and bHLH were reported to regulate the expression levels of several genes involved in the phenylpropanoid pathways [47,48]. Following UV-C treatment, the gene Myb14 has been shown to activate *STS* genes in grape [20,35,40,49]. In this study, we found 2461 TFs among the DEGs with members of MYB, bHLH and ERF being the most active after UV-C irradiation (Table S3) and might play crucial roles in UV-C induced responses in *P. cuspidatum*. We did not observe a clear trend in the differential expression of genes belonging to these families as many family members were either down-regulated or up-regulated after exposure to UV-C. This highlights the complex mechanisms of UV-C transcriptional regulation and renders the identification of potential master regulators of resveratrol biosynthetic genes very difficult. Gene co-expression analysis is a widely used technique to break down large transcriptome datasets into co-expressed modules in order to pinpoint key TFs modulating important structural genes [50–56]. For example, this approach has been used to discover WRKY TFs (WRKY24/28/29/37/41) that were co-expressed simultaneously with eight *STS* genes (*STS12/13/16/17/18/24/27/29*) in roots and leaves of *Vitis vinifera* [57]. Later on, TFs belonging to different families such as MYB, WRKY and ERF were shown to putatively contribute to STS regulation [58,59]. Therefore, we proposed further RNA-sequencing in *P. cuspidatum* using various genotypes and tissues to comprehensively undertake an integrated gene co-expression analysis in order to find out the key regulators of the resveratrol biosynthetic genes in this important medicinal plant species.

4. Materials and Methods

4.1. Plant Materials

Polygonum cuspidatum Sieb. et Zucc. was used as plant material in this study. Plants were maintained in the medicinal plant garden at Yangtze University, Jingzhou, China. Healthy, mature (30-days old) leaves of similar size were detached from the shoot and were immediately immersed in water and subsequently transferred into ddH_2O [60]. All leaves were incubated in the dark at 25 °C for 30 min. Then, the leaf abaxial surfaces were exposed for 10 min to 6 W/m^2 UV-C irradiation provided by a UV-C lamp (Model CJ-1450, Sujie Purification, China). Samples were collected at 0 h (before UV-C treatment, PC), 6 h (PC6H) and 12 h (PC12H) after the initiation of treatment. Each treatment was performed three times and each replication consisted of six leaves. After sampling, the leaves were ground into powder in liquid nitrogen and stored at −80 °C until analysis.

4.2. Quantification of Resveratrol in Polygonum cuspidatum Leaves

The sample preparation, extract analysis, resveratrol identification and quantification were performed at Wuhan MetWare Biotechnology Co., Ltd, Wuhan, China. (www.metware.cn) following their standard procedures and previously described by Zhang et al. [61]. The experimental measurements were carried out in triplicate and results were presented as average of three analyses ± standard deviation. Statistical analysis of the data was done using the (www.r-project.org) using the one-way analysis of variance for testing statistical significance between different samples. Mean comparisons were done using the Tukey HSD test.

4.3. RNA Extraction, cDNA Library Construction and Transcriptome Sequencing

The leaf samples from PC, PC6H and PC12H were used for total RNAs extraction employing the Spin Column Plant total RNA Purification Kit according to the manufacturer's protocol (Sangon Biotech, Shanghai, China). Purity of the extracted RNAs was assessed on 1% agarose gel followed by a NanoPhotometer spectrophotometer (IMPLEN, Westlake Village, CA, USA). RNA quantification was performed using the Qubit RNA Assay Kit in Qubit 2.0 Flurometer (Life Technologies, Carlsbad, CA, USA). Next, RNA integrity was checked by the RNA Nano 6000 Assay Kit of the Agilent Bioanalyzer 2100 system (Agilent Technologies, Santa Clara, CA, USA).

Sequencing libraries was created using NEB Next Ultra RNA Library Prep Kit following the manufacturer's instructions. The index codes were added to each sample. Briefly, the mRNA was purified from 3 µg total RNA of each of three replicate using poly-T oligo-attached magnetic beads and then broken into short fragments to synthesize first strand cDNA. The second strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H. PCR was carried out with Phusion High Fidelity DNA polymerase using universal PCR primers and index (X) primer. Finally, six paired-end cDNA libraries with an insert size of 150 bp were constructed for transcriptome sequencing and sequenced on HiSeqTM 2000 platform (Illumina Inc., San Diego, CA, USA).

4.4. Quality Check, Cleaning and de novo Assembly

The clean reads were retrieved after trimming adapter sequences, removal of low quality (containing >50% bases with a Phred quality score <10) and reads with unknown nucleotides (more than 5% ambiguous residues N) using the FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The high quality reads from all the nine libraries were de novo assembled into transcripts using the Trinity software (version r20140717, [62]) with the following parameters: -min-contig-length, 100-min-glue and 3–path-sing-distance-85-min-kmer-cov3. Next, the transcripts were realigned to construct unigenes and the software TGICL, version v2.1 [63] was employed to eliminate the redundant unigenes and to get the final unigene list based on the following parameters: -l, 40, -c, 10, -v, 25, -O, '-repeat_stringency, 0.95, -minmatch, 35 and -minscore 35'.

4.5. Functional Annotation of the Unigenes

The assembled unigenes were annotated by searching against various databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [64], Gene Ontology (GO) [65], Clusters of Orthologous Groups (COG) [66], Swissprot [67], NR [68], euKaryotic Orthologous Groups (KOG) [69] and NT using BLAST version:v2.2.26 [70] with a threshold of e-value $< 1 \times 10^{-5}$. Based on the functional annotation results, according to the database priorities of NR, SwissProt, KEGG and COG, we selected the best comparison fragment of unigenes as the CDS for the unigenes. For unigenes, which failed to return a hit, we then used ESTScan v3.0.2 [71] to make predictions using default parameters.

4.6. Analysis of Transcription Factors

To identify the gene encoding transcription factors (TF), we first collected the HMM profiles of known transcription factors from various databases (PlantTFDB [72], AnimalTFDB [73], FTFD [74] and DBD [75]) and the literature, and then used the hmmsearch of the HMMER package version 3.1b2 to compare the protein sequence to the HMM of known TFs, with the e-value $< 1 \times 10^{-5}$.

4.7. Gene Expression and Differential Expression Analysis

The clean reads were compared to the assembled unigenes using the tool Bowtie2 version 2.1.0 [76] and then the number of reads on each unigene was calculated using the RSEM version 1.2.21 [77]. The gene expression level was determined according to the fragments per kilobase of exon per million fragments mapped (FPKM).

The read count was normalized and DESeq2 tool version 1.4.5 [78] was used to detect the differential expressed genes (DEGs) between the control samples (PC) and the UV-C treated samples (PC6H and PC12H) with the fold change of >2 [79] and false discovery rate (FDR) correction set at $p < 0.01$. GO enrichment analysis was performed using the GOseq tool version 1.16.2 [80] with $p < 0.05$. KEGG pathway enrichment analysis of the DEGs was done using KOBAS2.0 [81] with FDR correction ($p < 0.05$).

4.8. Gene Expression Using Quantitative Real Time-PCR

The qRT-PCR was performed on RNA extracted from control and stressed leaf samples as described by Dossa et al. [82] using the *Actin* gene as the internal control. Specific primer pairs of 12 selected genes were designed using the Primer Premier 5.0 [83] (Table S4). The qRT-PCR was conducted on a Roche Lightcyler®480 instrument using the SYBR Green Master Mix (Vazyme), according to the manufacturer's protocol. Each reaction was performed using a 20 μL mixture containing 10 μL of 2× ChamQ SYBR qPCR Master Mix, 6 μL of nuclease-free water, 1 μL of each primer (10 mM) and 2 μL of four-fold diluted cDNA. All of the reactions were run in 96-well plates and each cDNA was analyzed in triplicate. The following cycling profile was used: 95 °C for 30 s, followed by 40 cycles of 95 °C/10 s, 60 °C/30 s. Data are presented as relative transcript level based on the $2^{-\Delta\Delta Ct}$ method [84].

5. Conclusions

In summary, we showed that UV-C treatments induced the accumulation of trans-resveratrol in *P. cuspidatum* leaves. Various UV-C treatments generated different levels of accumulation suggesting that there was room for optimization of the UV-C treatment protocol in order to yield maximum content of resveratrol. We further provided the putative genes participating in the resveratrol biosynthetic pathway and highlighted the key genes differentially expressed upon exposure to UV-C. It was evident that *P. cuspidatum* prioritized the resveratrol synthesis by strongly up-regulating the genes directly involved in this pathway and shutting down the expression of chalcone synthase genes. The results from this study provided insights into the mechanism of UV-C induced accumulation of resveratrol and probably its methylated forms in a species other than grape. It lays the foundation for further enhancement of stilbenes in *P. cuspidatum*.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/20/24/6185/s1. Figure S1 Gene ontology classification of all unigenes detected in this study; Figure S2 Correlation of the qRT-PCR expression profiles and the RNA-seq based on the selected genes. PC, PC6H and PC12H represent samples collected before, 6 h and 12 h after UV-C treatment, respectively; Table S1 List of the genes encoding enzymes involved in the resveratrol biosynthetic pathway; Table S2 Statistics of the read mapping results; Table S3 List of the differentially expressed transcription factors detected in this study; Table S4 The primer sequences of genes used for real time quantitative PCR.

Author Contributions: Conceptualization, H.Z., J.Q.; methodology, Z.L., J.X., X.W., Y.W.; validation, Y.L., D.W.; formal analysis, Z.L.; investigation, Z.L., J.X., X.W., Y.W.; resources, Y.L., D.W.; writing—original draft preparation, Z.L.; writing—review and editing, J.Q.; supervision, J.Q.; project administration, H.Z., J.Q.; funding acquisition, H.Z., J.Q.

Funding: This work was financially supported by the National Natural Science Foundation of China (Grant No. 81803670).

Conflicts of Interest: The authors declare no conflict of interest

References

1. Seiger, L. Mechanical control of Japanese knotweed (Fallopia japonica [Houtt.] Ronse Decraene): Effects of cutting regime on rhizomatous reserves. *Nat. Area. J.* **1997**, *17*, 341–345.
2. Bailey, J. Chapter 14. The rise and fall of Japanese knotweed? In *Invasive and Introduced Plants and Animals Human Perceptions, Attitudes and Approaches to Management*; Rotherham, I., Lambert, R., Eds.; Routledge: New York, NY, USA, 2013; pp. 221–232.

3. Jones, D.; Bruce, G.; Fowler, M.S.; Law-Cooper, R.; Graham, I.; Abel, A.; Street-Perrott, F.A.; Eastwood, D. Optimising physiochemical control of invasive Japanese knotweed. *Biol. Invasions* **2018**, *20*, 2091–2105. [CrossRef]
4. Yan, S.; Li, L.; Yu, S.; Xiao, P.G. Effect of Tabellae Polygonum Cuspidatum on Blood Lipids and Rheological Property in Rats. *China J. Chin. Mat. Med.* **1993**, *18*, 617–619.
5. Xiao, K.; Xuan, L.; Xu, Y.; Bai, D.; Zhong, D. Constituents from Polygonum cuspidatum. *Chem. Pharm. Bull. (Tokyo)* **2002**, *50*, 605–608. [CrossRef]
6. Peng, W.; Qin, R.; Li, X.; Zhou, H. Botany, phytochemistry, pharmacology, and potential application of Polygonum cuspidatum Sieb.et Zucc.: A review. *J. Ethnopharmacol.* **2013**, *148*, 729–745. [CrossRef]
7. Dong, J.; Wang, H.; Wan, L.; Hashi, Y.; Chen, S. Identification and determination of major constituents in Polygonum cuspidatum Sieb. et Zucc. by high performance liquid chromatography/electrospray ionization-ion traptime-of-flight mass spectrometry. *Se Pu* **2009**, *27*, 425–430.
8. Lachowicz, S.; Oszmiański, J. Profile of Bioactive Compounds in the Morphological Parts of Wild *Fallopia japonica* (Houtt) and *Fallopia sachalinensis* (F. Schmidt) and Their Antioxidative Activity. *Molecules* **2019**, *24*, 1436. [CrossRef]
9. Baur, J.A.; Sinclair, D.A. Therapeutic potential of resveratrol: The in vivo evidence. *Nat. Rev. Drug Discov.* **2006**, *5*, 493–506. [CrossRef]
10. Truong, V.L.; Jun, M.; Jeong, W.S. Role of resveratrol in regulation of cellular defense systems against oxidative stress. *BioFactors.* **2018**, *44*, 36–49. [CrossRef]
11. Renaud, S.; de Lorgeril, M. Wine, alcohol, platelets, and the French paradox for coronary heart disease. *Lancet.* **1992**, *339*, 1523–1526. [CrossRef]
12. Novelle, M.G..; Wahl, D.; Dieguez, C.; Bernier, M.; de Cabo, R. Resveratrol supplementation, where are we now and where should we go? *Ageing Res. Rev.* **2015**, *21*, 1–15. [CrossRef] [PubMed]
13. Nonomura, S.; Kanagawa, H.; Makimoto, A. Chemical constituents of polygonaceous plants. I. Studies on the components of Ko-jo-kon. *Yakugaku Zasshi.* **1963**, *83*, 988–990. [CrossRef] [PubMed]
14. Hao, D.; Ma, P.; Mu, J.; Chen, S.; Xiao, P.; Peng, Y.; Sun, C. De novo characterization of the root transcriptome of a traditional Chinese medicinal plant Polygonum cuspidatum. *Sci. China Life Sci.* **2012**, *55*, 452–466. [CrossRef] [PubMed]
15. Hasan, M.; Bae, H. An Overview of Stress-Induced Resveratrol Synthesis in Grapes: Perspectives for Resveratrol-Enriched Grape Products. *Molecules.* **2017**, *22*, 294. [CrossRef]
16. Jaillon, O.; Aury, J.M.; Noel, B.; Policriti, A.; Clepet, C.; Casagrande, A.; Choisne, N.; Aubourg, S.; Vitulo, N.; Jubin, C. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **2007**, *449*, U463–U465.
17. Halls, C.; Yu, O. Potential for metabolic engineering of resveratrol biosynthesis. *Trends Biotechnol.* **2008**, *26*, 77–81. [CrossRef]
18. Lanz, T.; Tropf, S.; Marner, F.J.; Schroder, J.; Schroder, G. The role of cysteines in polyketide synthases. Site-directed mutagenesis of resveratrol and chalcone synthases, two key enzymes in different plant-specific pathways. *J. Biol. Chem.* **1991**, *266*, 9971–9976.
19. Langcake, P.; Pryce, R.J. The production of resveratrol by Vitis vinifera and other members of the Vitaceae as a response to infection or injury. *Physiol. Plant Pathol.* **1976**, *9*, 77–86. [CrossRef]
20. Höll, J.; Vannozzi, A.; Czemmel, S.; D'Onofrio, C.; Walker, A.R.; Rausch, T.; Lucchin, M.; Boss, P.K.; Dry, I.B.; Bogs, J. The R2R3-MYB transcription factors MYB14 and MYB15 regulate stilbene biosynthesis in *Vitis vinifera*. *Plant Cell* **2013**, *25*, 4135–4149. [CrossRef]
21. Jiang, J.; Xi, H.; Dai, Z.; Lecourieux, F.; Yuan, L.; Liu, X.; Patra, B.; Wei, Y.; Li, S.; Wang, L. VvWRKY8 represses stilbene synthase genes through direct interaction with VvMYB14 to control resveratrol biosynthesis in grapevine. *J. Exp. Bot.* **2019**, *70*, 715–729. [CrossRef]
22. Langcake, P. Disease resistance of Vitis spp. and the production of the stress metabolites resveratrol, ″-viniferin, -viniferin, and pterostilbene. *Physiol. Plant Pathol.* **1981**, *18*, 213–226. [CrossRef]
23. Dercks, W.; Creasy, L.L. The significance of stilbene phytoalexins in the Plasmopara viticola-grapevine interaction. *Physiol. Mol. Plant Pathol.* **1989**, *34*, 189–202. [CrossRef]
24. Tassoni, A.; Fornale, S.; Franceschetti, M.; Musiani, F.; Michael, A.J.; Perry, B. Jasmonates and Na-orthovanadate promote resveratrol production in Vitis vinifera cv. Barbera cell cultures. *New Phytol.* **2005**, *166*, 895–905. [CrossRef] [PubMed]

25. Wen, P.F.; Chen, J.Y.; Kong, W.F.; Pan, Q.H.; Wan, S.B.; Huang, W.D. Salicylic acid induced the expression of phenylalanine ammonia-lyase gene in grape berry. *Plant Sci.* **2005**, *169*, 928–934. [CrossRef]
26. González-Barrio, R.; Beltrán, D.; Cantos, E.; Gil, M.I.; Espín, J.C.; Tomás-Barberán, F.A. Comparison of ozone and UV-C treatments on the postharvest stilbenoid monomer, dimer, and trimer induction in var. 'Superior' white table grapes. *J. Agric. Food Chem.* **2006**, *54*, 4222–4228. [CrossRef] [PubMed]
27. Schmidlin, L.; Poutaraud, A.; Claudel, P.; Mestre, P.; Prado, E.; Santos-Rosa, M. A stress-inducible resveratrol O-methyltransferase involved in the biosynthesis of pterostilbene in grapevine. *Plant Physiol.* **2008**, *148*, 1630–1639. [CrossRef] [PubMed]
28. Wang, L.J.; Ma, L.; Xi, H.F.; Duan, W.; Wang, J.F.; Li, S.H. Individual and combined effects of $CaCl_2$ and UV-C on the biosynthesis of resveratrols in grape leaves and berry skins. *J. Agric. Food Chem.* **2013**, *61*, 7135–7141. [CrossRef]
29. Banerjee, N.; Zhang, M.Q. Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.* **2003**, *31*, 7024–7031. [CrossRef]
30. Hollósy, F. Effects of ultraviolet radiation on plant cells. *Micron* **2002**, *33*, 179–197. [CrossRef]
31. Kasim, M.U.; Kasim, R.; Erkal, S. UV-C treatments on fresh-cut green onions enhanced antioxidant activity, maintained green color and controlled 'telescoping'. *J. Food Agric. Environ.* **2008**, *6*, 63–67.
32. Rivera-Pastrana, D.M.; Gardea, A.A.; Yahia, E.M.; Martínez-Téllez, M.A.; González-Aguilar, G.A. Effect of UV-C irradiation and low temperature storage on bioactive compounds, antioxidant enzymes and radical scavenging activity of papaya fruit. *J. Food Sci. Technol.* **2014**, *51*, 3821–3829. [CrossRef] [PubMed]
33. Crupi, P.; Pichierri, A.; Basile, T.; Antonacci, D. Postharvest stilbenes and flavonoids enrichment of table grape cv Redglobe (*Vitis vinifera* L.) as affected by interactive UV-C exposure and storage conditions. *Food Chem.* **2013**, *141*, 802–808. [CrossRef] [PubMed]
34. Wang, W.; Tang, K.; Yang, H.R.; Wen, P.F.; Zhang, P.; Wang, H.L.; Huang, W.D. Distribution of resveratrol and stilbene synthase in young grape plants (*Vitis vinifera* L. cv. Cabernet Sauvignon) and the effect of UV-C on its accumulation. *Plant Physiol. Biochem.* **2010**, *48*, 142–152. [CrossRef] [PubMed]
35. Wang, J.F.; Ma, L.; Xi, H.F.; Wang, L.J.; Li, S.H. Resveratrol synthesis under natural conditions and after UV-C irradiation in berry skin is associated with berry development stages in 'Beihong' (V. vinifera × V. amurensis). *Food Chem.* **2015**, *168*, 430–438. [CrossRef] [PubMed]
36. Deng, N.; Liu, C.; Chang, E.; Ji, J.; Yao, X.; Yue, J.; Bartish, I.V.; Chen, L.; Jiang, Z.; Shi, S. High temperature and UV-C treatments affect stilbenoid accumulation and related gene expression levels in Gnetum parvifolium. *Electron. J. Biotechnol.* **2017**, *25*, 43–49. [CrossRef]
37. Cantos, E.; Espin, J.C.; Fernandez, M.J.; Oliva, J.; Tomas-Barberan, F.A. Postharvest UV-C-irradiated grapes as a potential source for producing stilbene-enriched red wines. *J. Agric. Food Chem.* **2003**, *51*, 1208–1214. [CrossRef]
38. Guerrero, R.F.; Puertas, B.; Fernández, M.I.; Palma, M.; Cantos-Villar, E. Induction of stilbenes in grapes by UV-C: Comparison of different subspecies of Vitis. *Innov. Food Sci. Emerg. Technol.* **2010**, *11*, 231–238. [CrossRef]
39. Cantos, E.; Espin, J.C.; Tomas-Barberan, F.A. Postharvest stilbene-enrichment of red and white table grape varieties using UV-C irradiation pulses. *J. Agric. Food Chem.* **2002**, *50*, 6322–6329. [CrossRef]
40. Yin, X.; Singer, S.D.; Qiao, H.; Liu, Y.; Jiao, C.; Wang, H.; Li, Z.; Fei, Z.; Wang, Y.; Fan, C. Insights into the Mechanisms Underlying Ultraviolet-C Induced Resveratrol Metabolism in Grapevine (V. amurensis Rupr.) cv. "Tonghua-3". *Front Plant Sci.* **2016**, *7*, 503. [CrossRef]
41. Xi, H.; Ma, L.; Liu, G.; Wang, N.; Wang, J.; Wang, L. Transcriptomic analysis of grape (*Vitis vinifera* L.) leaves after exposure to ultraviolet C irradiation. *PLoS One* **2014**, *9*, e113772. [CrossRef]
42. Suzuki, M.; Nakabayashi, R.; Ogata, Y.; Sakurai, N.; Tokimatsu, T.; Goto, S. Multi omics in grape berry skin revealed specific induction of stilbene synthetic pathway by UV-C irradiation. *Plant Physiol.* **2015**, *168*, 47–59. [CrossRef] [PubMed]
43. Zhou, S.; Liu, P.; Chen, J.; Du, G.; Li, H.; Zhou, J. Characterization of mutants of a tyrosine ammonia-lyase from Rhodotorula glutinis. *Appl. Microbiol. Biotechnol.* **2016**, *100*, 10443–10452. [CrossRef] [PubMed]
44. Walle, T.F.; Hsieh, M.H.; DeLegge, J.E.; Oatis, U.K., Jr. Walle High absorption but very low bioavailability of oral resveratrol in humans Drug. *Metab. Dispos.* **2004**, *32*, 1377–1382. [CrossRef] [PubMed]

45. Jeong, Y.J.; An, C.H.; Woo, S.G.; Jeong, H.J.; Kim, Y.M.; Park, S.J.; Yoon, B.D.; Kim, C.Y. Production of pinostilbene compounds by the expression of resveratrol O-methyltransferase genes in Escherichia coli. *Enzyme Microb. Technol.* **2014**, *54*, 4–8. [CrossRef] [PubMed]
46. Martínez-Márquez, A.; Morante-Carriel, J.A.; Ramírez-Estrada, K.; Cusidó, R.M.; Palazon, J.; Bru-Martínez, R. Production of highly bioactive resveratrol analogues pterostilbene and piceatannol in metabolically engineered grapevine cell cultures. *Plant Biotechnol. J.* **2016**, *14*, 1813–1825. [CrossRef] [PubMed]
47. Gonzalez, A.; Zhao, M.; Leavitt, J.M.; Lloyd, A.M. Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in Arabidopsis seedlings. *Plant J.* **2008**, *53*, 814–827. [CrossRef] [PubMed]
48. Lloyd, A.; Brockman, A.; Aguirre, L.; Campbell, A.; Bean, A.; Cantero, A.; Gonzalez, A. Advances in the MYB–bHLH–WD repeat (MBW) pigment regulatory model: Addition of a WRKY factor and co-option of an anthocyanin MYB for betalain regulation. *Plant Cell Physiol.* **2017**, *58*, 1431–1441. [CrossRef]
49. Fang, L.; Hou, Y.; Wang, L.; Xin, H.; Wang, N.; Li, S. Myb14, a direct activator of STS, is associated with resveratrol content variation in berry skin in two grape cultivars. *Plant Cell Rep.* **2014**, *33*, 1629–1640. [CrossRef]
50. Childs, K.L.; Davidson, R.M.; Buell, C.R. Gene coexpression network analysis as a source of functional annotation for rice genes. *PLoS ONE* **2011**, *6*, e22196. [CrossRef]
51. Weston, D.J.; Karve, A.A.; Gunter, L.E.; Jawdy, S.S.; Yang, X.; Allen, S.M.; Wullschleger, S.D. Comparative physiology and transcriptional networks underlying the heat shock response in *Populus trichocarpa*, *Arabidopsis thaliana* and *Glycine max*. *Plant Cell Environ.* **2011**, *34*, 1488–1506.
52. Downs, G.S.; Bi, Y.M.; Colasanti, J.; Wu, W.; Chen, X.; Zhu, T.; Rothstein, S.J.; Lukens, L.N. A developmental transcriptional network for *Zea mays* defines coexpression modules. *Plant Physiol.* **2013**, *161*, 1830–1843. [CrossRef] [PubMed]
53. Ma, S.; Ding, Z.; Li, P. Maize network analysis revealed gene modules involved in development, nutrients utilization, metabolism, and stress response. *BMC Plant Biol.* **2017**, *17*, 131. [CrossRef] [PubMed]
54. Shen, P.; Hour, A.; Liu, L.D. Microarray meta-analysis to explore abiotic stress-specific gene expression patterns in Arabidopsis. *Botanical Studies* **2017**, *58*, 22. [CrossRef] [PubMed]
55. Shahan, R.; Zawora, C.; Wight, H.; Sittmann, J.; Wang, W.; Mount, S.M.; Liu, Z. Consensus coexpression network analysis identifies key regulators of flower and fruit development in wild strawberry. *Plant Physiol.* **2018**, *178*, 202–216. [CrossRef] [PubMed]
56. Dossa, K.; Mmadi, M.A.; Zhou, R.; Zhang, T.; Su, R.; Zhang, Y.; Wang, L.; You, J.; Zhang, X. Depicting the core transcriptome modulating multiple abiotic stresses responses in sesame (*Sesamum indicum* L.). *Int. J. Mol. Sci.* **2019**, *20*, 3930. [CrossRef] [PubMed]
57. Corso, M.; Vannozzi, A.; Maza, E.; Vitulo, N.; Meggio, F.; Pitacco, A.; Telatin, A.; D'Angelo, M.; Feltrin, E.; Negri, A.S.; et al. Comprehensive transcript profiling of two grapevine rootstock genotypes contrasting in drought susceptibility links the phenylpropanoid pathway to enhanced tolerance. *J. Exp. Bot.* **2015**, *66*, 5739–5752. [CrossRef] [PubMed]
58. Wong, D.C.J.; Schlechter, R.; Vannozzi, A.; Höll, J.; Hmmam, I.; Bogs, J.; Tornielli, G.B.; Castellarin, S.D.; Matus, J.T. A systems-oriented analysis of the grapevine R2R3-MYB transcription factor family uncovers new insights into the regulation of stilbene accumulation. *DNA Res.* **2016**, *23*, 451–466. [CrossRef]
59. Vannozzi, A.; Wong, D.C.J.; Höll, J.; Hmmam, I.; Matus, J.T.; Bogs, J.; Ziegler, T.; Dry, I.; Barcaccia, G.; Lucchin, M. Combinatorial Regulation of Stilbene Synthase Genes by WRKY and MYB Transcription Factors in Grapevine (*Vitis vinifera* L.). *Plant Cell Physiol.* **2018**, *59*, 1043–1059. [CrossRef]
60. Xi, H.F.; Ma, L.; Wang, L.N.; Li, S.H.; Wang, L.J. Differential response of the biosynthesis of resveratrols and flavonoids to UV-C irradiation in grape leaves. *New Zealand J. Crop Hortic. Sci.* **2015**, *43*, 163–172. [CrossRef]
61. Zhang, S.; Ying, H.; Pingcuo, G.; Wang, S.; Zhao, F.; Cui, Y.; Shi, J.; Zeng, H.; Zeng, X. Identification of Potential Metabolites Mediating Bird's Selective Feeding on *Prunus mira* Flowers. *Biomed Res. Int.* **2019**, *2019*, 8. [CrossRef]
62. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full length transcriptome assembly from RNA Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [CrossRef] [PubMed]

63. Pertea, G. , Huang, X.; Liang, F.; Antonescu, V.; Sultana, R.; Karamycheva, S.; Lee, Y.; White, J.; Cheung, F.; Parvizi, B.; Tsai, J.; Quackenbush, J. TIGR Gene Indices clustering tools (TGICL): A software system for fast clustering of large EST datasets. *Bioinformatics* **2003**, *19*, 651–652. [CrossRef] [PubMed]
64. Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **2004**, *2004. 32*, D277–D280. [CrossRef]
65. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef] [PubMed]
66. Tatusov, R.L.; Galperin, M.Y.; Natale, D.A. The COG database: A tool for genome scale analysis of protein functions and evolution. *Nucleic Acids Res.* **2000**, *28*, 33–36. [CrossRef]
67. Apweiler, R.; Bairoch, A.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; et al. UniProt: The Universal Protein knowledgebase. *Nucleic Acids Res. 2004. 32*, D115–D119. [CrossRef]
68. Deng, Y.Y.; Li, J.Q.; Wu, S.F.; Zhu, Y.P.; Chen, Y.W.; He, F.C. Integrated nr Database in Protein Annotation System and Its Localization. *Comput. Eng.* **2006**, *32*, 71–74.
69. Koonin, E.V.; Fedorova, N.D.; Jackson, J.D.; Jacobs, A.R.; Krylov, D.M.; Makarova, K.S.; Mazumder, R.; Mekhedov, S.L.; Nikolskaya, A.N.; Rao, B.S.; et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **2004**, *5*, R7. [CrossRef]
70. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *1997. 25*, 3389–3402. [CrossRef]
71. Iseli, C.; Jongeneel, C.V.; Bucher, P. ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1999**, 138–148.
72. Jin, J.; Zhang, H.; Kong, L.; Gao, G.; Luo, J. PlantTFDB 3.0: A portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res.* **2014**, *42*, D1182–D1187. [CrossRef] [PubMed]
73. Zhang, H.; Liu, T.; Liu, C.J.; Song, S.; Zhang, X.; Liu, W.; Jia, H.; Xue, Y.; Guo, A.Y. AnimalTFDB 2.0: A resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res.* **2015**, *43*, D76–D81. [CrossRef] [PubMed]
74. Park, J.; Park, J.; Jang, S.; Kim, S.; Kong, S.; Choi, J.; Ahn, K.; Kim, J.; Lee, S.; Kim, S. FTFD: An informatics pipeline supporting phylogenomic analysis of fungal transcription factors. *Bioinformatics.* **2008**, *24*, 1024–1025. [CrossRef] [PubMed]
75. Kummerfeld, S.K.; Teichmann, S.A. DBD: A transcription factor prediction database. *Nucleic Acids Res.* **2006**, *34*, D74–D81. [CrossRef] [PubMed]
76. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **2009**, *10*, R25. [CrossRef] [PubMed]
77. Li, B.; Colin, N.D. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [CrossRef] [PubMed]
78. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef]
79. Anders, S.; McCarthy, D.J.; Chen, Y.; Okoniewski, M.; Smyth, G.K.; Huber, W.; Robinson, M.D. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.* **2013**, *8*, 1765–1786. [CrossRef]
80. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biol.* **2010**, *11*, R14.
81. Xie, C.; Mao, X.; Huang, J.; Ding, Y.; Wu, J.; Dong, S.; Kong, L.; Gao, G.; Li, C.; Wei, L. KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* **2011**, *39*, W316–W322. [CrossRef]
82. Dossa, K.D.; Li, D.J.; Yu, J.L.; Wang, Y.; Zhang, J.; You, R.; Zhou, M.A.; Mmadi, A.; Li, D.; Fonceka, D.; et al. The genetic basis of drought tolerance in the high oil crop *Sesamum indicum*. *Plant Biotechnol. J.* **2019**, 1–16. [CrossRef] [PubMed]

83. Lalitha, S. Primer premier 5. *Biotech Softw. Internet Rep.* **2000**, *1*, 270–272. [CrossRef]
84. Livak, K.J.; Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* **2001**, *25*, 402–408. [CrossRef] [PubMed]

International Journal of
Molecular Sciences

Article

Pairing and Exchanging between *Daypyrum villosum* Chromosomes 6V#2 and 6V#4 in the Hybrids of Two Different Wheat Alien Substitution Lines

Xiaolan Ma [1], Zhiying Xu [1,2], Jing Wang [1], Haiqiang Chen [1], Xingguo Ye [1,3] and Zhishan Lin [1,3,*]

[1] Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China; 13264529812@163.com (X.M.); 13414916566@163.com (Z.X.); 13121260899@163.com (J.W.); haiqiang_chen@163.com (H.C.); yexingguo@caas.cn (X.Y.)

[2] Agricultural College, Guangdong Ocean University, Zhanjiang 524088, China

[3] National Key Facility of Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, Beijing 100081, China

* Correspondence: linzhishan@caas.cn

Received: 5 October 2019; Accepted: 27 November 2019; Published: 1 December 2019

Abstract: Normal pairing and exchanging is an important basis to evaluate the genetic relationship between homologous chromosomes in a wheat background. The pairing behavior between 6V#2 and 6V#4, two chromosomes from different *Dasypyrum villosum* accessions, is still not clear. In this study, two wheat alien substitution lines, 6V#2 (6A) and 6V#4 (6D), were crossed to obtain the F_1 hybrids and F_2 segregating populations, and the testcross populations were obtained by using the F_1 as a parent crossed with wheat variety Wan7107. The chromosomal behavior at meiosis in pollen mother cells (PMCs) of the F_1 hybrids was observed using a genomic in situ hybridization (GISH) technique. Exchange events of two alien chromosomes were investigated in the F_2 populations using nine polymerase chain reaction (PCR) markers located on the 6V short arm. The results showed that the two alien chromosomes could pair with each other to form ring- or rod-shaped bivalent chromosomes in 79.76% of the total PMCs, and most were pulled to two poles evenly at anaphase I. Investigation of the F_2 populations showed that the segregation ratios of seven markers were consistent with the theoretical values 3:1 or 1:2:1, and recombinants among markers were detected. A genetic linkage map of nine PCR markers for 6VS was accordingly constructed based on the exchange frequencies and compared with the physical maps of wheat and barley based on homologous sequences of the markers, which showed that conservation of sequence order compared to 6V was 6H and 6B > 6A > 6D. In the testcross populations with 482 plants, seven showed susceptibility to powdery mildew (PM) and lacked amplification of alien chromosomal bands. Six other plants had amplification of specific bands of both the alien chromosomes at multiple sites, which suggested that the alien chromosomes had abnormal separation behavior in about 1.5% of the PMCs in F_1, which resulted in some gametes containing two alien chromosomes. In addition, three new types of chromosome substitution were developed. This study lays a foundation for alien allelism tests and further assessment of the genetic relationship among 6V#2, 6V#4, and their wheat homoeologous chromosomes.

Keywords: wheat; *Dasypyrum villosum*; alien substitution line; GISH; molecular marker; marker-assisted selection

1. Introduction

Wheat (*Triticum aestivum* L. $2n = 6x = 42$, AABBDD) is one of the most widely cultivated crops in the world. Biotic and abiotic stresses often cause different degrees of decline in wheat yield and quality. There are many genes in wheat relatives with desirable traits, which have great potential to improve the resistance of common wheat to various stresses. For example, *Dasypyrum villosum* ($2n = 2x = 14$, VV), which originated in the Mediterranean and the Caucasus region [1,2], is a cross-pollinating annual plant that retained many characteristics that ordinary wheat lacks, such as resistances to powdery mildew (PM), stem rust, leaf rust, leaf blight, and scattered smut [1,2]. *D. villosum* can, therefore, be used as a potential resistance source for wheat breeding [3]. The resistance of the tertiary gene pool of wheat is also non-host resistance [4], which is usually wide spectrum and high efficiency. *Pm21*, a resistance gene to PM from *D. villosum*, was introgressed into common wheat using chromosome translocation, which was widely used in wheat production and became one of the most effective genetic loci introgressed into wheat from its wild species. Commercial wheat varieties carrying *Pm21* are widely applied in China, with an accumulative planting area of more than four million hectares [5]. In addition, *D. villosum* also contains many other excellent traits such as cold tolerance, strong tillering ability, dense spikes with many flowers, and high crude protein content in grain [6,7]. Currently, more than 300 *D. villosum* accessions were collected, and four of them were introgressed into wheat as additional lines, substitution lines, or translocation lines [8–11].

Chromosomal disproportionation in different *D. villosum* accessions can cause homologous chromosomes to be unpaired in their hybrids in a wheat background such as 6V#1 and 6V#2 [12]. Chromosome arms 6V#2S and 6V#4S can pair with each other only at a lower frequency in their F_1 hybrids between 6V#2S·6AL and 6V#4S·6DL translocation lines by cytological observation [10]. In order to rule out the possible effect on the pairing behavior of the alien arms that might be caused by the different dynamics of 6AL and 6DL, it is necessary to investigate the pairing behavior between chromosomes 6V#2 and 6V#4 in the non-translocation state. Chromosome 6V#2 carries *Pm21*, which encodes a typical coiled-coil - nucleotide-binding site - leucine-rich repeat (CC–NBS–LRR) protein that confers broad-spectrum resistance to PM [5,13]. Chromosome 6V#4 carries *PmV*, which is considered an allele of *Pm21* according to data on the National Center for Biotechnology Information (NCBI) website, but its allelism needs to be confirmed. Pairing and exchange between the two chromosomes are prerequisites for the allelism test. PM resistance is the only phenotypic trait that can be tracked in a wheat background; most of the plants showed resistance in the F_2 and testcross populations, but they cannot be distinguished from each other by phenotype.

A high collinear relationship between gramineous plant genomes was proven using comparative genomics strategies [14]. Specific markers to each chromosome of *D. villosum* were obtained using a comparative genomics method [10,15–21]. Some markers show polymorphisms in amplified length and presence or absence of specific amplified bands for the chromosomes of different *D. villosum* accessions. It should be possible to understand the affinity of the homologous chromosomes from different *D. villosum* accessions by investigating whether the exchange occurs among these markers in the F_2 populations. At the same time, the conservation of marker ranks on the homoeologous chromosomes can be well defined by comparing the genetic linkage map of the alien chromosome with the physical map of its homoeologous chromosomes in wheat and barley. These results will provide some important information for understanding the phylogenetic status of *D. villosum* in the grass family and the selection of breeding strategies to induce homoeologous chromosomal translocations between alien species and wheat.

In this study, the pairing behavior of two alien chromosomes, 6V#2 and 6V#4, was investigated in the F_1 hybrids derived from a cross between substitution lines 6V#2 (6A) and 6V#4 (6D) by genomic in situ hybridization (GISH). The genetic linkage map was drawn according to the exchange frequencies of the nine polymorphic PCR markers between the two alien chromosomes in the F_2 populations. Furthermore, the physical maps of homologous sequences of the nine markers in wheat and barley were compared. Combined with the molecular marker-assisted selection (MAS) of wheat chromosomes 6A and 6D, three new types of substitution lines were identified, which laid a foundation for further screening of novel resistance to PM wheat–*D. villosum* translocation lines between the alien chromosomes and different wheat chromosomes in the sixth homoeologous group, as well as evaluating their genetic effects.

2. Results

2.1. Pairing Behavior of 6V#2 and 6V#4 Chromosomes in Their F_1 Hybrids

In total, 107 pollen mother cells (PMCs) at meiosis in the F_1 hybrids of 6V#2 (6A) and 6V#4 (6D) substitution lines were observed. Based on the observation of the green fluorescent signal, two alien chromosomes paired with each other at diakinesis (Figure 1a) and metaphase I (Figure 1b,c) in most of the PMCs. Here, 6V#2 paired with 6V#4 to form a rod bivalent (64.18% of the bivalent chromosomes) (Figure 1b) or a ring bivalent (35.82% of the bivalent chromosomes) (Figure 1c) in 67 PMCs (79.76%). A few of the alien bivalent rings separated in advance (Figure 1d). In 17 PMCs (20.24%), the alien chromosomes did not pair, and presented as two univalent chromosomes (Figure 1e,f). At anaphase I, the alien chromosomes were evenly pulled to the two poles in most PMCs (Figure 1g,h). At metaphase I, most of the PMCs showed a trivalent and a univalent wheat chromosome in the same cell (Figure 1b,e,f). Because 6A and 6D chromosomes lacked a pairing partner in the F_1 hybrids, they should be present as two univalents, suggesting that there is a translocation involving 6D or 6A in wheat. At anaphase I, chromatin bridges were correspondingly observed in some PMCs (Figure 1g).

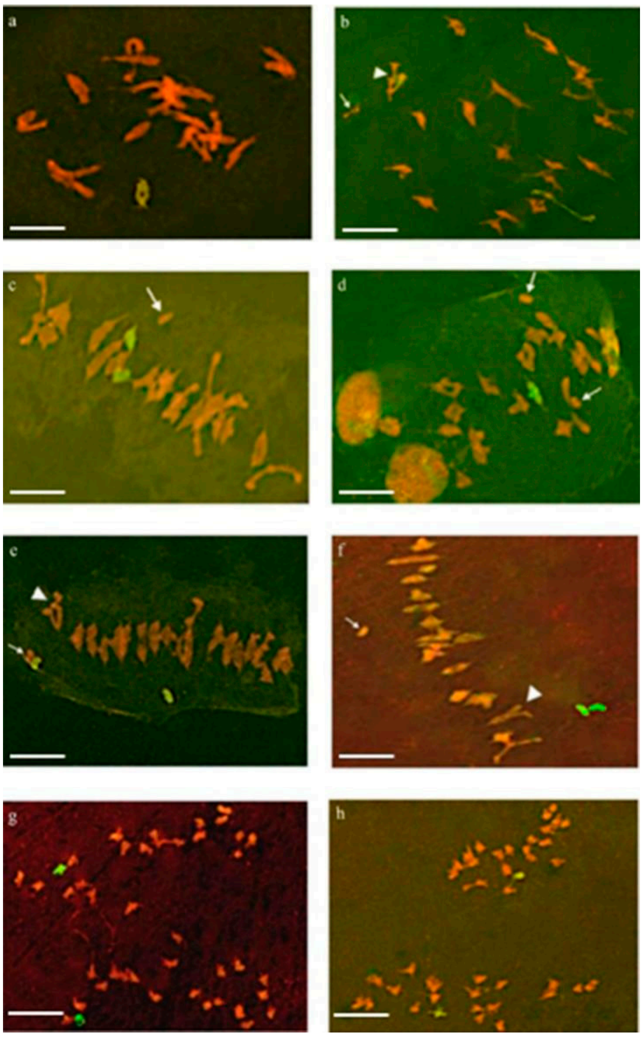


Figure 1. Chromosome pairing and separating behaviors in pollen mother cells (PMCs) of the F_1 hybrids between 6V#2 (6A) and 6V#4 (6D) substitution lines as revealed by genomic in situ hybridization (GISH) analysis at meiosis. Red indicates wheat chromosomes counterstained with propidium iodide (PI); green indicates alien chromosomes. The triangular arrow indicates heteromorphic bivalent chromosomes, and the arrow indicates wheat univalent chromosomes. (**a**) Alien chromosomes paired

with each other at diakinesis. (**b**) Alien chromosomes paired with each other to form bivalent rods at metaphase I. (**c**) Premature separation of alien chromosomes that formed bivalent rings. (**d**) Alien chromosomes paired with each other to form bivalent rings at metaphase I. (**e**,**f**) Unpaired univalent chromosomes 6V#2 and 6V#4. (**g**,**h**) Isolated homologous chromosomes 6V#2 and 6V#4 at anaphase I; wheat chromatin bridges can be seen. Bar = 10 μm.

2.2. Detection of Plant Genotypes by Nine 6VS-Specific Molecular Markers in the F_2 Populations

A total of 323 individual plants were randomly selected from the F_2 populations derived from the cross between the two substitution lines, and their genomic DNA samples were amplified by the nine 6V#2/6V#4-polymorphic PCR markers. The amplification of portions of the plants is shown in Figure 2; their amplified bands are listed in Table 1.

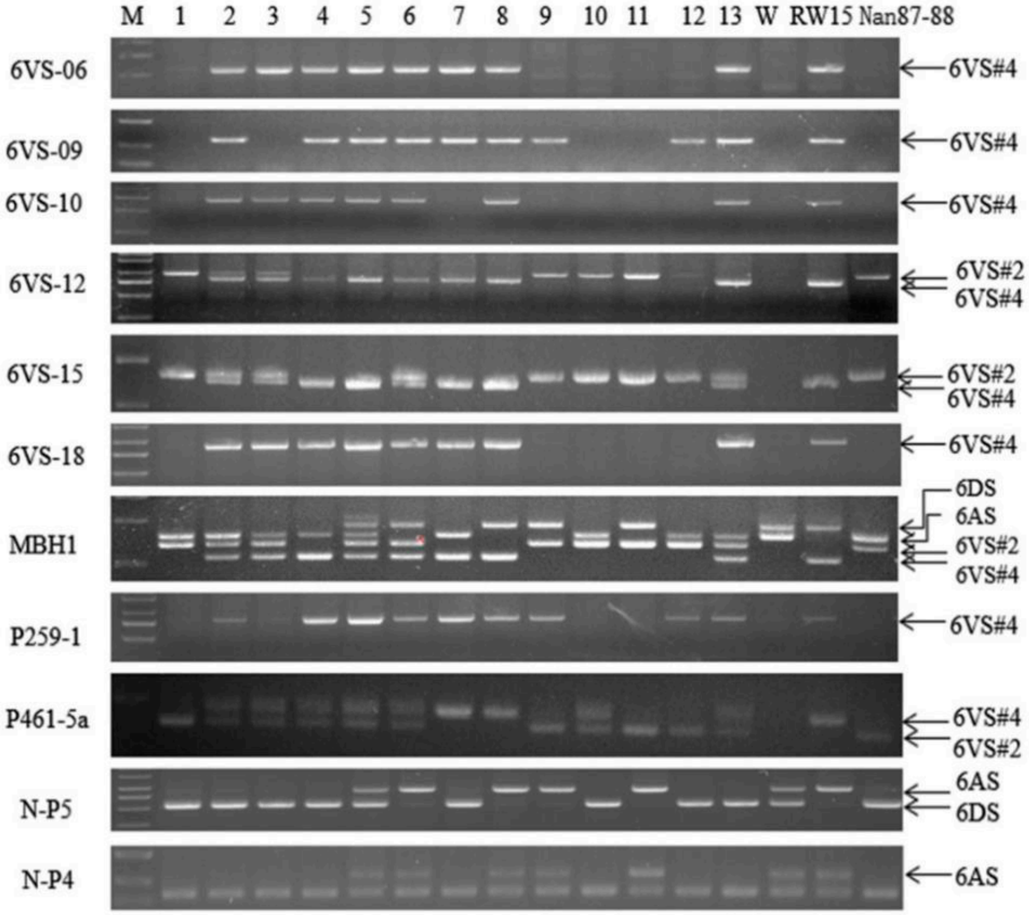


Figure 2. Amplification of portions of individual plants in F_2 hybridization populations derived from two substitution lines with different markers. Samples 1–13 are individual F_2 plants in 18GL101-11 from a cross of two substitution lines. M: DL2000+; W: Wan7107; RW15: substitution line 6VS#4 (6D); Nan87-88: substitution line 6VS#2 (6A).

Among all selected plants, 83 plants had amplified bands identical to one of the parental types by the nine markers, including 30 with 6V#2- and 53 with 6V#4-specific bands (sample 1 and sample 8 in Figure 2 and Table 1). A total of 239 plants showed heterozygotic genotypes, which had both 6V#2- and 6V#4-specific bands in all or some of the codominant marker loci (samples 2–6, 10, and 13 in Figure 2 and Table 1), or recombinant genotypes with 6V#2-specific bands in some loci while having 6V#4-specific bands in other loci (samples 7, 9, and 12 in Figure 2 and Table 1). In one plant, neither 6V#2 nor 6V#4 chromosomes were detected.

Table 1. Genotypes of portions of the F_2 plants from the crossing of two substitution lines amplified by different markers.

Materials	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**Wan 7107**	**RW15**	**Nan 87–88**
Markers																
6VS-06	-	4	4	4	4	4	4	4	-	-	-	-	4	-	4	-
6VS-09	-	4	-	4	4	4	4	4	4	-	-	4	4	-	4	-
6VS-10	-	4	4	4	4	4	-	4	-	-	-	-	4	-	4	-
6VS-12	2	2, 4	2, 4	4	4	4	4	4	2	2	2	-	4	-	4	2
6VS-15	2	2, 4	2, 4	4	4	2, 4	4	4	2	2	2	2	2, 4	-	4	2
6VS-18	-	4	4	4	4	4	4	4	-	-	-	-	4	-	4	-
MBH1	2	2, 4	2, 4	4	2, 4	2, 4	4	4	2	2	2	2	2, 4	-	4	2
P259-1	-	4	-	4	4	4	4	4	4	-	-	4	4	-	4	-
P461-5a	2	2, 4	2, 4	2, 4	2, 4	2, 4	4	4	2	2,4	2	2	2, 4	-	4	2
N-P4	-	-	-	-	6A	6A	-	6A	6A	-	6A	-	-	6A	6A	-
N-P5	6D	6D	6D	6D	6A, 6D	-	6D	-	-	6D	-	6D	6D	6A, 6D	6A	6D

Samples 1–13 are 13 individual F_2 plants of 18GL101-11 from the crossing of two substitution lines; "2" stands for specific band amplified from alien chromosome 6V#2S; "4" stands for specific band amplified from alien chromosome 6V#4S; "2,4" stands for specific bands amplified from alien chromosome 6V#2S and 6V#4S; "6A" stands for specific band amplified from wheat chromosome 6A; "6D" stands for specific band amplified from wheat chromosome 6D; "6A,6D" stands for specific bands amplified from wheat chromosomes 6A and 6D; "-" indicates the plant without the corresponding amplified band.

A chi-square test was performed on the separation ratios of the nine markers in the F_2 populations. The chi-square values of four dominant markers, 6VS-06, 6VS-10, 6VS-18, and P259-1, were all less than the X^2 value 6.635 of 3:1 at the 0.01 level (Table 2), consistent with the separation ratio of a Mendelian allele. The chi-square value of 6VS-09 was much larger than 6.635, indicating its deviation from the Mendelian ratio. The chi-square values of the codominant polymorphic marker 6VS-12 were higher than 9.210 at the 0.01 level (Table 2), indicating that the actual observed value was inconsistent with the theoretical value of 1:2:1. However, the chi-square values of MBH1, 6VS-15, and P461-5a were less than 9.210 at 0.01 level, which are consistent with the theoretical ratio of 1:2:1.

Table 2. Chi-square test with nine 6V#2S/6V#4S specific markers in the F_2 populations derived from the cross of two alien substitution lines.

Markers	**Total**	**#2$^+$**	**#4$^+$**	**#2$^+$#4$^+$**	**Chi-Square Value** $X^2_{3:1} = 6.635$, $X^2_{1:2:1} = 9.210$, $p = 0.01$
6VS-06	323	-	233	-	$1.41 < X^2_{3:1}$
6VS-09	323	-	198	-	$32.33 > X^2_{3:1}$
6VS-10	323	-	228	-	$3.35 < X^2_{3:1}$
6VS-18	323	-	229	-	$2.90 < X^2_{3:1}$
P259-1	323	-	236	-	$0.64 < X^2_{3:1}$
MBH1	316	84	79	153	$0.47 < X^2_{1:2:1}$
6VS-12	307	55	103	149	$15.27 > X^2_{1:2:1}$
6VS-15	125	36	30	59	$0.97 < X^2_{1:2:1}$
P461-5a	308	77	94	137	$5.63 < X^2_{1:2:1}$

2.3. *Construction of the 6VS Genetic Linkage Map and Collinearity Analysis*

To further determine the linear order of the nine markers on chromosome 6VS, the mapping software QTL IciMapping was used with an LOD value of 3.0. The genetic linkage map was drawn with a genetic recombination rate as the distance between the markers. The map spans a total distance of 142.58 cM (Figure 3).

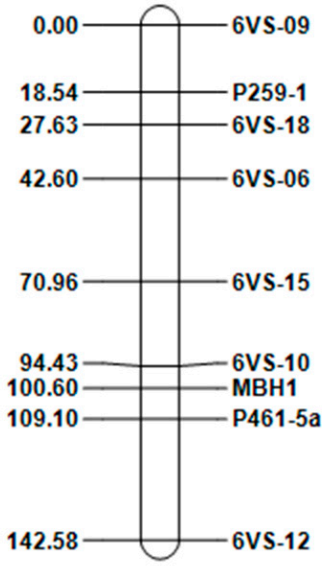


Figure 3. Genetic linkage map constructed based on the exchange frequencies of nine markers in the F_2 populations of two substitution lines.

Sequences of the nine markers were BLASTed on the website http://plants.ensembl.org/index.html to obtain the physical location information of the homologous sequences of the homoeologous group 6 of wheat and barley, and physical maps were drawn. The distribution of the nine markers on chromosomes of wheat homoeologous group 6 and barley 6H is shown in Figure 4. Compared with the physical maps of barley 6H and wheat 6A, 6B, and 6D, out of the nine homologous sequences, there were eight located on 6HS and 6BS, and six on 6AS and 6DS, while the others were found on the long arms of the corresponding chromosomes. The genetic map of the nine markers had the highest similarity in order on 6H and 6B, while the similarity in order on the three wheat chromosomes was 6B > 6A > 6D.

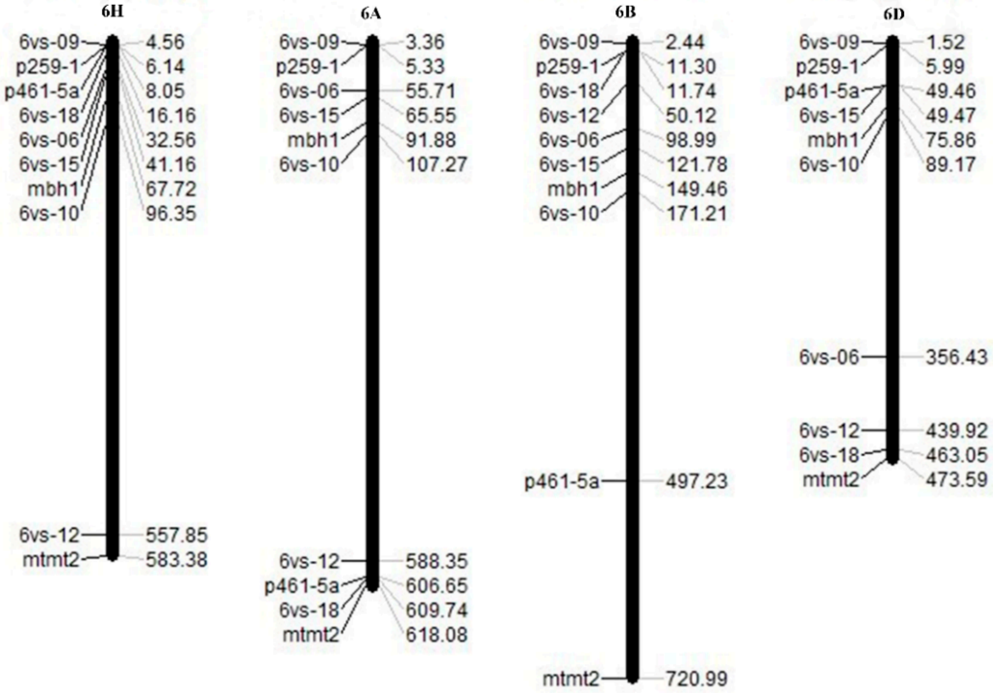


Figure 4. Physical maps of wheat 6A, 6B, and 6D and barley 6H based on homologous sequences of nine markers on 6VS.

2.4. Development of 6A- and 6D-Specific Molecular Markers and Detection of 6A and 6D Chromosomes in the F_2 Populations

For developing 6A- and 6D-specific molecular markers, 197 pairs of primers were designed. One pair of primers, N-P4 (Table 3), was amplified in CSN6BT6A, CSN6BT6D, CSN6DT6A, and CSN6DT6B, but not CSN6AT6B and CSN6AT6D (Figure 5a), indicating that it is a 6A-specific marker. Another pair of primers, N-P5 (Table 3), was amplified two bands in CSN6BT6A and CSN6BT6D, the larger one corresponding to the band amplified in CSN6AT6B and CSN6AT6D, and the smaller one corresponding to the band amplified in CSN6DT6A and CSN6DT6B (Figure 5c); therefore, it should be a codominant polymorphic marker of 6A and 6D. The two translocation lines 6V#2S·6AL and 6V#4S·6DL, lacking 6AS and 6DS, respectively, were amplified by N-P4 and N-P5. The results showed that the marker N-P4 could not amplify a band in the translocation line 6V#2S·6AL, but it could amplify a band in the translocation line 6V#4S·6DL (Figure 5b), indicating that N-P4 was a specific marker to 6AS. N-P5 amplified two bands in a wheat line Wan7107, corresponding to the bands in the two translocation lines (Figure 5d), which showed that N-P5 was a real polymorphic marker of 6AS and 6DS. The other two 6D-specific molecular markers ND-P7 and ND-P8 were developed using a similar method, and their amplifications are shown in Figure 5e–h.

Table 3. The molecular markers used in this study.

Primer	Forward Sequence (5′→3′)	Reverse Sequence (5′→3′)	Specific Type	Reference
N-P4	TTAAGAATGTAAGATCGTTGACCCGTAGAC	GGACTGTGACTTGTGAGCATGATGT	6AS	Present study
N-P5	GCGACCTGTTAGAATGCTATTACGATTAC	ATGCTACTCTACCGATGCTTTGAACC	6AS, 6DS	Present study
ND-P7	AACTCTAAGCTCCGCATCATCAATCAT	CTGCTGCCTCATCCAGTTACCAAG	6DS	Present study
ND-P8	GCAACTCTAAGCTCCGCATCATCA	CTGCTGCCTCATCCAGTTACCAAG	6DS	Present study
P1	AGAGTATTTGGTTCCGGATATG	TTTCTGCACACTTGCTGAGGAT	6V#2 (*Pm21*)	Present study
P3	CATACGGAATAGATTTTCCTACCGAAT	TAGCCCTATCTGAAACTGCATGTC	6V#2 (*Pm21*)	Present study
P4	AGTCTGAGGGAGCTGAGGCTTTACA	GACCACATTCATAGAACTGAGGGGAA	6V#2 (*Pm21*)	Present study
P7	GTATGTCAAGGTTTCTGCTTCATACGG	ATATCCTCTAGCGAGATGTGCTCCATA	6V#2 (*Pm21*)	Present study
P9	GTATGTCAAGGTTTCTGCTTCATACGG	AGCTCGCCAAGGCCATTAATATCC	6V#2 (*Pm21*)	Present study
V4	TTTGGTTCCGGACCCTGCCC	GACAACCGTGGCAAGCAGACAA	6V#4 (*Pm21* allele)	Present study
V6	TTTGGTTCCGGACCCTGCCC	ACATGGACGGAGATGAAGAGGAAGAT	6V#4 (*Pm21* allele)	Present study
V11	TTTGGTTCCGGACCCTGCCC	AGTACAGGAGACATGGACGGAGATG	6V#4 (*Pm21* allele)	Present study
6VS-06	GACAGGCAGCTATGAGGC	AATCGTCGTTTGGAGTGG	6V#4S	[10]
6VS-09	GTAAGAACAAGAGGCTAAACAG	CCAGATGACGGTTATTACATAG	6V#4S	[10]
6VS-10	GCCATAAGTGACGCTGAT	GCATCCTGTGAAGTTGTTG	6V#4S	[10]
6VS-12	TGTTGCCTCTCCTCATCA	ATTGCTGTCCGCTCATAC	6V#4S, 6V#2S	[10]
6VS-15	AGGACCATACATTCACAGAG	TTCCATGAGCAGATTAGCA	6V#4S, 6V#2S	[10]
6VS-18	AGCCAGTAAGATTCCGTATG	TCTAACCTTCCTCACAACAC	6V#4S	[10]
P461-5a	GCGTCATCCGCGCCCGTCAGGT	GAGTGCTAATGATAGATGTG	6V#4S, 6V#2S	[22]
P259-1	CGTGATTCAGGAAATGCGATAC	TTGCGCCGCCATGTTAG	6V#4S	[22]
MBH1	GCCATTATAGTCAAGAGTGCACTAGCTGT	AGCTCCTCTCGTTCTCCAATGCT	6AS, 6DS, 6V#4S, 6V#2S	[23]

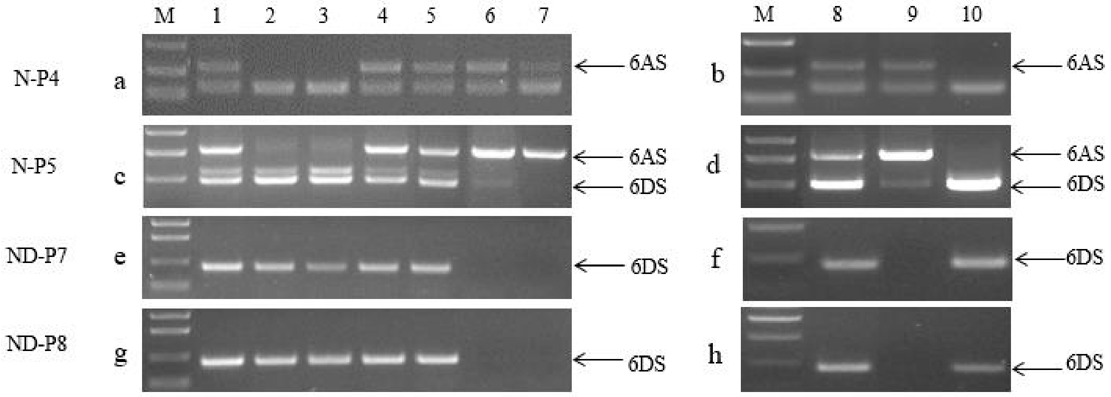


Figure 5. Amplification patterns of 6AS-/6DS-specific molecular markers in wheat, nulli-tetrasomic lines of 6A, 6B, and 6D, and translocation lines. (**a**,**b**) Samples were amplified by marker N-P4. (**c**,**d**) samples were amplified by marker N-P5. (**e**,**f**) samples were amplified by marker ND-P7. (**g**,**h**) samples were amplified by marker ND-P8. M: DL5000; 1 and 8, Wan7107; 2, CSN6AT6B; 3, CSN6AT6D; 4, CSN6BT6A; 5, CSN6BT6D; 6, CSN6DT6A; 7, CSN6DT6B; 9, 6V#4S·6DL translocation line Pm97033; 10, 6V#2S·6AL translocation line Nanyi.

According to a previous report [23], MBH1 can specifically amplify 6V#2S/6V#4S/6AS/6DS chromosomes simultaneously. We carefully investigated the amplification bands of MBH1 in the F_2 populations, and further verified them with the markers N-P4 and N-P5 developed in this study. Based on the specific amplified bands of 6A and 6D, the transmission rate of 6D was higher than that of 6A in the F_2 populations (Table 4).

Table 4. Frequencies of chromosomes 6A and 6D in the F_2 populations of two alien substitution lines.

Total Plants Investigated	$6A^+6D^-$	$6A^-6D^+$	$6A^+6D^+$	$6A^-6D^-$
319 100%	86 26.96%	121 37.93%	97 30.41%	15 4.70%

2.5. Development of Pm21 and Its Allele-Specific Molecular Markers and Detection of Both Genes in the Testcross Populations

In order to develop specific markers to detect the PM resistance gene *Pm21* on 6V#2 and its allele on the 6V#4 chromosome in a common wheat background, a multiple sequence alignment was conducted between the two genes and their homologous sequences in wheat. Primers were designed among the polymorphic sites. The schematic diagram of the distribution of primers on both alien genes is shown in Figure 6a. Primers V4, V6, and V11 are linked with the *Pm21* allele, while primers P1, P3, P4, P7, and P9 are linked with *Pm21*; their amplifications are shown in Figure 6b.

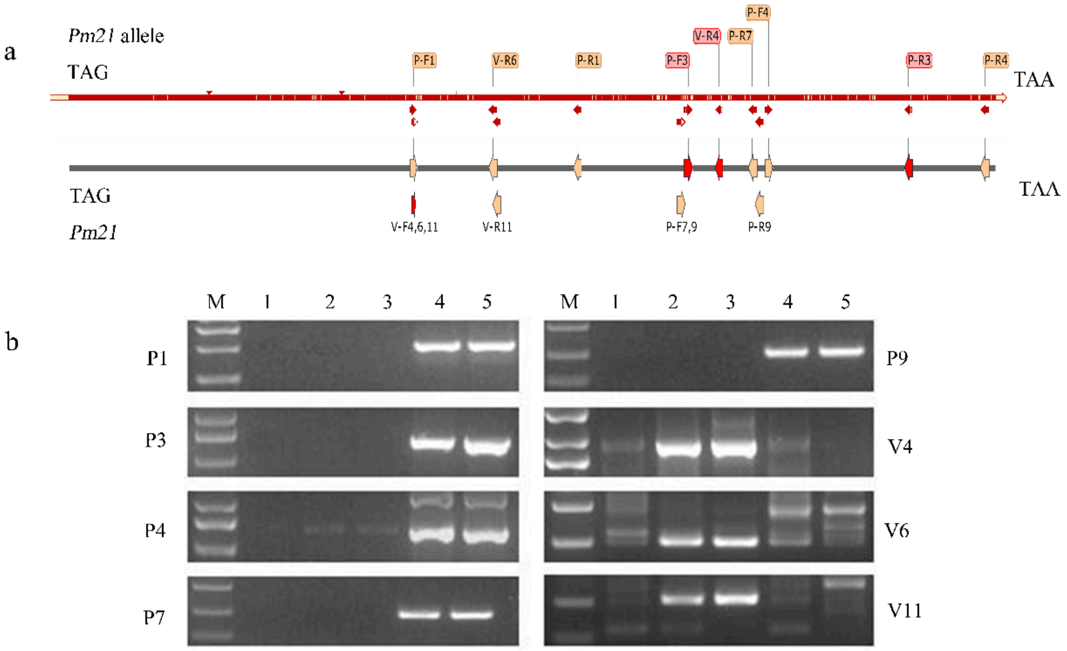


Figure 6. *Pm21* and its allele-specific markers and their chromosomal distribution. (**a**) The schematic diagram of the distribution of all primers for the *Pm21* gene and its alleles. The *Pm21* gene and its alleles were amplified with primers (P3 and V4) marked by red arrows in this experiment. (**b**) Development of molecular markers of the *Pm21* gene and its alleles. M: DL5000; 1: Wan7107; 2: translocation line 6V#4S·6DL; 3: Hv-s (6V#4); 4: translocation line 6V#4S·6AL; 5: CMM (6V#2).

The newly developed primers were used to trace the alien chromosomes in the testcross generations with 482 individuals. *Pm21* or its allele-specific markers were amplified in most of the plants. However, there were 16 plants where no alien-specific bands were amplified, of which seven showed susceptibility and nine showed resistance to PM. Moreover, there were seven other plants where both *Pm21* and its allele were amplified. In order to confirm the results, eight additional 6V#2/6V#4-polymorphic PCR markers were used to amplify bands in the individuals that contain both genes. The results indicated that 6V#2- and 6V#4-specific bands were also simultaneously amplified by three co-dominant markers except in one plant, which only had bands specific to the 6V#4 chromosome, suggesting that it was a recombinant of *Pm21* and its allele or a heterozygote of two alien chromosomes (sample 2 in Figure 7).

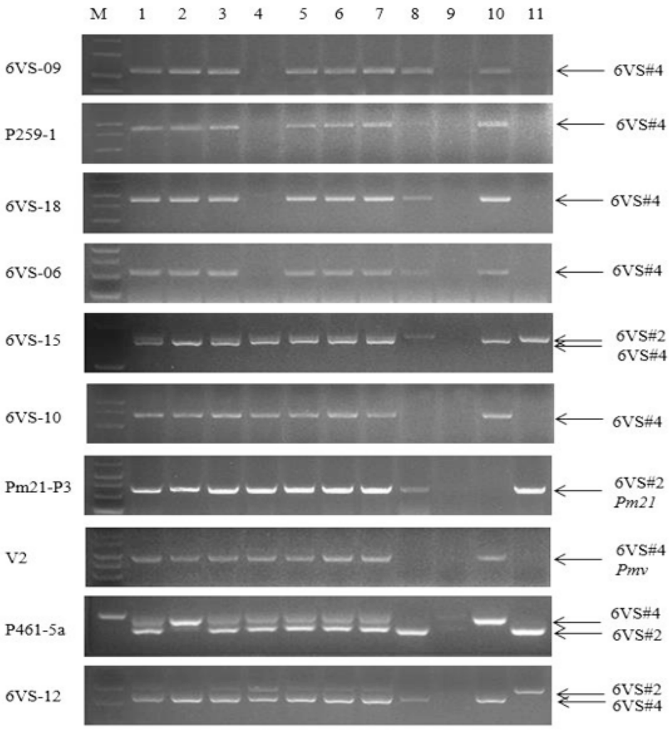


Figure 7. Amplification of some individuals in the testcross generations. Samples 1–8 are individuals from a testcross generation. M: DL5000; 9: Wan7107; 10: translocation line 6V#4S·6DL; 11: translocation line 6V#2S·6AL.

2.6. Identification of New Genotypes

According to the amplified bands by exogenous chromosome-specific markers and A and D chromosome-specific markers in group 6, we found that, in addition to the parental substitution types 6V#2 (6A) (sample 1 in Figure 2 and Table 1) and 6V#4 (6D) (sample 8 in Figure 2 and Table 1), there were three new types of substitution lines of 6V#2 (6D) (sample 11 in Figure 2 and Table 1), 6V#4 (6A) (sample 7 in Figure 2 and Table 1), and 6D or 6A chromosomes substituted by 6V#2#4 recombinant chromosome (samples 9 and 12 in Figure 2 and Table 1) in the F_2 populations.

3. Discussion

3.1. Distribution of Amplified Sequences on Chromosomes and Their Separation Ratios in Hybrid Progenies

In this study, nine molecular markers were used to detect exchange events between the two alien chromosomes 6V#2 and 6V#4 in the F_2 populations derived from the two alien substitution lines. The separation ratio of the nine markers was investigated based on the chi-square test. Among them, seven markers segregated consistently with the theoretical ratios 3:1 or 1:2:1 of Mendel's separation law at the 0.01 level, suggesting that the alien chromosomes had regular separating behavior, which is consistent with cytological observations of F_1 hybrids.

However, the separation ratios of two markers, 6VS-09 and 6VS-12, did not fit the theoretical ratios. Different competitiveness of gametes or genetic disorder in distant hybridization was suggested as the cause for distorted segregation [18,19,24,25]. For PCR-based molecular markers, we believe that the separation ratio might be closely related to the copy number and distribution mode of primer amplification regions on the genome or chromosome, and the sequences of primers.

Due to lacking genomic sequence information of *D. villosum*, the copy number of the amplified regions on 6V and the distance between different copies are not clear. However, BLAST performed on the sequences of the nine aforementioned molecular markers showed considerable variation in the copies of the homologous sequences on wheat chromosomes 6A, 6B, and 6D, and barley chromosome 6H and their chromosomal distributions. Therefore, if these markers have more than one copy of primer-matched binding sequences, and exchange occurs at the same time between these copies, it may affect the separation ratio of the molecular markers in the segregation populations. In addition, other factors might lead to significant differences between the actual statistical results and the theoretical results, including distribution differences of homologous sequences on the two alien chromosomes, and competitive amplification caused by the complexity of some primers binding to DNA template (for example, when some primers amplified the mixed DNA samples of translocation lines 6V#2S·6AL and 6V#4S·6DL, occasionally, priority was given to the short target fragment).

3.2. Two Alien Chromosomes 6V#2 and 6V#4 Pair and Exchange in Their Hybrids

Both 6V#2S and 6V#4S are known to contain powdery mildew (PM) resistance genes, which is the only phenotypic trait that can be tracked in a wheat background, but they cannot be distinguished by phenotype from each other. Therefore, it is impossible to determine the existence of exchange and recombination by phenotype in F_2 hybrids. In addition, the feasibility of the allelic test can only be confirmed by verifying that exchange and recombination events can occur.

Some alien chromosomes cannot be properly paired and exchanged in a wheat background, on account of the large genetic differences between homologous chromosomes from different *D. villosum* sources. For example, Qi et al. [12] selected 40 restriction fragment length polymorphism (RFLP) probes specific to the sixth homoeologous chromosomes of wheat to analyze disomic substitution lines 6V#2 (6A) and 6V#1 (6A), for which 25 probes could detect the differences between 6V#2 and 6V#1. Using a cytological technique, the authors observed that chromosomes 6V#2 and 6V#1 did not pair, indicating that they had significant structural differences. According to our previous research, chromosome-pairing frequency of 6V#2S and 6V#4S in the F_1 hybrids between translocation lines 6V#2S·6AL and 6V#4S·6DL was 18.9% [26], and, in most cases, they were not paired. This might be related to different structures of the two alien chromosome arms and the kinetic difference during the meiosis process of each translocated wheat chromosome.

In this study, two alien substitution lines, 6V#2 (6A) and 6V#4 (6D) were hybridized to construct F_2 segregating populations. Whether the two alien chromosomes 6V#2 and 6V#4 can normally pair and exchange without the involvement of translocation between alien and wheat chromosomes was investigated. Using the nine markers to detect the F_2 generations, we found that non-exchanged parental type was 25.70%, exchange type was 73.99%, and wheat type without alien chromosomes was 0.31%, indicating that pairing and exchanging between the two alien chromosomes occurred at a rational frequency. However, it was observed that the percentage of the alien bivalent rod was much higher than that of the alien bivalent ring formed at meiosis in the F_1 hybrids, indicating that the pairing between 6V#2S and 6V#4S is still limited. Therefore, the allelism test should be carried out in a larger populations.

The type of gametes formed in the F_1 hybrids can be detected in the testcross populations. Theoretically, each individual in the testcross populations should contain only one alien chromosome. According to our results, out of 482 individuals, two exogenous chromosomes, 6V#2 and 6V#4, were simultaneously amplified in seven individuals by *pm21* and its allele-specific markers and three other molecular markers, 6VS-12, 6VS-15, and P461-5a. Moreover, seven PM-susceptible individuals without alien chromosomes were detected as well. These results suggested that nearly 1.5% of the alien homologous chromosomes or the sister chromosomes did not separate, and both moved to the same pole during meiosis anaphase I or II, resulting in the other pole lacking the corresponding alien chromosomes, thus forming the types we detected. This can also be supported by the fact that 20.24% of alien chromosomes are not paired during the meiosis of F_1 hybrids. In the testcross populations, we also detected nine PM-resistant plants lacking alien chromosome-specific markers. A possible explanation for this might be other unknown PM resistance gene(s) in the wheat genome.

3.3. Screening of Molecular Markers for Novel Alien Substitution Lines

At present, substitution lines are mainly derived from five methods including spontaneous substitution, monosomic substitution, telosomic substitution, nullisomic backcrossing, and tissue culture [27–30]. In theory, the homoeologous chromosomes of three wheat subgenomes can be replaced with alien chromosomes, forming three different types of alien substitution lines. The type of alien substitution lines might be related to the developing technique. Recently, we identified 6V#4 (6D) alien substitution lines from the progeny of *T. durum–D. villosum* amphidiploids crossed and backcrossed with common wheat (unpublished). In this case, genomes A and B have a homologous pairing partner during meiosis in F_1 and backcrossing generations, and they can normally separate at anaphase I. Genomes V and D lack their homologous partner chromosomes and, thus, D genome chromosomes are more easily replaced by V chromosomes.

In this study, two different substitution lines, 6V#4 (6D) and 6V#2 (6A), were used to hybridize; theoretically, in the F_1 hybrids, 6V#2 and 6V#4, as homologous chromosomes, can be paired and exchanged with each other to form new recombinant chromosomes if they have high affinity without significant structural differences. In contrast, chromosomes 6A and 6D, due to a lack of homologous chromosomes, became two univalents at meiosis metaphase I. In addition, the two univalents separated randomly at anaphase I, and were independently assorted in the F_2 offspring, which provided a chance for the formation of new substitution types (Figure 8). Unexpectedly, a trivalent chromosome and a univalent chromosome instead of two univalent chromosomes were observed at metaphase I, suggesting that a 6A or 6D chromosome was involved in the trivalent. At the same time, the frequency of chromosome 6D in the F_2 generation was higher than that of chromosome 6A. Considering that 6A or 6D was involved in the trivalent at meiosis metaphase I, 6D should be the target participating in the translocation event. Using the MAS technique, we obtained three new types of substitution lines and new alien chromosome-exchange types in F_2 and F_3, which lays the foundation for further genetic evaluation, such as studying the differences of genetic effects between 6V#4 (6A) and 6V#4 (6D) or 6V#2 (6A) and 6V#2 (6D).

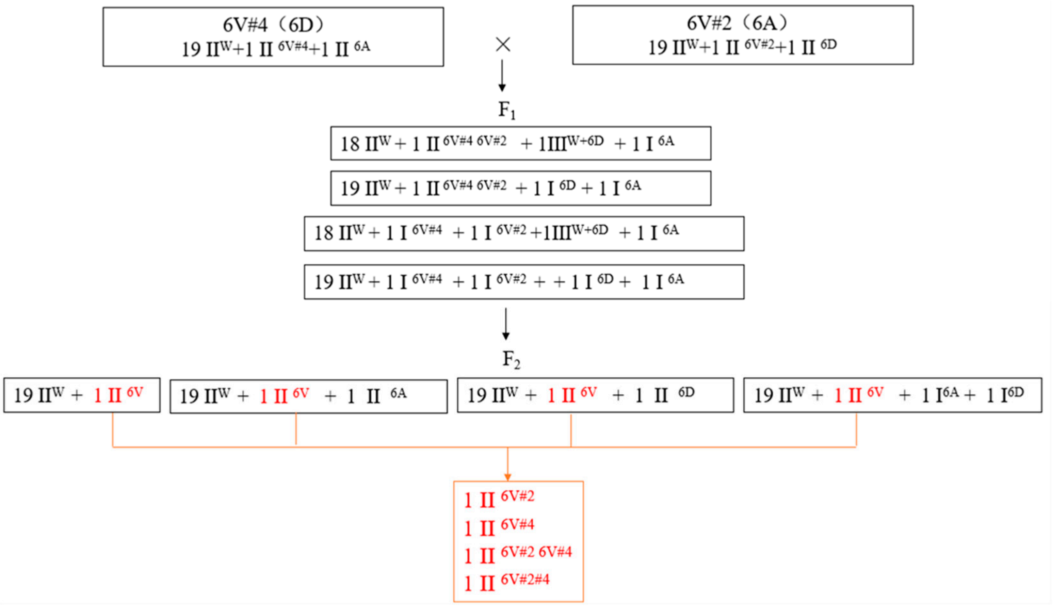

6V#4（6D）
19 II^W+1 $II^{6V\#4}$+1 II^{6A}
×
6V#2（6A）
19 II^W+1 $II^{6V\#2}$+1 II^{6D}
F_1
18 II^W + 1 $II^{6V\#4\ 6V\#2}$ + 1III^{W+6D} + 1 I^{6A}
19 II^W + 1 $II^{6V\#4\ 6V\#2}$ + 1 I^{6D} + 1 I^{6A}
18 II^W + 1 $I^{6V\#4}$ + 1 $I^{6V\#2}$ +1III^{W+6D} + 1 I^{6A}
19 II^W + 1 $I^{6V\#4}$ + 1 $I^{6V\#2}$ + + 1 I^{6D} + 1 I^{6A}
F_2
19 II^W + 1 II^{6V}
19 II^W + 1 II^{6V} + 1 II^{6A}
19 II^W + 1 II^{6V} + 1 II^{6D}
19 II^W + 1 II^{6V} + 1 I^{6A} + 1 I^{6D}
1 $II^{6V\#2}$
1 $II^{6V\#4}$
1 $II^{6V\#2\ 6V\#4}$
1 $II^{6V\#2\#4}$

Figure 8. Development of new exchange types in this study. W indicates wheat chromosomes; 6V indicates alien chromosomes. There are four possible combinations of II^{6V} in F_2 shown in red, which are $II^{6V\#2}$, $II^{6V\#4}$, $II^{6V\#26V\#4}$ (heterozygote), and $II^{6V\#2\#4}$ (recombinant).

3.4. Importance of Genetic Linkage Map for Genera That Lack Genome Sequence Information

Genetic linkage maps provide breeders with the order of markers and information of target genes on the chromosomes [31–35]. Molecular markers are suitable for constructing genetic linkage maps. These molecular markers not only reflect the genetic diversity at the DNA level, but also show high polymorphism, and co-dominance features that can identify homozygous and heterozygous genotypes [36]. The basic principle of constructing genetic linkage maps is generated from eukaryotes undergoing meiosis, in which the chromosomes are recombined and exchanged, and the exchanging and recombination rates are also affected by the relative distance between any two sites on chromosomes. Qi et al. [37] mapped more than 16,000 expressed sequence tag (EST) loci to specific physical intervals on different wheat chromosomes. These EST sequences and localization provided information for constructing a genetic linkage map of target genes.

At present, the genome databases of wheat, barley, and other cereal crops are gradually improving with the rapid development and cost reduction of sequencing technology [38–42], which provides great convenience for the study of wild wheat species without genomic information. The *ph1b* gene can induce pairing and exchange between alien chromosomes and their homoeologous chromosomes in wheat backgrounds [43,44]. The paired homoeologous chromosomes need to have high collinearity, and structural variation of chromosomes including the inversion of fragments may influence the chromosomes pairing and exchanging. In this study, a genetic linkage map was constructed using nine markers, and the order of the markers on the genetic linkage map 6V was nearly consistent with those on barley 6H and wheat 6B, followed by 6A and 6D. The quantity of markers used in this study was limited; thus, we need to develop additional molecular markers and increase the density of markers on chromosome 6VS. Increasing marker density is necessary for making more accurate maps and better evaluating the conservation of the linear ordering of the sequences on alien chromosomes with those of their wheat homoeologous chromosomes. Such studies are of great significance for investigating the phylogenetic relationship between *D. villosum* and other cereal crops, transferring beneficial genes into wheat, and obtaining an ideal genetic compensation effect in translocation materials.

4. Materials and Methods

4.1. Plant Materials

D. villosum accession *D.v#2* from Cambridge Botanical Garden in England and its derived T6V#2S·6AL translocation line Nanyi, and 6V#2(6A) alien substitution line Nan87-88 were provided by Peidu Chen at Nanjing Agricultural University, China. Chinese Spring group 6 nulli-tetrasomic stocks CSN6AT6B, CSN6AT6D, CSN6BT6A, CSN6BT6D, CSN6DT6A, and CSN6DT6B were provided by Dr. Steven Xu (USDA-ARS, Northern Crop Science Laboratory, USA) and preserved at the Institute of Crop Science (ICS), Chinese Academy of Agricultural Sciences (CAAS). *D. villosum* accession No. 1026 from Russia and its derived T6V#4S·6DL translocation line Pm97033, 6V#4(6D) alien substitution line RW15, and wheat cultivar Wan7107 were provided by professor Chen Xiao at ICS, CAAS. The F_1 populations were derived from a cross between Nan87-88 and RW15, and the F_2 populations were obtained by self-crossing of the F_1 plants. Testcross generations were obtained from a cross between F_1 populations and wheat cultivar Wan7107. The phenotype of wheat powdery mildew resistance was evaluated at the seedling stage and mature plant stage.

4.2. Molecular Markers

Nine 6V-specific molecular markers developed previously in our laboratory [10,22] and by Bie et al. [23] were used in the present study. Four 6AS- and 6DS-specific markers were developed for this study as follows: sequences of chromosome 6A/6B/6D in Fasta format were downloaded from http://plants.ensembl.org/Triticum_aestivum/Info/Index. The files were opened in SnapGene software (https://www.snapgene.com/), and a fragment from one chromosome with a certain size at the exon was selected. BLAST analyses were conducted using the Ensembl website, and the sequences with the highest identity to this fragment on the short chromosome arms of the other two genomes in the sixth homoeologous group were selected. Primers were designed based on the region showing the polymorphic sequences among the three genomes using Primer premier 6.0 software (http://www.premierbiosoft.com/primerdesign/), and synthesized by Sangon Biotech company (Shanghai, China).

4.3. DNA Extraction and PCR Amplification Conditions

Genomic DNA was extracted from 1 g of the leaves of two-month-old seedlings using Nuclear Plant Genomic DNA Kit (CW Bio Inc., Beijing, China). The DNA pellet was dissolved into a concentration of 100 ng·µL^{-1} in sterile water, and stored at −20 °C. PCR was carried out in a 15-µL reaction volume consisting of 100 ng of template DNA, 0.3 µL of each primer (10 µmol·L^{-1}), 5.9 µL of water, and 7.5 µL of 2× Taq Master Mix (containing Mg^{2+} and dNTPs, Vazyme Biotech Co., Ltd., Nanjing, China). PCR amplification was performed in a Biometra Professional thermal cycler (Göttingen, Germany) with an initial denaturation at 95 °C for 5 min, 35 cycles of 20 s at 95 °C, 30 s at 54–60°C (annealing temperatures varied with the primer), 40–60 s (extension times varied with the product size) at 72 °C, and a final extension of 8 min at 72 °C. PCR products were separated on 1.5% agarose gels with a standard TAE buffer (40 mM Tris, 20 mM acetic acid, 1 mM EDTA) and visualized in a gel doc system (Bio-Rad, Hercules, CA, USA) after Genecolour II staining.

4.4. Construction of a Linkage Map and Comparison Maps

In this study, 323 plants in the aforementioned F_2 populations were randomly selected and analyzed using the nine polymorphic molecular markers. QTL IciMapping software [45] (http://www.isbreeding.net/) was used to draw a genetic linkage map.

The sequences of the nine 6VS-specific markers were BLASTed on the website http://plants.ensembl.org/index.html. According to the physical position information of the homologous sequences on wheat 6A, 6B, 6D and barley 6H, the physical maps were drawn and compared.

4.5. Cytological Procedures

Anthers at meiosis stages were fixed in Carnoy's Fluid for 24 h and then transferred to 70% ethanol and stored at −20 °C. Before cytological preparation, the anthers were washed with water three times, then put into 1× (A + B) solution (0.1 M citric acid + 0.1 M tri-sodium citrate) for 11 min, and transferred to enzymatic hydrolysate (2% cellulase R-10 + 2% pectinase Y-23, Japan) at 37 °C for 6.5 min. The reaction was terminated by adding 1× (A + B) solution. Anthers were put on slides, and one drop of 45% glacial acetic acid was added; the anther was lightly mashed with a dissecting needle, and then covered with a coverslip and pressed vertically with the thumb. The slide was then frozen in liquid nitrogen and stored at −20 °C after removing the coverslip.

Genomic in situ hybridization was carried out according to the method introduced by Wei et al. [46]. DIG-Nick Translation Mix Kit (Roche, Mannheim, Germany) was used for probe labeling, the total genomic DNA of *D. villosum* was labeled to use as a probe, and Chinese Spring genomic DNA was included as blocking DNA. Hybridization signals were identified by using a fluorescein isothiocyanate (FITC)-conjugated Anti-Digoxigenin Fluorescein Fab Fragments Kit (Roche, Mannheim, Germany). Observation of chromosome pairing configuration was carried out with a fluorescence microscope (BX51, Olympus Co., Ltd., Tokyo, Japan).

5. Conclusions

Firstly, 6V#2 and 6V#4, chromosomes from different *Dasypyrum villosum* accessions, were found to pair and exchange. This was confirmed by cytological observation in their F_1 hybrids, and by analyzing nine molecular markers in their F_2 populations, which allows genes controlling desirable traits from different alien chromosomes to be pyramided, and provides important information for the alien allelism test of some vital genes. Secondly, the genetic linkage map of nine markers on 6VS was compared with the physical maps of wheat and barley based on homologous sequences of the markers, which showed that the conservation of sequence order compared to 6V was 6H, 6B > 6A > 6D. Finally, three wheat–*D. villosum* substitutions 6V#4 (6A), 6V#2 (6D), and the alien recombinants 6V#2#4 (6A/6D) with novel resistance to powdery mildew were selected using 6V-, 6A-, and 6D-specific molecular markers.

Author Contributions: The experiment was designed by Z.L.; X.M., Z.X., J.W., and H.C. conducted the cytologic experiment; X.M. and Z.X. conducted the PCR experiment; X.M. analyzed the data; X.M. developed the wheat, *Pm21*, and its allele-specific molecular markers. The manuscript was drafted by X.M., Z.L., and X.Y., and corrected and approved by all authors.

Funding: This research was financially supported by the National Key Research and Development Program of China (2016YFD0102001 and 2016YFD0102002), and the Agricultural Science and Technology Innovation Program (ASTIP) of the Chinese Academy of Agricultural Sciences.

Acknowledgments: We are grateful to Yiping Yan at ICS-CAAS for managing wheat plants in the growth chamber and assistance with sampling. We also thank Alia Anwar for linguistic modification.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

Abbreviations

GISH	genomic in situ hybridization
MAS	marker-assisted selection
PMCs	pollen mother cells
FITC	fluorescein isothiocyanate
PCR	polymerase chain reaction
PM	powdery mildew

References

1. Grądzielewska, A. The genus Dasypyrum-part 2. *Dasypyrum villosum*-a wild species used in wheat improvement. *Euphytica* **2006**, *152*, 441–454.
2. Pace, C.D.; Snidaro, D.; Ciaffi, M.; Vittori, D.; Ciofo, A.; Cenci, A.; Tanzarella, O.A.; Qualset, C.O.; Scarascia Mugnozza, G.T. Introgression of *Dasypyrum villosum* chromatin into common wheat improves grain protein quality. *Euphytica* **2011**, *117*, 67–75. [CrossRef]
3. Qi, L.; Friebe, B.; Zhang, P.; Gill, B.S. Homoeologous recombination, chromosome engineering and crop improvement. *Chromosome Res.* **2007**, *15*, 3–19. [CrossRef]
4. Li, K.; Hegarty, J.; Zhang, C.; Wan, A.; Wu, J.; Guedira, G.B.; Chen, X.; Muñoz-Amatriaín, M.; Fu, D.; Dubcovsky, J. Fine mapping of barley locusRps6conferring resistance to wheat stripe rust. *Theor. Appl. Genet.* **2016**, *129*, 845–859. [CrossRef]
5. Xing, L.; Hu, P.; Liu, J.; Witek, K.; Zhou, S.; Xu, J.; Zhou, W.; Gao, L.; Huang, Z.; Zhang, R.; et al. *Pm21* from *Haynaldia villosa* Encodes a CC-NBS-LRR that Confers Powdery Mildew Resistance in Wheat. *Mol. Plant.* **2018**, *11*, 874–878. [CrossRef]
6. Blanco, A.; Resta, P.; Simeone, R.; Parmar, S.; Shewry, P.R.; Sabelli, P.; Lafiandra, D. Chromosomal location of seed storage protein genes in the genome of *Dasypyrum villosum* (L.) Candargy. *Theor. Appl. Genet.* **1991**, *82*, 358–362. [CrossRef]
7. Chen, P.; Liu, D. Cytogenetic studies of hybrid progenies between triticum aestivum and haynaldia villosa. *J. Nanjing Agric. Univ.* **1982**, *4*, 1–16.
8. Blanco, A.; Simeone, R.; Resta, P. The addition of *Dasypyrum villosum* (L.) Candargy chromosomes to durum wheat (*Triticum durum* Desf.). *Theor. Appl. Genet.* **1987**, *74*, 328–333. [CrossRef]
9. Li, G.; Zhao, J.; Li, D.; Yang, E.; Huang, Y.; Liu, C.; Yang, Z. A Novel Wheat-*Dasypyrum breviaristatum* Substitution Line with Stripe Rust Resistance. *Cytogenet. Genome Res.* **2014**, *143*, 280–287. [CrossRef]
10. Li, S.; Lin, Z.; Liu, C.; Wang, K.; Du, L.; Ye, X. Development and comparative genomic mapping of *Dasypyrum villosum* 6V#4S-specific PCR markers using transcriptome data. *Theor. Appl. Genet.* **2017**, *130*, 2057–2068.
11. Li, S.; Wang, J.; Wang, K.; Chen, J.; Wang, K.; Du, L.; Ni, Z.; Lin, Z.; Ye, X. Development of PCR markers specific to *Dasypyrum villosum* genome based on transcriptome data and their application in breeding *Triticum aestivum-D. villosum*#4 alien chromosome lines. *BMC Genom.* **2019**, *20*, 289.

12. Qi, L.; Wang, S.; Chen, P.; Liu, D.; Gill, B.S. Identification and physical mapping of three *Haynaldia villosa* chromosome-6V deletion lines. *Theor. Appl. Genet.* **1998**, *97*, 1042–1046. [CrossRef]
13. He, H.; Zhu, S.; Zhao, R.; Jiang, Z.; Ji, Y.; Ji, J.; Qiu, D.; Li, H.; Bie, T. *Pm21* encoding a typical CC-NBS-LRR protein, confers broad-spectrum resistance to wheat powdery mildew disease. *Mol. Plant* **2018**, *11*, 879–882. [CrossRef] [PubMed]
14. Liu, Z.; Zhu, J.; Hua, W.; Yang, Z.; Sun, Q.; Liu, Z. Comparative Genomics Analysis and Constructing EST Markers Linkage Map of Powdery Mildew Resistance Gene *pm42* in Wheat. *Acta Agron. Sin.* **2011**, *37*, 1569–1576. (In Chinese) [CrossRef]
15. Ando, K.; Krishnan, V.; Rynearson, S.; Rouse, M.N.; Danilova, T.; Friebe, B.; See, D.; Pumphrey, M.O. Introgression of a Novel Ug99-Effective Stem Rust Resistance Gene into Wheat and Development of *Dasypyrum villosum* Chromosome-Specific Markers via Genotyping-by-Sequencing (GBS). *Plant Dis.* **2019**, *103*, 1068–1074. [CrossRef]
16. Zhang, R.; Zhang, M.; Wang, X.; Chen, P. Introduction of chromosome segment carrying the seed storage protein genes from chromosome 1V of *Dasypyrum villosum* showed positive effect on bread-making quality of common wheat. *Theor. Appl. Genet.* **2014**, *127*, 523–533.
17. Zhang, R.; Hou, F.; Feng, Y.; Zhang, W.; Zhang, M.; Chen, P. Characterization of a *Triticum aestivum-Dasypyrum villosum* T2VS·2DL translocation line expressing a longer spike and more kernels traits. *Theor. Appl. Genet.* **2015**, *128*, 2415–2425. [CrossRef]
18. Zhang, J.; jiang, Y.; Wang, Y.; Guo, Y.; Long, H.; Deng, G.; Chen, Q.; Xuan, P. Molecular markers and cytogenetics to characterize a wheat-*Dasypyrum villosum* 3V (3D) substitution line conferring resistance to stripe rust. *PLoS ONE* **2018**, *13*, e0202033. [CrossRef]
19. Wang, H.; Dai, K.; Xiao, J.; Yuan, C.; Zhao, R.; Doležel, J.; Wu, Y.; Cao, A.; Chen, P.; Zhang, S.; et al. Development of intron targeting (IT) markers specific for chromosome arm 4VS of *Haynaldia villosa* by chromosome sorting and next-generation sequencing. *BMC Genom.* **2017**, *18*, 167. [CrossRef]
20. Zhang, R.; Sun, B.; Chen, J.; Cao, A.; Xing, L.; Feng, Y.; Lan, C.; Chen, P. Pm55, a developmental-stage and tissue-specific powdery mildew resistance gene introgressed from *Dasypyrum villosum* into common wheat. *Theor. Appl. Genet.* **2016**, *129*, 1975–1984. [CrossRef]
21. Cao, Y.; Cao, A.; Wang, X.; Chen, P. Screening and Application of EST-Based PCR Markers Specific to Individual Chromosomes of *Haynaldia villosa*. *Acta Agron. Sin.* **2009**, *35*, 1–10.
22. Liu, C.; Li, S.; Wang, K.; Ye, X.; Lin, Z. Developing of Specific Transcription Sequences P21461 and P33259 on *Dasypyrum villosum* 6VS and Application of Molecular Markers in Identifying Wheat-*D. villosum* Breeding Materials with Powdery Mildew Resistance. *Acta Agron. Sin.* **2017**, *43*, 983–992. (In Chinese) [CrossRef]
23. Bie, T.; Zhao, R.; Zhu, S. Development and characterization of marker MBH1 simultaneously tagging genes *Pm21* and *PmV* conferring resistance to powdery mildew in wheat. *Mol. Breed* **2015**, *35*, 189. [CrossRef]
24. Kreiner, J.M.; Kron, P.; Husband, B.C. Evolutionary Dynamics of Unreduced Gametes. *Trends Genet.* **2017**, *33*, 583–593. [CrossRef]
25. Mason, A.S.; Pires, J.C. Unreduced gametes: meiotic mishap or evolutionary mechanism? *Trends Genet.* **2015**, *31*, 5–10. [CrossRef]
26. Liu, C.; Ye, X.; Wang, M.; Li, S.; Lin, Z. Genetic behavior of *Triticum aestivum-Dasypyrum villosum* translocation chromosomes T6V#4S·6DL and T6V#2S·6AL carrying powdery mildew resistance. *J. Integr. Agric.* **2017**, *16*, 2136–2144.
27. Zhang, X.; Chen, S. Production and Utilization of Alien Substitution Lines of Common Wheat. *Hereditas* **1990**, *12*, 40–44. (In Chinese)
28. Sears, E.R. Nullisomic Analysis in Common Wheat. *Am. Nat.* **1953**, *87*, 245–252. [CrossRef]
29. Ghazali, S.; Mirzaghaderi, G.; Majdi, M. Production of a novel Robertsonian translocation from *Thinopyrum bessarabicum* into bread wheat. *Tsitol Genet* **2015**, *49*, 38–42. [CrossRef]
30. Danilova, T.V.; Friebe, B.; Gill, B.S.; Poland, J.; Jackson, E. Chromosome Rearrangements Caused by Double Monosomy in Wheat-Barley Group-7 Substitution Lines. *Cytogenet. Genome Res.* **2018**, *154*, 45–55. [CrossRef]
31. Zhou, S.; Zhang, J.; Che, Y.; Liu, W.; Lu, Y.; Yang, X.; Li, X.; Jia, J.; Liu, X.; Li, L. Construction of *Agropyron Gaertn*. genetic linkage maps using a wheat 660K SNP array reveals a homoeologous relationship with the wheat genome. *Plant Biotechnol. J.* **2018**, *16*, 818–827. [CrossRef] [PubMed]

32. Danilova, T.V.; Friebe, B.; Gill, B.S. Development of a wheat single gene FISH map for analyzing homoeologous relationship and chromosomal rearrangements within the Triticeae. *Theor. Appl. Genet.* **2014**, *127*, 715–730. [CrossRef] [PubMed]
33. Zhang, Y.; Zhang, J.; Huang, L.; Gao, A.; Zhang, J.; Yang, X.; Liu, W.; Li, X.; Li, L. A high-density genetic map for P genome of *Agropyron Gaertn*. based on specific-locus amplified fragment sequencing (SLAF-seq). *Planta* **2015**, *242*, 1335–1347. [CrossRef] [PubMed]
34. Luo, Z.; Hackett, C.A.; Bradshaw, J.E.; McNicol, J.W.; Milbourne, D. Construction of a genetic linkage map in tetraploid species using molecular markers. *Genetics* **2001**, *157*, 1369–1385.
35. Qi, P.; Eudy, D.; Schnable, J.C.; Schmutz, J.; Raymer, P.L.; Devos, K.M. High Density Genetic Maps of Seashore Paspalum Using Genotyping-By-Sequencing and Their Relationship to The Sorghum Bicolor Genome. *Sci. Rep.* **2019**, *9*, 12183. [CrossRef]
36. Jia, J. Molecular Germplasm Diagnostics and Molecular Marker assisted Breeding. *Scientia Agricultura Sinica* **1996**, *29*, 1–10. (In Chinese)
37. Qi, L.; Echalier, B.; Chao, S.; Lazo, G.R.; Butler, G.E.; Anderson, O.D.; Akhunov, E.D.; Dvorák, J.; Linkiewicz, A.M.; Ratnasiri, A.; et al. A Chromosome Bin Map of 16,000 Expressed Sequence Tag Loci and Distribution of Genes Among the Three Genomes of Polyploid Wheat. *Genetics* **2004**, *168*, 701–712. [CrossRef]
38. Zimin, A.V.; Puiu, D.; Hall, R.; Kingan, S.; Clavijo, B.J.; Salzberg, S.L. The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *Gigascience* **2017**, *6*, 1–7. [CrossRef]
39. Lu, F.H.; McKenzie, N.; Kettleborough, G.; Heavens, D.; Clark, M.D.; Bevan, M.W. Independent assessment and improvement of wheat genome sequence assemblies using Fosill jumping libraries. *Gigascience* **2018**, *7*, giy053. [CrossRef]
40. International Barley Genome Sequencing Consortium; Mayer, K.F.; Waugh, R.; Brown, J.W.; Schulman, A.; Langridge, P.; Platzer, M.; Fincher, G.B.; Muehlbauer, G.J.; Sato, K.; et al. A physical, genetic and functional sequence assembly of the barley genome. *Nature* **2012**, *491*, 711–716.
41. Ariyadasa, R.; Mascher, M.; Nussbaumer, T.; Schulte, D.; Frenkel, Z.; Poursarebani, N.; Zhou, R.; Steuernagel, B.; Gundlach, H.; Taudien, S.; et al. A sequence-ready physical map of barley anchored genetically by two million single-nucleotide polymorphisms. *Plant Physiol.* **2014**, *164*, 412–423. [CrossRef] [PubMed]
42. Mascher, M.; Gundlach, H.; Himmelbach, A.; Beier, S.; Twardziok, S.O.; Wicker, T.; Radchuk, V.; Dockter, C.; Hedley, P.E.; Russell, J.; et al. A chromosome conformation capture ordered sequence of the barley genome. *Nature* **2017**, *544*, 427–433. [CrossRef] [PubMed]
43. Zhao, R.; Wang, H.; Xiao, J.; Bie, T.; Cheng, S.; Jia, Q.; Yuan, C.; Zhang, R.; Cao, A.; Chen, P.; et al. Induction of 4VS chromosome recombinants using the CS ph1b mutant and mapping of the wheat yellow mosaic virus resistance gene from *Haynaldia villosa*. *Theor. Appl. Genet.* **2013**, *126*, 2921–2931. [CrossRef] [PubMed]
44. Li, H.; Deal, K.R.; Luo, M.; Ji, W.; Distelfeld, A.; Dvorak., J. Introgression of the *Aegilops speltoides* Su1-Ph1 Suppressor into Wheat. *Front. Plant. Sci.* **2017**, *8*, 2163. [CrossRef] [PubMed]
45. Meng, L.; Li, H.; Zhang, L.; Wang, J. QTL IciMapping: Integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* **2015**, *3*, 269–283. [CrossRef]
46. Wei, W.H.; Qin, R.; Song, Y.C.; Ning, S.B.; Guo, L.Q.; Gu, M.G. Location and analysis of introgressed segments in the parthenogenetic progenies of Zea mays×Z. diploperennis by GISH. *Acta Bot. Sin.* **2002**, *44*, 373–376.

International Journal of
Molecular Sciences

MDPI

Article

Analysis of *Centranthera grandiflora* Benth Transcriptome Explores Genes of Catalpol, Acteoside and Azafrin Biosynthesis

Xiaodong Zhang [1,2,†], Caixia Li [1,2,†], Lianchun Wang [1], Yahong Fei [3] and Wensheng Qin [4,*]

[1] College of Chemistry Biology and Environment, Yuxi Normal University, Yuxi 653100, China; zxd95@yxnu.edu.cn (X.Z.); lcx@yxnu.edu.cn (C.L.); wanglianchun@yxnu.edu.cn (L.W.)
[2] Food and Bioengineering College, Xuchang University, Xuchang 461000, China
[3] Yuxi Flyingbear Agricultural Development Company Limited, Yuxi 653100, China; feiyahong@gmail.com
[4] Department of Biology, Lakehead University, Thunder Bay, ON P7B 5E1, Canada
* Correspondence: wqin@lakeheadu.ca; Tel.: +1-807-343-8467
† These authors contribute to this article equally.

Received: 14 September 2019; Accepted: 27 November 2019; Published: 29 November 2019

Abstract: Cardiovascular diseases (CVDs) are a major cause of health loss in the world. Prevention and treatment of this disease by traditional Chinese medicine is a promising method. *Centranthera grandiflora* Benth is a high-value medicinal herb in the prevention and treatment of CVDs; its main medicinal components include iridoid glycosides, phenylethanoid glycosides, and azafrin in roots. However, biosynthetic pathways of these components and their regulatory mechanisms are unknown. Furthermore, there are no genomic resources of this herb. In this article, we provide sequence and transcript abundance data for the root, stem, and leaf transcriptome of *C. grandiflora* Benth obtained by the Illumina Hiseq2000. More than 438 million clean reads were obtained from root, stem, and leaf libraries, which produced 153,198 unigenes. Based on databases annotation, a total of 557, 213, and 161 unigenes were annotated to catalpol, acteoside, and azafrin biosynthetic pathways, respectively. Differentially expressed gene analysis identified 14,875 unigenes differentially enriched between leaf and root with 8,054 upregulated genes and 6,821 downregulated genes. Candidate MYB transcription factors involved in catalpol, acteoside, and azafrin biosynthesis were also predicated. This work is the first transcriptome analysis in *C. grandiflora* Benth which will aid the deciphering of biosynthesis pathways and regulatory mechanisms of active components.

Keywords: *Centranthera grandiflora* Benth; transcriptome; catalpol biosynthesis; acteoside biosynthesis; azafrin biosynthesis

1. Introduction

Cardiovascular diseases (CVDs) are an important reason for death in the world which hinder sustainable development of human beings [1]. In China, CVDs were also the leading cause of death due to lifestyle changes, urbanization, and the accelerated process of aging, and the figures have exceeded 42% of all deaths in both rural and urban regions, which was much higher than deaths caused by cancer or any other diseases in 2014 [2]. Traditional Chinese medicine has been used for more than 2000 years and has displayed the explicit role in preventing and treating CVDs, although the detailed pharmacological mechanisms have seldomly been clarified [3]. *Centranthera grandiflora* Benth, also known as broad bean *Ganoderma lucidum*, wild broad bean root, Huaxuedan, Golden Cat's Head, and Xiaohongyao, is a medicinal plant widely used for preventing and treating CVDs among Miao Nationality of Yunnan in China. In taxonomy, it belongs to the Centranthera, Scrophulariaceae family. Distinguished as a rare and endangered medicinal plant, *C. grandiflora* Benth usually grows well with *Cyperus rotundus* and is mainly distributed in Yunnan, Guizhou, and Guangxi in China as well as parts of India, Myanmar, and Vietnam [4–7]. Its roots possess many functions, such as to promote blood circulation, to regulate menstruation, to dispel blood stasis, and to relieve pain, and has known coagulation, antibacterial, and anticancer properties [6,8–10]. Therefore, it is mainly used to treat amenorrhea, dysmenorrhea, metrorrhagia, fall-related injuries, rheumatic bone pain, traumatic hemorrhage, and cardiovascular and cerebrovascular diseases [6,8–10].

So far, studies on this herb have mainly focused on the isolation and identification of its chemical constituents and pharmacological effects, while the discovery of genes related to biosynthesis of active secondary metabolites has not been reported. Azafrin and D-mannitol were first isolated and identified from the roots of *C. grandiflora* Benth in 1984 [8]. Then, aeginetin and azalea were isolated from the roots of *C. grandiflora* Benth, and their coagulation, antimicrobial, and anticancer functions were verified in 2012 [10]. In the same year, nine iridoid glycosides including aucubin, mussaenoside, 8-epiloganin, 8-epiloganic acid, mussaenosidic acid, catalpol, gardoside methyl ester, geniposidic acid, and 6-O-methylaucubin were isolated from roots of *C. grandiflora* Benth [6]. In 2014, another 17 compounds, including six new ones: centrantheroside A to E and neomelasmoside; phenylethanoid glycosides: plantainoside A, calceolarioside A, acteoside, and isoacteoside; monoterpenoid glycosides: melasmoside and rehmaionoside C; Di-O-methylcrenatin; azafrin; β-sitosterol; mannitol; and β-daucosterol were isolated from *C. grandiflora* Benth roots [5,7]. Studies have shown that iridoid glycosides, phenylethanoid glycosides, and azafrin are the main substance bases for their pharmacodynamics [5,7]. In 2017, tissue culture of *C. grandiflora* Benth was also successfully developed [11].

At present, *C. grandiflora* Benth roots sold in public markets are mainly collected from wild resources, while its artificial cultivation has just started [5]. So far, the cost of annual *C. grandiflora* Benth planting is about $0.13 million per hectare and the worth of annual yield is about $0.64 million per hectare [5]. Therefore, to explore the biosynthetic pathways and regulatory mechanisms of the main active ingredients of *C. grandiflora* Benth will lay a scientific foundation for breeding new varieties of this herb and for producing its medicinal chemical constituents by synthetic biology.

Iridoid glycosides belong to monoterpenoids, and their biosynthesis in plants can be divided into three stages. The first stage is precursor formation, which includes the plastidial 2-C-methyl-D-erythritol-4-phosphate (MEP) pathway and the cytoplasmic mevalonate (MVA) pathway to produce isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) [12,13]. The second stage is the formation of a carbon skeleton structure [13–16]. The third stage is the post-modification of terpenoids: hydroxylation, methylation, isomerization, demethylation, glycosylation, etc. [16]. So far, most of the biosynthesis pathways of iridoid glycosides remain unclear. However, the complete catalpol biosynthetic pathway was first elicited in *Picrorhiza kurroa* [17], and it was partially decoded in *Rehmannia glutinosa* [18]. In *P. kurroa*, the catalpol biosynthetic pathway contains 29 steps including 14 steps for the MEP and MVA pathways and 15 steps for the iridoid pathway [17]. As the MEP and MVA pathways has been widely and intensively studied and they are conserved in plants [19], here, we mainly focused on the iridoid pathway. So far, two iridoid pathways including secoiridoid pathway (Route I) and decarboxylated iridoid pathway (Route II) have been reported, and the early enzymatic steps containing geranyl diphosphate synthase (GPPS), geraniol synthase (GES), geraniol 10-hydroxylase (G10H), 8-hydroxygeraniol oxidoreductase (8HGO), iridoid synthase (IS), iridoid oxidase (IO), and UDP-glucosyltransferase (UGT) are common to both pathways [20], has been verified in *Catharanthus roseus* and *P. kurroa* [14,21,22], and proposed in *Gardenia jasminoides* [23,24]. The remaining steps were first deduced by chemical intermediates [20], and then the corresponding enzymes were predicted and discovered by transcriptome analysis [17]. In *P. kurroa*, another seven enzymes containing aldehyde dehydrogenase (ALD), flavanone 3-dioxygenase/hydoxylase (F3D), 2-hydroxyisoflavanone dehydratase (2FHD), deacetoxycephalosporin-C hydroxylase (DCH), uroporphyrinogen decarboxylase/UDP-glucuronic acid decarboxylase (UPD/UGD), and squalene monooxygenase (SQM) have been proposed to catalyze the remaining seven steps in catalpol biosynthesis [17].

Acteoside, belonging to phenylethanoid glycosides, is composed of two parts: caffeoyl CoA and hydroxytyrosol glucoside [25]. Feeding and inhibition experiments showed that hydroxytyrosol glucoside moiety is derived from tyrosine while caffeoyl CoA moiety is derived from phenylalanine via the cinnamate pathway and that both tyrosine and phenylalanine come from the shikimate pathway [26,27]. In *Ole europae* and *R. glutinosa*, phenylalanine is converted into caffeoyl CoA via four enzymes including phenylalanine ammonia-lyase (PAL), cinnamate-4-hydroxylase (C4H), coumarate-3-hydroxylase (C3H), and 4-coumarate-CoA ligase (4CL) [18,25,28]. Simultaneously, tyrosine is transformed into hydroxytyrosol glucoside through two alternative pathways: one is via *L*-dopa, dopamine, and hydroxytyrosol with the enzymes polyphenol oxidase (PPO), tyrosine decarboxylase (TDC), copper-containing amine oxidase (CuAO), alcohol dehydrogenase (ADH), and UGT; the other is via tyramine, tyrosol, and salidroside with the enzymes TDC, CuAO, ADH, UGT, and PPO [18,25,28]. Finally, caffeoyl CoA and hydroxytyrosol glucoside can be converted into acteoside by Shikimate O-hydroxycinnamoyltransferase (HCT) and UGT [18,25]. Recent studies have verified that acteoside possesses pharmacological properties: antioxidant, anti-inflammatory, antidepressant, antitumor, antidiabetes, and hepatoprotection [29–32].

Azafrin, belonging to carotenoid derivative, is one of the most abundant active ingredients in *C. grandiflora* Benth roots and plays an important role in myocardial protection [33]. Carotenoids are ubiquitous pigments in plants, and they confer plants with bright yellows, oranges, and reds [34]. In higher plants, carotenoids are synthesized through isoprene-like pathways in plastids, including condensation, dehydrogenation, cyclization, hydroxylation, and epoxidation reactions, while lycopene acts as an important branch point of both synthesis of α-carotene and β-carotene [35]. In the α-carotene pathway, α-carotene is synthesized by lycopene ε-cyclase (LCY-ε) and lycopene β-cyclase (LCY-β) and is then converted to lutein by ε-hydroxylase (LUT1) and β-cyclohexylase (LUT5) [17,18]. In the β-carotene pathway, LCY-β catalyzes the synthesis of β-carotene, which can be converted into strigolactone, astaxanthin, capsanthin, capsorubin, and violaxanthin under the catalysis of different enzymes, while violaxanthin can be further converted into abscisic acid [36,37]. However, studies have shown that azafrin is an apocarotenoid which is generated by cleavage of carotenoids at the C9′–C10′ [38,39]. In the strigolactone pathway, β-carotene is converted into carlactone through 9-cis-carotene, 10′-apo-β-carotenal by enzymes DWARF27, carotenoid cleavage dioxygenase 7 (CCD7), and CCD8 [39]. The intermediate product 10′-apo-β-carotenal is very similar to azafrin in structure except one terminal carboxyl group and two hydroxyl groups. Therefore, the hypothesis that azafrin is synthesized via 10′-apo-β-carotenal is proposed in this article.

Thus, the aim of this research is to characterize globally for the first time the transcriptomes of the root, stem, and leaf of *C. grandiflora* Benth using the Illumina Hiseq2000. To explore the genes involved in the catalpol, acteoside, and azafrin biosynthesis pathways and regulatory mechanisms, transcripts from leaves, stems, and roots of *C. grandiflora* Benth were screened out, quantified, and annotated. The results obtained here will facilitate further molecular studies in *C. grandiflora* Benth.

2. Results

2.1. Sequencing and Assembly

To figure out which genes are involved in the biosynthesis of active components in *C. grandiflora* Benth, nine sequencing libraries including roots (C_R1, C_R2, and C_R3), stems (C_S1, C_S2, and C_S3), and leaves (C_L1, C_L2, and C_L3) were prepared and sequenced with the Illumina Hiseq2000 platform. As a result, more than 45 million clean reads per library were obtained after cleaning and quality examination. Quality assessments of the sequencing data are shown in Table 1. The error rate of all libraries was 0.03%, while Q20 and Q30 were over 94.99% and 88.32%, respectively, indicating that these data are suitable for further analysis. The raw data from the nine libraries have been deposited into the Short Reads Archive (SRA) database under the accession numbers: SRX6654843–SRX6654851.

Table 1. Quality assessment of the sequencing data.

Sample	Raw Reads	Clean Reads	Clean Bases	Error (%)	Q20 (%)	Q30 (%)	GC (%)
C_R1	49187308	48514118	7.28G	0.03	95.57	89.01	47.36
C_R2	49633082	48674546	7.3G	0.03	97.80	93.85	49.17
C_R3	48837996	47783570	7.17G	0.03	97.94	94.13	48.63
C_S1	55750138	55002348	8.25G	0.03	95.53	89.08	47.82
C_S2	49324046	48163930	7.22G	0.03	97.47	93.33	49.39
C_S3	50001598	48875538	7.33G	0.03	97.63	93.61	49.65
C_L1	50405504	49123820	7.37G	0.03	95.66	89.19	48.69
C_L2	45678578	45075310	6.76G	0.03	97.59	93.45	49.89
C_L3	47606624	46899750	7.03G	0.03	94.99	88.32	49.60

Note: C_R1, C_R2, and C_R3: three root samples; C_S1, C_S2, and C_S3: three stem samples; and C_L1, C_L2, and C_L3: three leaf samples.

The clean reads were combined and assembled by Trinity with min_kmer_cov set to 2 and all other default parameters [40]. Assembled sequences were subjected to cluster using the Trinity algorithm. As a result, 153,300 contigs were clustered into 173,851 trinity components. Each Trinity component contained a set of transcripts derived from the same gene, and a unigene was designated as the longest transcript in each trinity component. A total of 173,851 transcripts and 153,198 genes were assembled, with 69,421 (39.93%) transcripts and 69,419 (45.31%) genes being over 2 Kb in length (Figure 1). The average length of transcripts and genes were 1895 bp and 2115 bp, respectively (Table 2), and the N50 for transcripts and genes were 2902 and 2936 bp, respectively (Table 2).

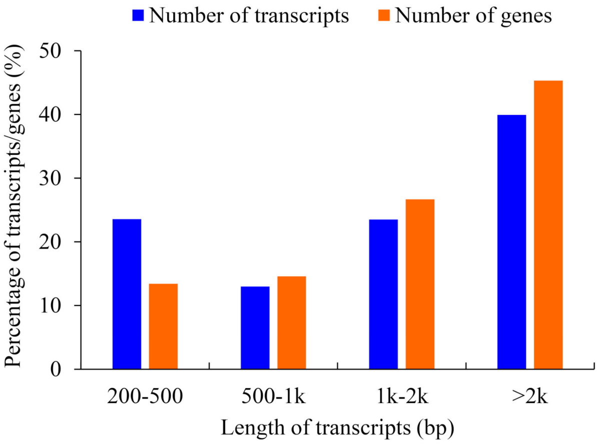


Figure 1. Length distribution frequency of spliced transcripts and deduced genes.

Table 2. Length frequency distribution of the spliced transcripts and genes.

	Min Length	Mean Length	Median Length	Max Length	N50	N90	Total Nucleotides
Transcripts	201	1895	1574	16,816	2902	1089	329,518,919
Genes	201	2115	1824	16,816	2936	1195	323,991,974

2.2. Gene Function Annotation and Classification

All the 153,198 assembled putative unigenes were aligned using the BLAST (Basic Local Alignment Search Tool) program against the seven classic databases including NR (nonredundant protein sequences), NT (Nucleotide collection), PFAM (Protein family), SwissProt, KOG (euKaryotic Orthologous Groups), KEGG (Kyoto Encyclopedia of Genes and Genomes), and GO (Gene Ontology) databases with *e*-value cutoffs of 10^{-5}, 10^{-5}, 10^{-2}, 10^{-5}, 10^{-3}, 10^{-10}, and 10^{-6}, respectively. A total of 26,652 unigenes (17.39%) were annotated to the above seven databases in common, while 132,896 unigenes (86.74%) were annotated in at least one database (Table 3). Among them, 127,767 unigenes (83.39%) showed high similarity with sequences in the NR database (*e*-value = 10^{-5}), 96,216 unigenes (62.80%) matched to protein sequences in NT, and 103,257 unigenes (67.40%) showed homology with known genes in SwissProt. The detailed results are shown in Table 3 and Tables S1–S3. Based on the top-hit species distribution of the homology results against NR database, the highest matches were genes from *Sesamum indicum* (43.77%), followed by *Handroanthus impetiginosus* (22.58%) and *Erythranthe guttata* (13.07%) (Figure 2).

Table 3. Statistical results of gene annotation.

Item	Number of Unigenes (n)	Percentage (%)
Annotated in NR	127,767	83.39
Annotated in NT	96,216	62.80
Annotated in SwissProt	103,257	67.40
Annotated in PFAM	98,364	64.20
Annotated in GO	98,364	64.20
Annotated in KOG	44,170	28.83
Annotated in KEGG	57,190	37.33
Annotated in all Databases	26,652	17.39
Annotated in at least one Database	132,896	86.74
Total unigenes	153,198	100

Note: NR (nonredundant protein sequences), NT (Nucleotide collection), PFAM (Protein family), GO (Gene Ontology), KOG (euKaryotic Orthologous Groups), KEGG (Kyoto Encyclopedia of Genes and Genomes).

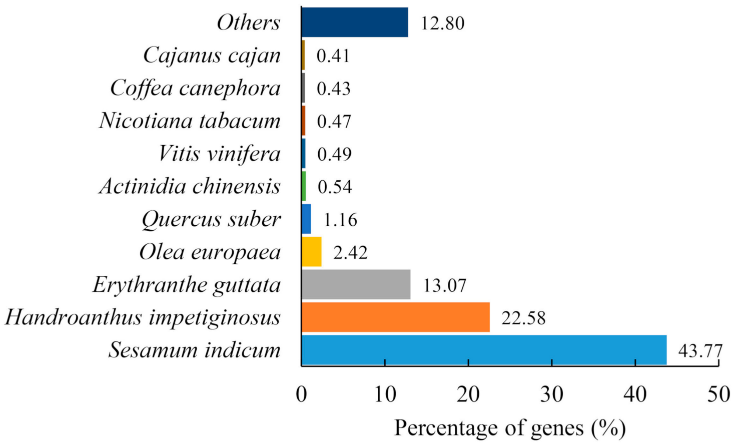


Figure 2. Species distribution of top 10 BLASTx hits against the NR database.

In coding sequence prediction analysis, unigenes were aligned first to the NR and then Swissprot database. If aligned, ORF (open reading frame) information of transcripts was extracted from the alignment results and the sequence of the coding region was translated into amino acid sequences according to the standard codon table. If failed, ESTSCAN (Expression Sequence Tag Scan) software was adopted to predict the ORF of the unigenes. As a result, a total of 157,392 peptides were predicted by BLASTx and the peptide length mainly ranged from 38 to 1059 (Figure 3a) while 35,339 peptides were predicted by ESTSCAN and the peptide length was from 15 to 1092 (Figure 3b).

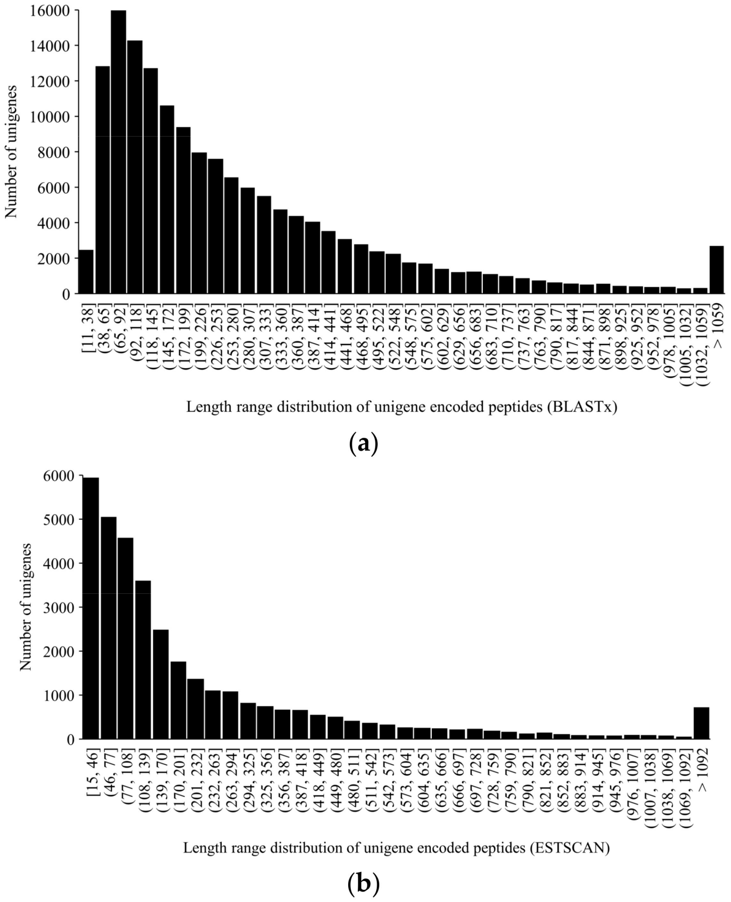


Figure 3. Length range distributions of unigene encoded peptides: (**a**) Peptides predicted by BLASTx searches against NR and Swissprot databases; (**b**) peptides predicted by software ESTSCAN 3.0.3.

To figure out the biological processes that our unigenes are involved in as well as their molecular functions and the cellular environments they reside in, all unigenes were searched against the GO database with software BLAST2GO. Out of 153,198 unigenes, 98,364 (64.20%) were successfully annotated and classified into three GO categories—biological process (BP), cellular component (CC), and molecular function (MF)—and then assigned to 55 functional groups (Figure 4). As shown in Figure 4, assignments which fell under BP (273,598, 47.25%) ranked the highest, followed by CC (175,650, 30.34%) and MF (129,742, 22.41%). Similar to *R. glutinosa* [41] and adventitious roots in *Panax ginseng* [42], "cellular process" (60,529, 61.54%) and "metabolic process" (56,468, 57.41%) were the two most representative subcategories in the BP category, which suggested that lots of important cellular processes and metabolic activities took place in *C. grandiflora* benth. Unlike adventitious roots in *P. ginseng* [42], although unigenes related to "cell" (33,946, 34.51%) and "cell part" (33,946, 34.51%) were dominant in the CC category, the percentages in *C. grandiflora* Benth were far less than in *P. ginseng*, which implied that many tissues and organs in *C. grandiflora* Benth were in construction at a slow speed. In the MF category, the majority of unigenes were involved in "binding" (59,157, 60.14%) and "catalytic activity" (49,089, 49.91%) in *C. grandiflora* Benth, and this was somewhat similar to *R. glutinosa* [41], in which unigenes annotated into "binding" were about 20% more than "catalytic activity", indicating that more catalytic reactions may occur in the form of protein complexes.

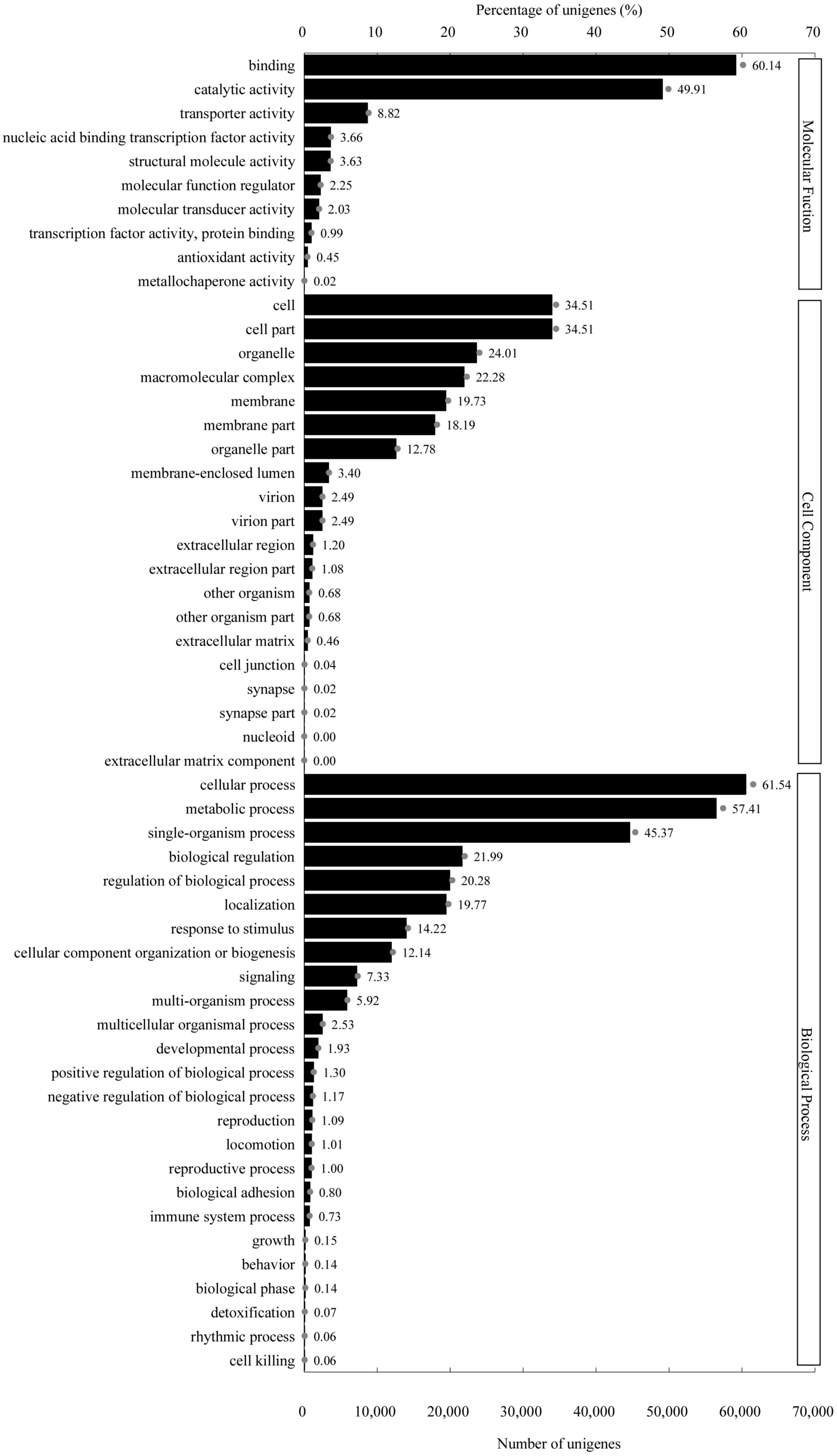


Figure 4. GO classification map: The ordinate represents the next-level GO term of the three GO categories, while the abscissa represents the number of genes annotated into the corresponding term and its proportion of the total number of annotated genes. Three basic categories of GO term, from top to bottom, are the molecular function, cell components, and biological processes.

KOG refers to clusters of orthologous groups from different eukaryotic species, and genes from the same ortholog are assumed to have the same function. To further classify our unigenes, KOG annotation was performed with software diamond. A total of 44,170 unigenes were classified into 26 KOG groups (Figure 5), where the “posttranslational modification, protein turnover, and chaperon” (5516, 12.49%) category accounted for the most frequent group, “general function prediction only” (5445, 12.33%) was the second largest group, and “translation, ribosomal structure, and biogenesis” (4404, 9.97%) and “intracellular trafficking, secretion, and vesicular transport” (3504, 7.93%) were tied for the third largest. In addition, 773 unigenes were assigned to “secondary metabolites biosynthesis, transport, and catabolism”, implying that catalpol, acteoside, and carotenoid biosynthesis may take place in *C. grandiflora* Benth.

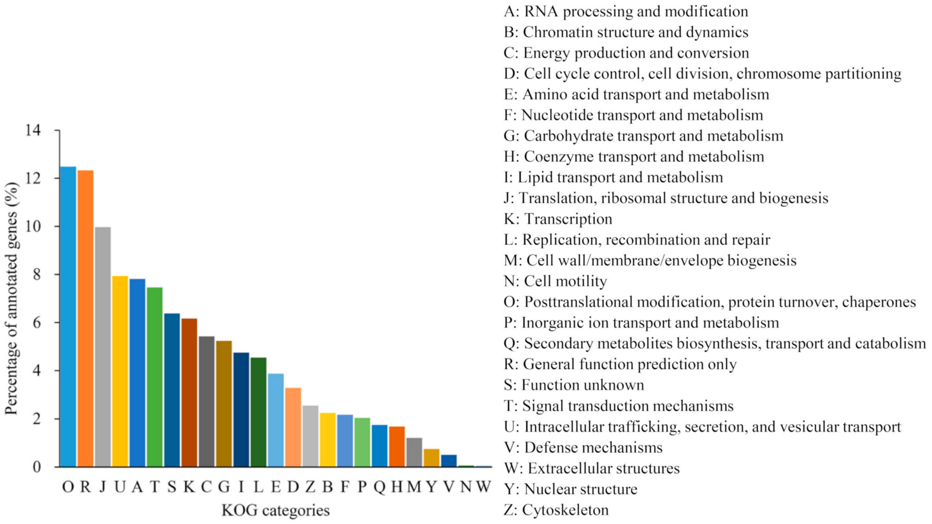


Figure 5. KOG classification map: The abscissa represents KOG groups, while the vertical axis represents the percentage of annotated genes.

KEGG is a database in which gene products and compounds of the cellular metabolic pathways and the functions of these gene products were systematically analyzed. To figure out the active pathways in growing *C. grandiflora* Benth, KEGG annotation of our unigenes were performed with KAAS (KEGG Automatic Annotation Server). A total of 57,190 (37.33%) unigenes were annotated into five categories (level 1; cellular processes, environmental information processing, genetic information processing, metabolism, and organismal systems), 19 subcategories (level 2, Figure 6), and 130 pathways (level 3). Similar to *R. glutinosa* [41], the five pathways with the largest number of genes were “carbohydrate metabolism” (4996, 8.74%), “overview” (3455, 6.04%), “amino acid metabolism”, “lipid metabolism”, and “energy metabolism” in the metabolism category, indicating that primary metabolism was very important to the growth of *C. grandiflora* Benth. In the category of genetic information processing, the two pathways with the largest number of genes were “translation” (5427, 9.49%) and “folding, sorting, and degradation” (4178, 7.31%), indicating that protein biosynthesis and processing were more active in *C. grandiflora* Benth (Figure 6). The numbers of unigenes for “amino acid metabolism”, “metabolism of terpenoids and polyketides”, and “biosynthesis of other secondary metabolites” were 2927, 1164, and 882, respectively. These results indicate that the amino acid pathway and terpenoid pathways were active in growing *C. grandiflora* Benth and that the corresponding genes would be good candidate genes for catalpol, acteoside, and carotenoid biosynthesis.

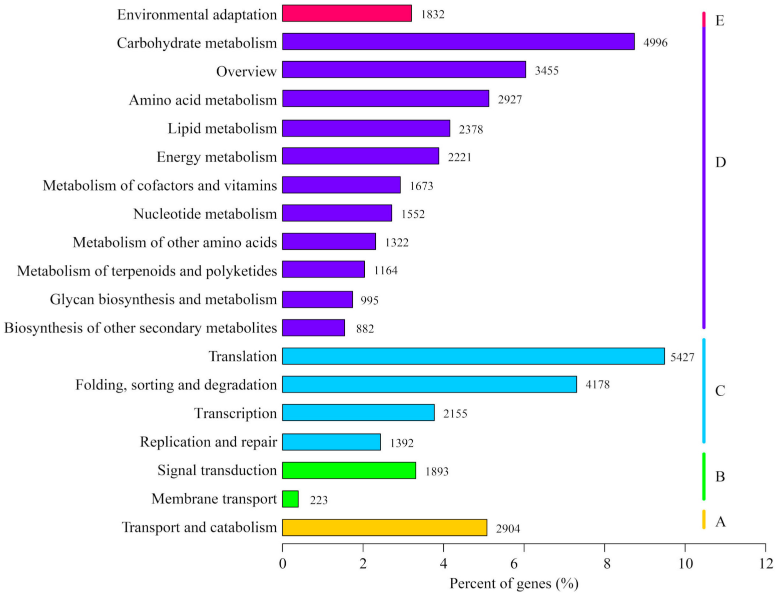


Figure 6. KEGG classification map: The ordinate is the pathway, and the abscissa is the proportion of genes belonging to the corresponding pathway. These genes were divided into five categories: **A**. Cellular Processes; **B**. Environmental Information Processing; **C**. Genetic Information Processing; **D**. Metabolism; and **E**. Organismal Systems.

2.3. Identification of Differentially Expressed Genes (DEGs), GO, and KEGG Enrichment Analysis

Gene expression level which is transformed from read counts using RSEM (RNA-Seq by Expectation-Maximization) software was analyzed with the FPKM (expected number of Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced) method [43]. According to the criteria $p < 0.05$ and $\log_2$(FoldChange) > 1, 14,875 genes (9.71% of all genes) were identified as significant DEGs between leaves and roots, which comprised 8054 upregulated genes (54.14%) and 6821 downregulated genes (45.86%) in leaves (Figure 7a, Table S4). There were 4126 genes (2.69% of all genes) identified as significant DEGs between leaves and stems, which comprised 2251 upregulated genes (54.56%) and 1875 downregulated genes (45.44%) in leaves (Figure 7b, Table S5). A total of 9115 genes (5.95% of all genes) were identified as significant DEGs between stems and roots, which comprised 5290 upregulated genes (58.04%) and 3825 downregulated genes (41.96%) in stems (Figure 7c, Table S6). Using a Venn diagram, we compared these three data sets from different comparison groups (C_L vs. C_R, C_S vs. C_R, and C_L vs. C_S). In all three comparison groups, 829 DEGs were identified as being in common (Figure 7d). Specifically, 4839 DEGs were identified in both "C_L vs. C_S" and "C_S vs. C_R" comparisons; 1918 DEGs were identified in both "C_L vs. C_R" and "C_L vs. C_S" comparisons; and 551 DEGs were identified in both "C_L vs. C_S" and "C_S vs. C_R" comparisons (Figure 7d).

GO and KEGG enrichment analysis on all DEGs were performed to find the enriched pathways. The GO enrichment is shown in Tables S7–S9. In the KEGG enrichment analysis, the top two enriched pathways were flavonoid biosynthesis with 37 DEGs and 68 background unigenes, and flavone and flavonol biosynthesis with 10 DEGs and 14 background unigenes in the C_L vs. C_R comparison (Figure 8a, Table S10). In the C_L vs. C_S comparison, the top two pathways were flavone and flavonol biosynthesis with 5 DEGs and 14 background unigenes, and the stilbenoid, diarylheptanoid, and gingerol biosynthesis with 11 DEGs and 48 background unigenes (Figure 8b, Table S11). Finally, in the L_S vs. L_R comparison, the top two pathways were stilbenoid, diarylheptanoid, and gingerol biosynthesis with 15 DEGs and 48 background unigenes, and flavonoid biosynthesis with 3 DEGs and 11 background unigenes (Figure 8c, Table S12). Notably, pathways for both phenylpropanoid biosynthesis and carotenoid biosynthesis were enriched in all three comparisons. The top 20 KEGG enrichment pathways are shown in Figure 8.

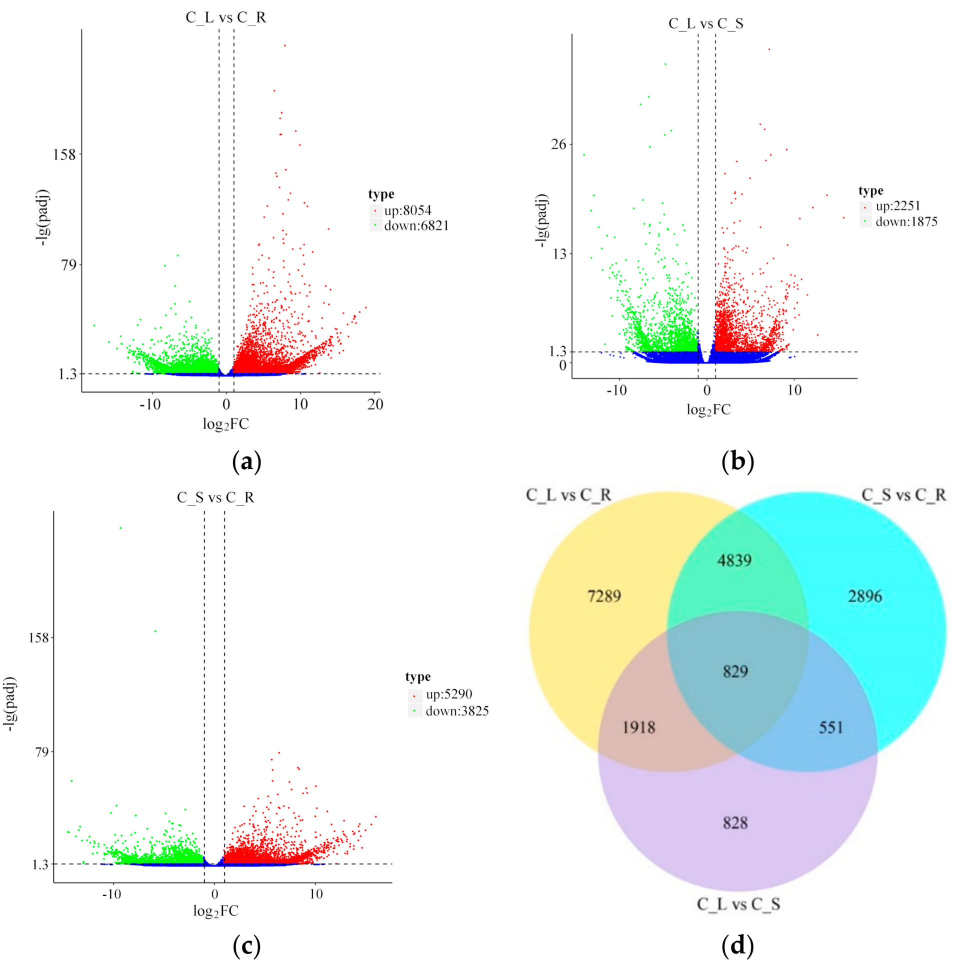


Figure 7. Differentially expressed genes (DEGs) in different comparisons. Volcano plots of the DEGs in different comparisons: The red dots mean significantly upregulated genes, and the green dots represent significantly downregulated genes. The black dots represent non-DEGs. (**a**) C_L vs. C_R volcano; (**b**) C_L vs. C_S volcano; and (**c**) C_S vs. C_R volcano. (**d**) Venn diagram of DEGs in different comparisons: All DEGs are clustered into three comparison groups represented by three circles. Overlapping parts of the different circles represent the number of DEGs in common among those comparison groups.

2.4. Biosynthetic Genes of the Terpenoid Backbone and Catalpol in *C. grandiflora* Benth

Terpenoids are produced from the universal precursor IPP (a five-carbon building material) and its isomer DMAPP [44]. In plants, IPP is synthesized via the cytoplasmic MVA pathway from acetyl-CoA and through the plastidial MEP pathway from glyceraldehyde 3-phosphate and pyruvate; IPP isomerase (IDI) catalyzes the interconversion between IPP and DMAPP [44] (Figure 9a). When examining the annotation of unigenes against the KEGG database, 239 unigenes were assigned to the terpenoid backbone biosynthesis pathway, including 74 unigenes encoding 6 enzymes in the MVA pathway and 165 unigenes encoding 8 enzymes in the MEP pathway (Table 4). Among these genes, the largest number is *DXS* (67), followed by *IDI* (28), *AACT* (24), and *CMK* (24), and the lowest number is *MCS* (1). Transcriptome profiling data showed that the MEP pathway is more active in leaves, while the MVA pathway is more active in stems due to the high expression levels of corresponding pathway genes (Figure 9b).

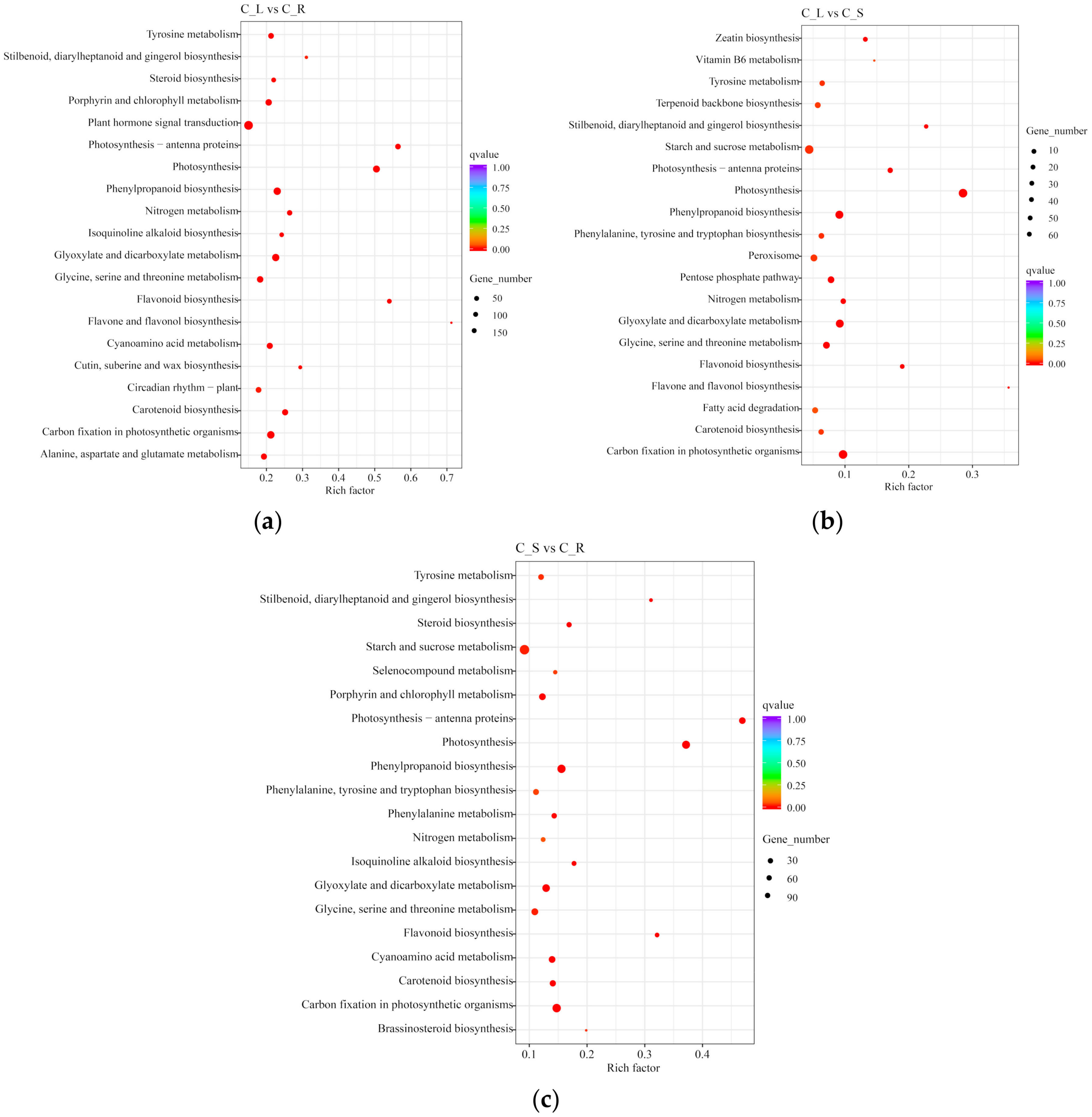


Figure 8. Top 20 of KEGG pathway enrichment of DEGs: The y-axis indicates the pathway name, and the x-axis indicates the enrichment factor corresponding to the pathway. The q-value is represented by the color of the dot. The number of DEGs is represented by the size of the dots. (**a**) C_L vs. C_R; (**b**) C_L vs. C_S; and (**c**) C_S vs. C_R.

Catalpol, belonging to iridoid glucoside, is usually found in Scrophulariaceae plants [6,45,46], and iridoid glucosides are derived from MEP and MVA pathways [41]. Based on feeding experiments with isotope labeling and transcriptome analysis, the draft biosynthesis pathway of catalpol was first proposed in *R. glutinosa* in 2012 [47,48]. Then, the complete pathway of catalpol was clarified for the first time in *P. kurroa* in 2015 [17].

According to the KEGG and Swissprot annotation, a total of 368 unigenes were assigned to the catalpol biosynthetic pathway, with 60 unigenes upregulated and 39 unigenes downregulated in leaf vs. root (Figure 9a and Table 4). The unigenes encoding 13 enzymes involved in catalpol biosynthesis are listed in Table 4. Among these genes, the largest number was *ALDH* (76), followed by *8HGO* (53), *IO* (44), *GPPS* (32), and UGT (30), and the lowest number was *F3D* (2) (Table 4). Like the terpenoid backbone pathway, most genes in the catalpol biosynthesis pathway possess at least two unigenes, displaying the redundancy of the plant genes and adding the difficulty of deciphering the pathway

(Table 4). In our transcriptome, unigenes of catalpol pathways are more abundant in leaves, as revealed by much higher expression level of *GES*, *G10H*, *IS*, *IO*, and *F3D* in leaves than in roots (Figure 9b, Table 4). It is worth noting that there were only two *F3D* genes in our transcriptome and that it was only expressed in leaves and stems, but not roots, which indicated that the catalpol biosynthesis was active in aboveground growth at this developmental stage (Figure 9a, Table 4). Therefore, while it is the roots of *C. grandiflora* Benth that are used as medicinal materials, our results imply that the catalpol is first synthesized in the leaves and then transported and stored in the roots. Furthermore, *DCH* gene functioning in the conversion of deoxygeniposidic acid to geniposidic acid was not found in our transcriptome, which may be a result of low expression or low homology with the known *DCH* genes.

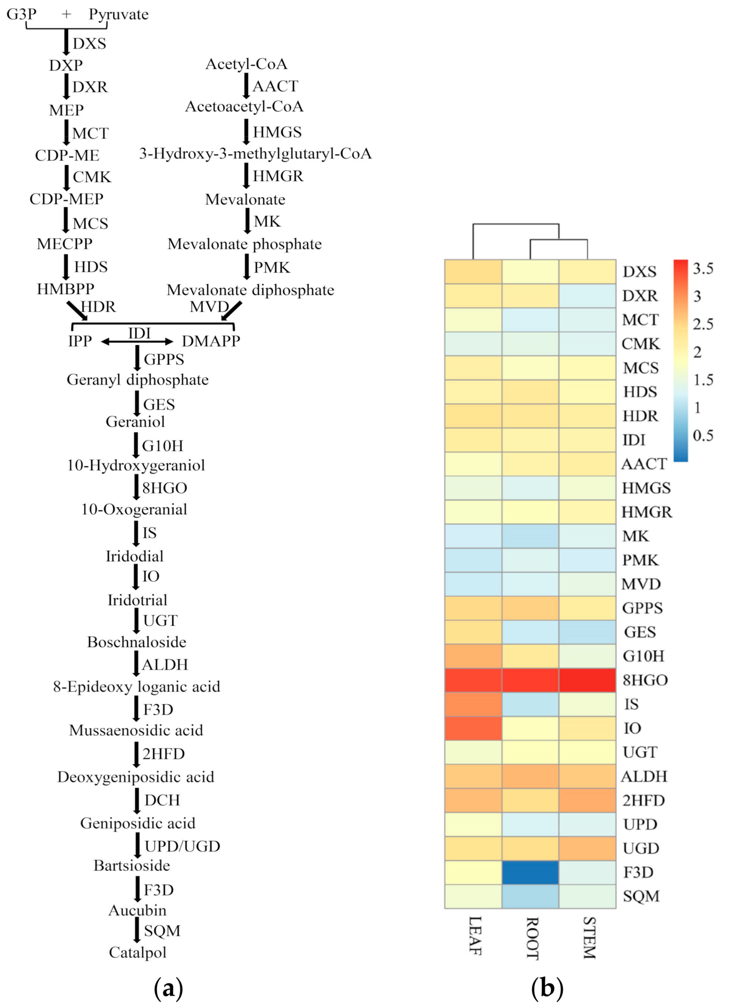


Figure 9. Expression of unigenes in the putative pathway of terpenoid backbone and catalpol biosynthesis in *C. grandiflora* Benth: (**a**) Proposed biosynthetic pathway of terpenoid backbone and catalpol. (**b**) Heatmap based on the expression level of unigenes involved in terpenoid backbone and catalpol biosynthesis across three tissues in *C. grandiflora* Benth. The expression level is the sum of all the unigenes for each gene, and $\log_{10}$(sum(FPKM)+1) was used to plot the heatmap. Candidate unigenes were selected according to the annotation. Abbreviations: G3P, Glyceraldehyde 3-phosphate; DXP, 1-deoxy-D-xylulose-5-phosphate; MEP, 2-C-methyl-D-erythritol 4-phosphate; CDP-ME, 4-(Cytidine 5′-diphospho)-2-C-methyl-D-erythritol; MECPP, 2-C-methyl-D-erythritol-2,4-cyclodiphosphate; HMBPP, 1-hydroxy-2-methyl-2-butenyl 4-diphosphate; IPP, isopentenyl pyrophosphate; DMAPP, dimethylallyl pyrophosphate.

Table 4. Putative genes of the mevalonate (MVA), 2-C-methyl-D-erythritol-4-phosphate (MEP), and catalpol biosynthesis pathways.

Pathway	Gene	Gene Name	EC	Number	Upregulated ($\log_2$(FC) > 1, L vs. R)	Downregulated ($\log_2$(FC) > 1, L vs. R)
MVA	*AACT*	acetyl-CoA C-acetyltransferase	2.3.1.9	24		2
	HMGS	hydroxymethylglutaryl-CoA synthase	2.3.3.10	8	2	
	HMGR	hydroxymethylglutaryl-CoA reductase	1.1.1.34	9		
	MK	mevalonate kinase	2.7.1.36	6	1	
	PMK	phosphomevalonate kinase	2.7.4.2	21		3
	MVD	diphosphomevalonate decarboxylase	4.1.1.33	6		
MEP	*DXS*	1-deoxy-D-xylulose-5-phosphate synthase	2.2.1.7	67	12	
	DXR	1-deoxy-D-xylulose-5-phosphate reductoisomerase	1.1.1.267	10		
	MCT	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	2.7.7.60	3	1	
	CMK	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	2.7.1.148	24		
	MCS	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	4.6.1.12	1	1	
	HDS	(*E*)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase	1.17.7.1/1.17.7.3	21		
	HDR	4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase	1.17.7.4	11		
	IDI	isopentenyl pyrophosphate isomerase	5.3.3.2	28	2	
Catalpol	*GPPS*	geranyl diphosphate synthase	2.5.1.1	32	5	6
	GES	geraniol synthase	3.1.7.11	9	3	
	G10H	geraniol 10-hydroxylase	1.14.13.152	14	8	2
	8HGO	8-hydroxygeraniol oxidoreductase	1.1.1.324	53	13	11
	IS	iridoid synthase	1.3.1.99	5	3	
	IO	iridoid oxidase		44	3	2
	UGT	UDP-glucosyl transferase	2.4.1.	22	4	4
	ALDH	aldehyde dehydrogenase	1.2.1.3	76	7	10
	F3D	flavanone 3-dioxygenase	1.14.11.9	2	1	
	2HFD	2-hydroxyisoflavanone dehydratase	4.2.1.105	10	2	
	UPD	uroporphyrinogen decarboxylase	4.1.1.37	23	3	
	UGD	UDP-glucuronic acid decarboxylase	4.1.1.35	70	3	4
	SQM	squalene monooxygenase	1.14.13.132	8	5	

Note: FC represents fold change.

2.5. Biosynthetic Genes of Acteoside in C. grandiflora Benth

Studies have shown that acteoside is widely distributed in more than 150 plant species and has medicinal properties including antioxidant, anti-inflammation, anti-nephritis, cell regulation, hepatoprotection, immunoregulation, and neuroprotection [49]. Upstream regions of the acteoside biosynthetic pathway, including the phenylalanine-derived pathway and tyrosine-derived pathway, was first clarified in *Olea europaea* using feeding experiments, while the downstream is largely unknown [26]. The downstream region was partially deciphered with elicitor inducing and transcriptome sequencing in *R. glutinosa* [18,28].

Based on the KEGG annotation in this study, a total of 213 unigenes were assigned to the acteoside biosynthetic pathway, with 40 unigenes significantly upregulated and 16 unigenes significantly downregulated in leaves vs. roots (Figure 10a). The unigenes encoding key enzymes involved in acteoside biosynthesis are listed in Table 5. Among these genes, the largest number was *CuAO* (53), followed by *4CL* (35), *ADH* (26), and *UGT* (22), and the lowest number was *HCT* (6) (Table 5). In the DEGs analysis, four genes including *PAL*, *C4H*, *C3H*, and *4CL* were upregulated in leaves and stems compared with roots (Figure 10b, Table 5), which implies that the phenylalanine-derived pathway is active in aerial parts.

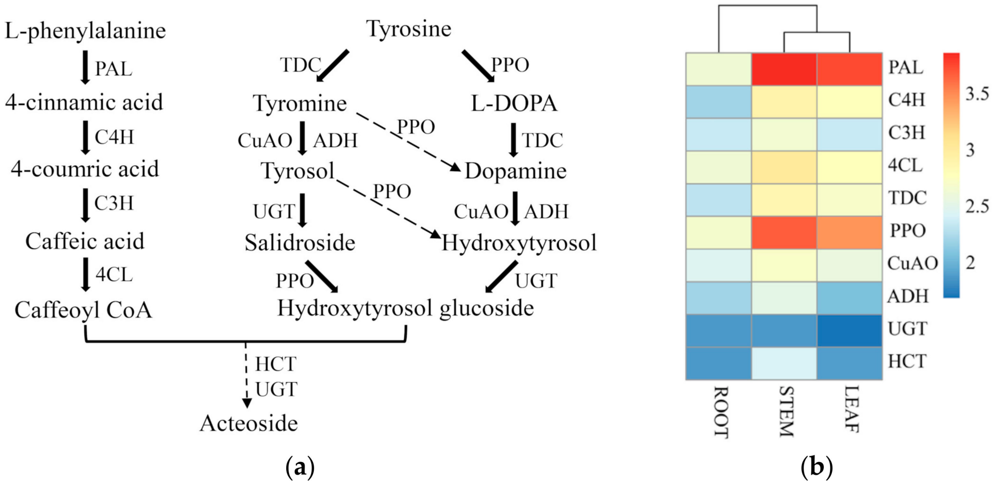


Figure 10. Expression of unigenes in putative pathways of acteoside biosynthesis in *C. grandiflora* Benth: (**a**) Proposed biosynthetic pathway of acteoside. The solid arrow represents the known steps, and the dashed arrows denote the putative steps. (**b**) Heatmap based on the expression level of unigenes involved in acteoside biosynthesis across three tissues: Candidate unigenes were selected according to the annotation.

Table 5. Putative genes of acteoside biosynthesis pathways.

Gene	Gene name	EC	Number	Upregulated ($\log_2$(FC) > 1, L vs. R)	Downregulated ($\log_2$(FC) > 1, L vs. R)
PAL	phenylalanine ammonia-lyase	4.3.1.24	19	4	1
C4H	cinnamate-4-hydroxylase	1.14.14.91	12	5	
C3H	coumarate-3-hydroxylase	1.14.14.96	10	3	1
4CL	4-coumarate-CoA ligase	6.2.1.12	35	9	
TDC	tyrosine decarboxylase	4.1.1.25	19	3	
CuAO	copper-containing amine oxidase	1.4.3.21	53	4	2
ADH	alcohol dehydrogenase	1.1.1.1	26	3	3
UGT	UDP-glucose glucosyltransferase	2.4.1.35	22	4	4
PPO	polyphenol oxidase	1.14.18.1	11	4	5
HCT	Shikimate O-hydroxycinnamoyltransferase	2.3.1.133	6	1	

2.6. Biosynthetic Genes of Azafrin in C. grandiflora Benth

Recent studies have shown that azafrin can significantly improve myocardial contractile function during myocardial ischemia via activation of the Nrf2-ARE (Nuclear factor-erythroid 2-related factor-Antioxidant Response Element) pathway in rats [7,33]. So far, biosynthesis and chemical synthesis pathways of azafrin are still unknown. Azafrin is a derivative of carotenoid, a tetraterpenoid compound [38]. In higher plants, carotenoids are manufactured in plastid with IPP generated by the MEP pathway [50]. The putative carotenoid biosynthesis pathway, including the MEP part, lutein branch, strigolactone branch, capsanthin/capsorubin branch, abscisic acid branch, and without the azafrin pathway, has been established in plants (Figure 11a) [35,51]. However, studies have shown that the substrate β-carotene can be directly converted into 10′-apo-β-carotenal and ionone by β-carotene-9′,10′-oxygenase (BCO2) in non-plants or can be indirectly converted into 10′-apo-β-carotenal and ionone by DWARF27 and carotenoid cleavage dioxygenases 7 (CCD7) in plants [39,52]. The differences between azafrin and 10′-apo-β-carotenal are one terminal carboxyl group and two hydroxyl groups in the cyclohexane skeleton. From the aspect of biochemistry, acetaldehyde dehydrogenase (ALDH) can transform aldehyde into carboxylic acid and the cytochrome P450 monooxygenases (CYP450s) are capable of inserting oxygen atoms into inert hydrophobic molecules to make them more hydrophilic [53]. There is also a report that post-modification of terpenoid derivatives is mostly initiated by oxidation and that most of them are catalyzed by CYP450s and then other post-modification reactions are carried out on the basis of oxidation products [54]. Then, we hypothesize that azafrin is produced in two continuous steps: 10′-apo-β-carotenal is first oxidized by ALDH and then is hydrolyzed by CYP450. The detailed biosynthetic pathway of azafrin is shown in Figure 11a.

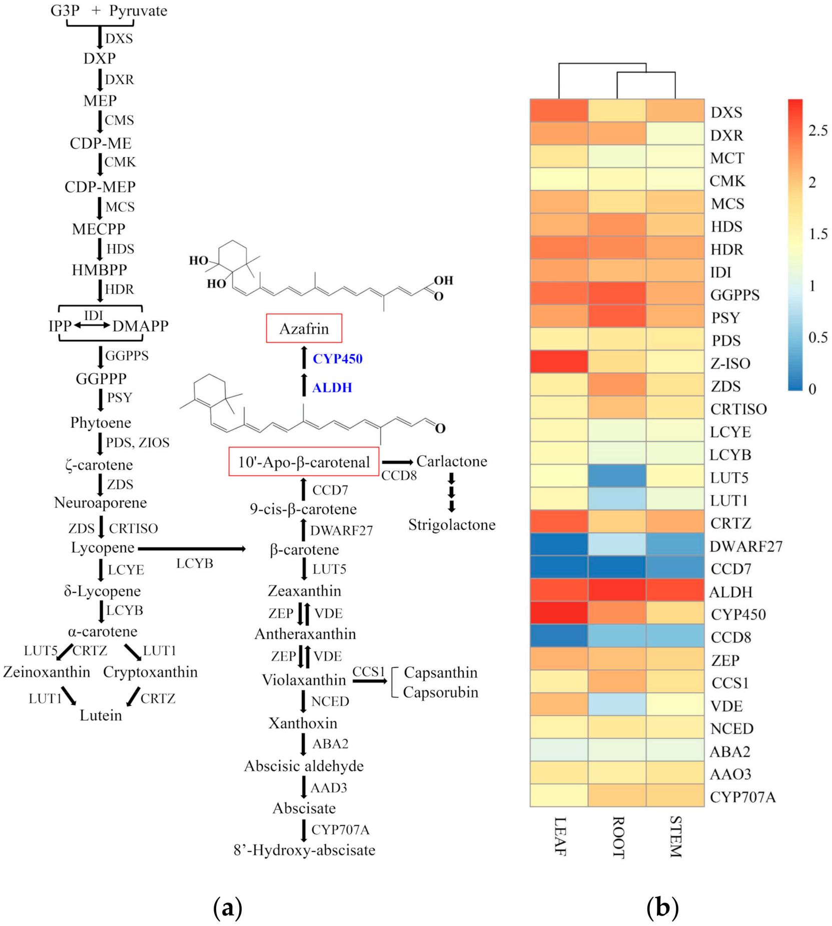


Figure 11. The putative carotenoid biosynthesis pathway and the heatmap of corresponding genes in *C. grandiflora* Benth: (**a**) Proposed biosynthetic pathway of carotenoid. Here, the hypothesis that 10′-apo-β-carotenal can be converted into azafrin by ALDH and CYP450 is proposed. (**b**) Heatmap based on the expression level of unigenes involved in carotenoid biosynthesis across three tissues. For CYP450, only genes of Log$_2$(FC) > 10 (leaf vs. root) were selected for map. Candidate unigenes were selected according to the annotation.

Based on the KEGG annotation and NR annotation in this study, a total of 356 unigenes were correlated with the carotenoid biosynthesis, of which 161 unigenes were assigned to the azafrin biosynthetic pathway with 20 unigenes upregulated and 33 unigenes downregulated in leaves vs. roots (Table 6). For the MEP portion, it was active in leaves, stems, and roots in general as it is known to provide the universal precursor for the terpenoids (Figure 11b). For the lutein pathway, it was more active in leaves and stems while somewhat inactive in roots due to the low expressions of *LUT5* and *LUT1* genes (Figure 11b). For the azafrin and strigolactone branch, it was slightly active in stem, as neither *DWARF27* and *CCD7* were expressed in leaves nor *CCD7* were expressed in roots (Figure 11b). For the capsanthin/capsorubin and abscisic acid branch, it is more active in leaves and stems and somewhat blocked by the low expression of the *LUT5* gene (Figure 11b). What should be noted here is that there is only one gene encoding the CCD7 enzyme, which may be a rate-limiting enzyme (Table 6).

Table 6. Putative genes of carotenoid biosynthesis pathways.

Gene	Gene Name	EC	Number	Upregulated ($\log_2$(FC) > 1, L vs. R)	Downregulated ($\log_2$(FC) > 1, L vs. R)
GGPPS	geranylgeranyl diphosphate synthase	2.5.1.29	15	3	6
PSY	phytoene synthase	2.5.1.32	9	1	2
PDS	phytoene desaturase	1.3.5.5	7	2	1
Z-ISO	zeta-carotene isomerase	5.2.1.12	3		2
ZDS	zeta-carotene desaturase	1.3.5.6	26	1	4
crtISO	carotenoid isomerase	5.2.1.13	10		2
LCYE	lycopene epsilon-cyclase	5.5.1.18	49	3	1
LCYB	lycopene beta-cyclase	5.5.1.19	2	1	
LUT5	beta-ring hydroxylase	1.14.-.-	18	8	
LUT1	carotenoid epsilon hydroxylase	1.14.99.45	14	2	
CRTZ	beta-carotene 3-hydroxylase	1.14.13.129	11	6	5
DWARF27	beta-carotene isomerase	5.2.1.14	2		1
CCD7	9-cis-beta-carotene 9′,10′-cleaving dioxygenase	1.13.11.68	1		
ALDH	aldehyde dehydrogenase	1.2.1.3	76	7	10
CYP450	cytochrome P450		10	5	5
CCD8	carlactone synthase	1.13.11.69	11		
ZEP	zeaxanthin epoxidase	1.14.15.21	17	1	4
CCS1	capsanthin/capsorubin synthase	5.3.99.8	4		1
VDE	violaxanthin de-epoxidase	1.23.5.1	12	3	
NCED	9-cis-epoxycarotenoid dioxygenase	1.13.11.51	33	6	7
ABA2	xanthoxin dehydrogenase	1.1.1.288	5	1	
AAD3	abscisic-aldehyde oxidase	1.2.3.14	11		
CYP707A	(+)-abscisic acid 8′-hydroxylase	1.14.14.137	10	4	6

2.7. Identification of Transcription Factors (TFs)

Transcription factors can activate or inhibit the expression of functional genes in the biosynthetic pathway of plant metabolites, thereby effectively regulating the synthesis and accumulation of secondary metabolites. According to gene sequence alignment to the PFAM database, referring to the Hidden Markov Model files of various TFs, the HMMER3.0 software was used to search the transcriptome database of *C. grandiflora* Benth. The results showed that, in our transcriptome, 4888 unigenes were annotated as TFs belonging to 78 categories. The top three TFs with the largest numbers were MYB (avian myeloblastosis viral oncogene homolog, 356, accounting for 7.28%), WRKY (WRKY domain-containing protein, 301, accounting for 6.16%), and orphans (234, accounting for 4.79%), followed by HB (homeobox, 223, accounting for 4.56%), C3H (Cys3His zinc finger domain-containing protein, 209, accounting for 4.28%), and bHLH (basic Helix-Loop-Helix, 201, accounting for 4.11%) (Figure 12, Table 7, and Table S13). There were also TFs ERF and bZIP (basic region-leucine zipper, Table 7). Among these TFs, most were expressed in both root and leaf tissues, with 121 and 132 showing significantly upregulation and downregulation in leaves, respectively (Table 7).

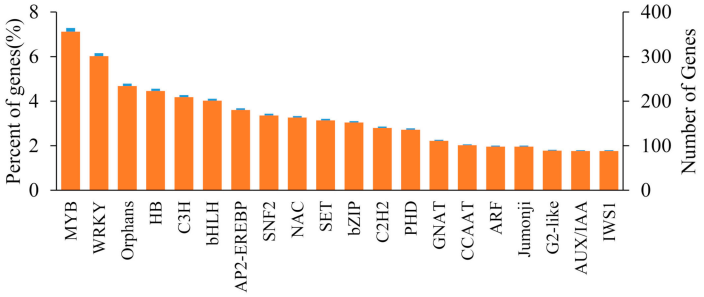


Figure 12. Top 20 transcription factors in *C. grandiflora* Benth.

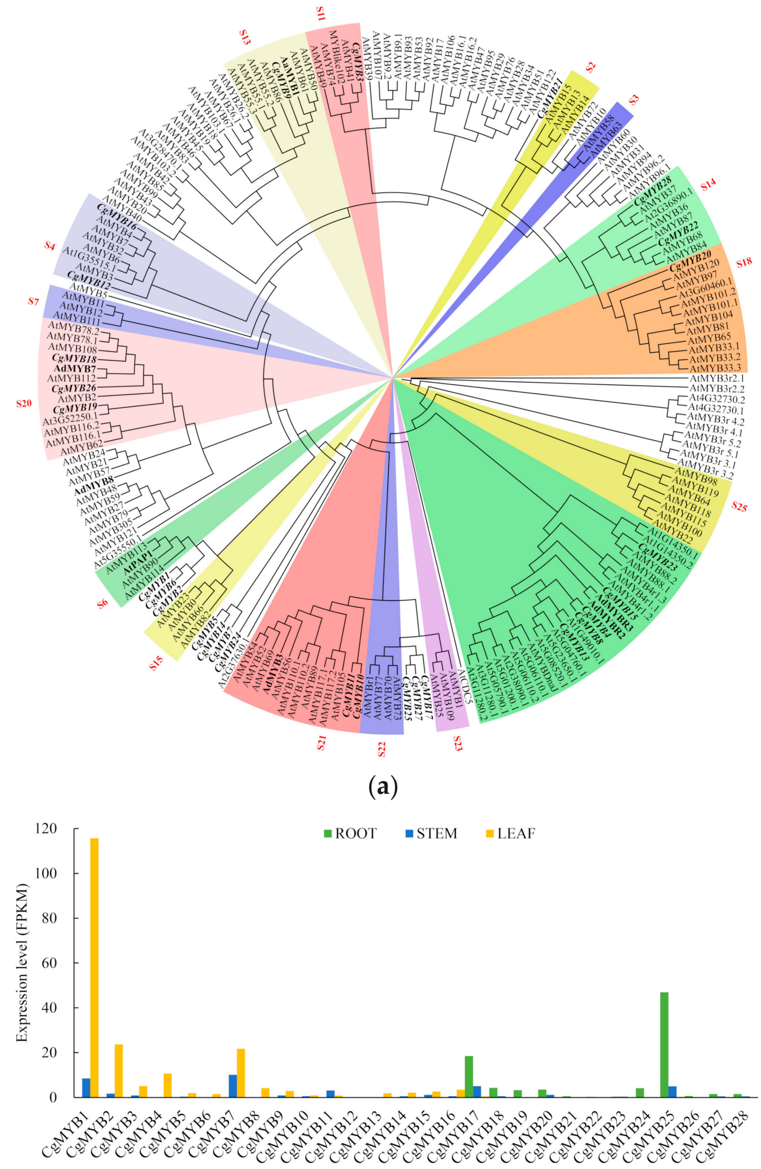
Table 7. Summary of transcription factor unigenes of *C. grandiflora* Benth.

TF Family	Number of Genes Detected	UpRegulated in Leaves ($\log_2$(FC) > 2)	Upregulated in Roots ($\log_2$(FC) > 2)
MYB	356	25	21
WRKY	301	16	44
Orphans	234	14	11
HB	223	25	4
C3H	209	6	7
bHLH	201	21	9
ERF	180	9	21
bZIP	152	5	15
Total	1856	121	132

Note: TF (Transcription Factor), MYB (avian myeloblastosis viral oncogene homolog), WRKY (WRKY domain-containing protein), HB(homeobox), C3H(Cys3His zinc finger domain-containg protein), bHLH (basic Helix-Loop-Helix), ERF(Ethylene Responsive Factor), bZIP (basic region-leucine zipper).

Studies have also shown that the active components of medicinal plants are regulated by many TFs and that the number of genes regulated by a specific TF varies widely. There may be even crosstalk between regulations. In *Artemisia annua*, only AaHD1 (Homeodomain-leucine zipper) and AaGSW1 (Glandular trichome-Specific WRKY 1) can activate transcription of the *CYP71AV1* gene [55,56]; AaWRKY1, AabHLH1, and ERF (Ethylene Response Factor) TFs including AaTAR1 (Trichome and Artemisinin Regulator 1), AaERF1, AaERF2, and AaORA (Octadecanoid-derivative Responsive *Apetala*2 domain) can activate the transcription of both the amorpha-4,11-diene synthase (*ADS*) gene and the *CYP71AV1* gene and then facilitates artemisinin biosynthesis [57–61]; AabZIP1 is responsible for the activation of the *ADS*, *CYP71AV1*, and *AaGSW1* genes [62], while AaMYC2 is responsible for the *CYP71AV1*, *DBR2* (*Double-Bond Reductase* 2), and *AaGSW1* genes [63,64]; and AaMYB2 may regulate the *ADS*, *CYP71AV1*, *DBR2*, and *ALDH1* (*Aldehyde Dehydrogenase 1*) genes [65]. All the abovementioned TFs were found in our transcriptome (Table 7).

In order to figure out which TFs are involved in catalpol, acteoside, and carotenoid biosynthesis in *C. grandiflora* Benth, MYB TFs of which the $\log_2$(FC) > 4 were selected for performing phylogenetic analysis with 168 MYBs from *Arabidopsis thaliana*. As a result, a total of 28 MYBs, including 16 upregulated and 12 downregulated, were screened out in leaf vs. root. In kiwifruit, *AdMYB7*, *AdMYB8*, *AdMYBR2*, and *AdMYBR3* play important roles in regulating carotenoid accumulation in tissues through transcriptional activation of metabolic pathway genes [35,66]. In our analysis, CgMYB18, CgMYB26, CgMYB19, and AdMYB7 were all clustered into the S20 subgroup, while AdMYB8 was near the S20; CgMYB15, CgMYB4, CgMYB8, CgMYB13, AdMYBR2, and AdMYBR3 were in the same clade; and *CgMYB18*, *CgMYB26*, *CgMYB19*, *CgMYB15*, *CgMYB4*, *CgMYB8*, and *CgMYB13* were all upregulated in the root, which indicates that they may regulate carotenoid biosynthesis in the roots of *C. grandiflora* Benth. (Figure 13a). In *A. annua*, overexpression of *AaMYB1* exclusively in trichomes or in whole plants both increased the expression of the *FDS* (*farnesyl diphosphate synthase*), *ADS*, *CYP71AV1*, *DBR2*, and *ALDH1* genes and increased the accumulation of artemisinin [67]. CgMYB9 and AaMYB1 were clustered into S13 subgroup and *CgMYB9* was upregulated in leaves, which showed that it may be a candidate regulatory gene in catalpol and carotenoid biosynthesis [66]. Overexpression of *AtPAP1* (*Production of Anthocyanin Pigment1*) in rose plants enhanced production of phenylpropanoid and terpenoid scent compounds by transcriptional activation of their respective pathway genes [68], and AtPAP1, AtMYB90 (AtPAP2), AtMYB113, and AtMYB114 all regulated anthocyanin biosynthesis [69]. In our data, CgMYB1, CgMYB2, CgMYB6, AtPAP1, and AtMYB90 were all in the S6 subgroup and *CgMYB1*, *CgMYB2*, and *CgMYB6* were all significantly upregulated in the leaves (Figure 13b), suggesting that they are candidate regulatory genes in catalpol, acteoside, and carotenoid biosynthesis.



(**a**)

(**b**)

Figure 13. Phylogenetic analysis and expression level of MYBs from *C. grandiflora* Benth: (**a**) Phylogenetic analysis of CgMYBs. Amino acid sequences were aligned using the ClustalX2 program, and evolutionary distances were calculated using phyML software with the maximum likelihood statistical method. The sequences of *C. grandiflora* Benth are listed in Table S14. The sequences of *Arabidopsis thaliana* come from PLANTTFDB (https://planttfdb.cbi.pku.edu.cn), while that of *Artemisia annua* and *Actinidia deliciosa* come from NCBI (https://www.ncbi.nlm.nih.gov). (**b**) Expression level of CgMYBs: Expression level for each gene is represented by the average RPKM in roots, stems, and leaves.

2.8. Expression Correlation Analysis of Selected Genes

To verify our transcriptome results, five terpenoid-related genes including three upregulated genes (*MCS*, *GES*, and *IS*) and two downregulated genes (*8HGO* and *HMGR2*) in leaf vs. root were selected for correlation analysis. All the selected genes possessed the same expression trend although the expression levels varied between RNA-Seq and qRT-PCR, especially for the *8HGO* and *HMGR2* genes (Figure 14a). The overall correlation coefficient was about 0.84, which indicates that our transcriptome is valid (Figure 14b).

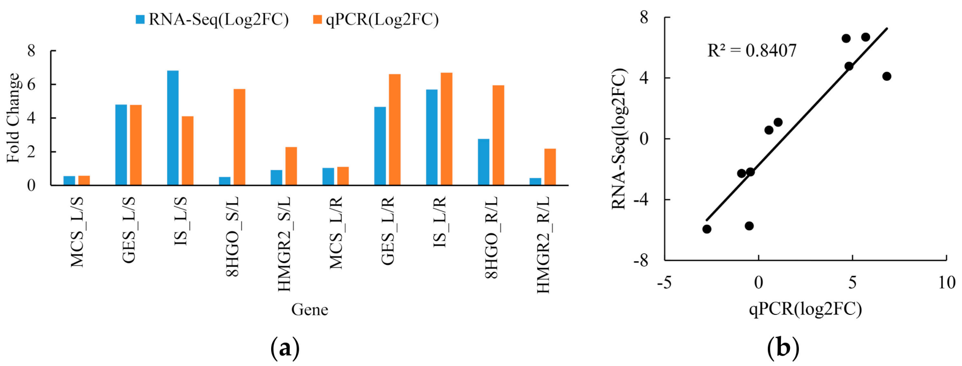


Figure 14. The correlation analysis of gene expression pattern between RNA-Seq and qPCR in roots, stems, and leaves in *C. grandiflora* Benth. (**a**) Expression patterns of *MCS*, *GES*, *IS*, *8HGO* and *HMGR2* gene by RNA-Seq and qPCR; (**b**) Correlation analysis of gene expression between RNA-Seq and qPCR. Each qPCR was biologically repeated three times.

3. Discussion

Cardiovascular diseases (CVDs) remain a major cause of health loss for all regions of the world in the past 25 years [70]. In China, the incidence of CVDs is continuously rising and will keep an upward trend in the next decade [2]. Therefore, to find the herbs with effective treatment of CVDs is imminent. *C. grandiflora* Benth is one of the most precious herbs in the area of Miao Nationality in Yunnan, China. It is widely used in folk medicine because of its multifunctional medicinal values, especially in the aspect of the prevention and treatment of CVDs [7]. Although it has been collected in the Chinese Materia Medica, up to now, it has not been included in the Chinese Pharmacopoeia because of limited researches [9]. The current situation is high market prices and overexploitation of wild resources, which has not only prevented the herbal medicine from being widely used but has destroyed species diversity. Synthetic biology will provide solutions for the abovementioned problems through the biosynthetic pathway elucidations of the main pharmacodynamic components.

So far, de novo transcriptome analysis is an important method in gene discovery of biosynthesis pathways, especially for species without reference genomes [13]. In this research, the transcriptomes of three tissues with three biological repeats were sequenced by illumine Hiseq2000, and 438,112,930 clean reads were assembled into 173,851 transcripts and 153,198 unigenes. This suggests that one gene may have different transcripts which may come from variable splicing, alleles, different copies of the same gene, homologs, orthologs, etc. The mean length of transcripts and genes were 1895 bp and 2115 bp, respectively, and the N50 of the transcripts and genes were 2902 and 2936 bp, respectively (Table 2), which were higher than that in *Dendrobium huoshanense*, *Persea Americana*, and *R. glutinosa* [18,71,72]. These results implied that our assembly quality was suitable for subsequent analyses. In the species distribution analysis of unigenes, more than 43.77% of unigenes were matched to *Sesamum indicum* (Figure 2), which is similar to *R. glutinosa*, a plant of Scrophulariaceae; these results implied that they shared the closer genetic relationship, similar chemical substances, and similar biosynthetic pathways. Catalpol, acteoside, and azafrin are three medicinal ingredients in *C. grandiflora* Benth; however, their biosynthetic pathway is unexplored.

So far, catalpol biosynthesis containing terpenoid backbone pathway and iridoid pathway has not been fully deciphered due to the deficiency of detailed information on genetic and molecular levels [20]. In 1993, Damtoft found that 8-epi-deoxyloganic acid, bartsioside, and aucubin are intermediates of catalpol biosynthesis by feeding experiments [48]. Then, Jensen et al. confirmed that catalpol is synthesized via decarboxylated iridoids pathway (Route II), which involved 8-epi-iridodial, 8-epi-iridotrial, and 8-epi-deoxyloganic acid [73]. In 2013, the more detailed route II was proposed in *R. glutinosa* and *P. kurrooa* [21,41]. In 2015, the complete catalpol biosynthesis pathway was hypothesized in *P. kurrooa* according to data of the transcriptome mining, gene expression, and picroside content [17]. In our transcriptomes, 368 unigenes were annotated to the catalpol biosynthetic pathway with

60 unigenes upregulated in leaves and 39 unigenes in roots; simultaneously combined with the fact that *F3D* gene was not expressed in roots, we deduced that catalpol biosynthesis was mainly active in leaves. A recent article showed that, in wild *C. grandiflora* Benth, the content of catalpol is far higher in leaves than in stems and roots [74], which also implied that catalpol is mainly synthesized in leaves other than roots. The discovery of rate-limiting enzymes is essential for synthetic biology; therefore, some genes are discussed here. Catalpol biosynthesis begins with the terpenoid backbone pathway, which contains the MEP and MVA pathways. In the MEP pathway, the DXS enzyme is the first and rate-limiting enzyme, and in *A. annua*, among the three AaDXSs, only AaDXS2 might participate in artemisinin biosynthesis [75]. Contrary to *A. annua*, the *DXS*s were more abundant in our transcriptome, which seems that DXS was not a limiting enzyme in *C. grandiflora* Benth. Further studies are needed to clarify which DXS functions in MEP pathway. A recent report showed that plastidial IDI plays an important role in optimizing the ratio between IPP and DMADP as precursors for different downstream isoprenoid pathways while mutation of *IDI1* reduced the content of carotenoids in fruits, flowers, and cotyledons (except mature leaves) [44]. In our transcriptome, there were 28 *IDI* genes with two upregulated in leaves compared with roots, which highlights their importance in terpenoid backbone biosynthesis (Table 4). However, there were no significant differences for the overall expression of *IDI* genes in roots, stems, and leaves in our transcriptome (Figure 9b). What is interesting is that there was only one *MCS* gene in our transcriptome; however, its expression levels in roots, stems, and leaves were all relatively high, which directly denied that *MCS* was a rate-limiting enzyme gene. According to the expression profile, *MCT* may be a rate-limiting enzyme for roots (Figure 9b). In addition, the relative contribution of the MEP and MVA pathways for a specific pathway is a focus scientist paying attention to. In *P. kurroa*, the biosynthesis of picroside-I is contributed solely by the MEP pathway [17]. In *Taxus baccata*, the MEP pathway provides the main source of universal terpenoid precursor IPP [76]. However, in *C. grandiflora* Benth, the contribution of the MEP and MVA pathways for catalpol biosynthesis remains to be clarified and it will be resolved by the inhibition experiments in the future.

Acteoside biosynthesis was first studies in an *O. europaea* cell with feeding experiments, which outline the basic pathway profile: caffeoyl moiety was synthesized through the phenylalanine-derived pathway including intermediates cinnamic acid, p-coumaric acid, and caffeic acid, while hydroxytyrosol moiety was formed via the tyrosine-derived pathway including two alternative routes [26]. Then, HCT enzyme which connects the caffeoyl moiety and the hydroxytyrosol moiety, UGT enzymes, and the corresponding enzymes of the phenylalanine-derived pathway and tyrosine-derived pathway were hypothesized in *R. glutinosa* [28]. All of the acteoside pathway genes were found in our transcriptome of *C. grandiflora* Benth. Expression profiles showed that genes involved in both the phenylalanine-derived pathway and the tyrosine-derived pathway were more abundant in leaves and stems compared to roots, especially for the *PAL* and *PPO* genes (Figure 10b). This is consistent with the reports that, in *Harpagophytum procumbens*, the content of acteoside was higher in leaves and stems than in roots and that, in *Sesamum indicum*, the content of acteoside in leaves is far higher than in stems and roots [25,77].

Studies have shown that PAL is an entry-point enzyme which can convert *L*-Phe into trans-cinnamic acid and that it plays a vital role in channeling carbon flux from primary metabolism into the phenylpropanoid pathway [78]. So far, *PAL* gene has been cloned from many medicinal plants, such as *Ocimum basilicum* [79], *Ginkgo biloba* [80], *Salvia miltiorrhiza* [81], and *A. annua* [82]. In *G. biloba*, the highest expression of *GbPAL* gene was found in leaves, followed by stems, and the lowest expression was in roots; transcription levels of *GbPAL* were closely related to flavonoid accumulation [80]. In *R. glutinosa*, the *RgPAL* gene (CL1389.Contig1) shared the same expression pattern as in *G. biloba* [28]. In *A. annua*, the highest expression of the *AaPAL* gene was found in young leaves and the lowest expression of that was in roots [82]. In plants, *PAL* gene is a multi-gene family and the gene number ranges from 4 in *A. thaliana* to more than 12 in tomato and potato [83]. For example, there are 6 *PAL* genes in *R. glutinosa* [28]. Recently, three different redundancy phenomena including active compensation in ligand plus passive compensation in receptor in tomato, passive compensation in ligand plus active compensation in receptor in Arabidopsis, and active compensation in both in corn have been figured out [84]; however, which type does the *CgPAL* genes belong to and whether they benefit the plants themselves in *C. grandiflora* Benth remain to be discovered. Unlike potato, the *PAL* gene family is highly redundant but underutilized due to the highly silencing mechanism in tomato [83]. In our transcriptome, there are 19 *PAL* genes and their highest expressions are found in leaves and stems with the lowest expression in roots (Figure 10b), which is similar to that in *G. biloba*, *R. glutinosa*, and *A. annua* [28,80,82]. Our transcriptome profiling data showed that 10 of 19 *CgPAL* genes were not expressed or slightly expressed in roots, stems, and leaves (Figure S1), which implied that gene silencing was also active in *C. grandiflora* Benth, and DNA cytosine methylation may account for this phenomenon [83]. A recent report showed that functional redundancy among *BZR/BEH* (*BRASSINAZOLE-RESISTANT/BRI1-EMS-SUPRESSOR1/BRASSINAZOLE-RESISTANT1 HOMOLOG*) gene family members is not necessary for trait robustness [85]. Even in tomato, only *PAL5* was expressed under environmental stimuli [83]. Therefore, *PAL* genes including the 3 significantly upregulated and 1 significantly downregulated in leaf vs. root in *C. grandiflora* Benth played important roles in acteoside biosynthesis (Table 5).

Polyphenol oxidase is usually undesirable in fruit and vegetable due to the browning, while it is desirable in tea, coffee, cocoa, etc. for the pigmentation [86]. Polyphenol oxidase (1,2-benzenediol: oxygen oxidoreductase), also known as tyrosinase, catechol oxidase, and laccase according to the specific substrate and reaction mechanism, is a group of copper-containing proteins [86,87]. A typical PPO protein contains three conservative regions: an N-terminal transit peptide that is responsible for the import of PPO into the thylakoid lumen; a di-copper center, each with three histidine residues to bind a copper atom; and a C-terminal region [88]. Polyphenol oxidases can catalyze two quite different types of reactions: monophenol monooxygenases (E.C. 1.14.18.1) activity and o-diphenol oxidation reactions including catechol oxidases (E.C. 1.10.3.1) and laccases (E.C. 1.10.3.2) activity [87]. In plants, polyphenol oxidase is localized in chloroplasts and the reaction product accumulated in thylakoid [89]. The number of *PPO* gene ranges from 1 to 13 in land plants with 0 for green algae and *A. thaliana*, and tandem duplications of the *PPO* gene family is common in dicotyledon [88]. In our transcriptome, 11 *PPO* genes were clustered into three groups. Expression levels of the upper group including *PPO7*, *PPO9*, *PPO10*, and *PPO11* were higher in leaves and stems compared with roots, while that of the bottom group including *PPO1*, *PPO2*, and *PPO3* were higher in roots and stems than in leaves with the somewhat low expressions in middle group including *PPO4*, *PPO5*, *PPO6*, and *PPO8* (Figure S2). Phylogenetic analysis of 11 CgPPOs with 6 PPOs of *Solanum melongena* and 6 PPOs of *Solanum lycopersicum* showed that all of our CgPPO proteins are clustered into one clade and that the other 12 PPO proteins formed another two clades (Figure S3). These species-specific PPO clades were also found in four major land plant lineages including *Populus trichocarpa*, *Glycine max*, *Vitis vinifera*, and *Aquilegia coerulea*, which implied that *CgPPO* genes were also formed by independent burst of gene duplication [88].

Azafrin ($C_{27}H_{38}O_4$) derivates from tetraterpenoids (C_{40}). It has been found in many medicinal plants such as rhizome of *Alectra chitrakutensis*, *Bergenia ciliate*, *Caralluma umbellate*, and *Alectra parasitica*, and it has the functions of being antimicrobial, anti-inflammatory, analgesic, antioxidant, treatment of cardiovascular diseases [90–93]. Roots of *C. grandiflora* Benth display orange-yellow color, which is largely due to the presence of abundant azafrin as *A. parasitica* [93]. So far, the biosynthetic pathway of azafrin is not established from perspectives of chemistry and biology. There are studies implying that excentric cleavage of carotenoid compounds is a possible route [94]. CCD7 can catalyze β-carotene (C_{40}) into 10′-apo-β-carotenal (C_{27}) and ionone (C_{13}) to support the above hypothesis [39]. In the view of molecular structure, the differences between 10′-apo-β-carotenal and azafrin are one terminal carboxyl group and two hydroxyl groups in cyclohexane skeleton. Therefore, two reactions are indispensable from 10′-apo-β-carotenal to azafrin: one is to convert the aldehyde group into carboxyl groups, and the other is to insert two oxygen atoms into cyclohexane skeleton to generate two hydroxyl groups. The ALDH superfamily comprises a group of enzymes involved in the NAD$^+$ (Nicotinamide Adenine Dinucleotide) or NADP$^+$ (Nicotinamide Adenine Dinucleotide Phosphate)-dependent conversion of various aldehydes to their corresponding carboxylic acids [95]. Although there are only 76 NAD$^+$-dependent *ALDH* genes in our transcriptome, they are candidate genes for azafrin biosynthesis. In plant, CYP450s are responsible for many oxidative reactions such as hydroxylation, epoxidation, dealkylation, and dehydration, and the reactions catalyzed by CYP450s are irreversible [53]. There are 413 CYP450 unigenes in our transcriptome, of which 5 are significantly upregulated ($\log_2$(FC) > 10) and 5 are significantly downregulated in leaf vs. root (Table 6). They can be candidate genes of azafrin biosynthesis. The key enzymes determine the flux of the pathway, and the expression of the key enzyme gene dominates the number of enzymes. In marigold, the expression level of the *LCYE* gene in petals and *LCYB* gene in leaves were positively correlated with the lutein content [96]. In *Momordica cochinchinensis*, transcriptional regulation of genes including *HMGR*, *HDS*, *PSY*, *PDS*, *ZDS*, *CRTISO,* and *LCYE* may determine the alteration of carotenoid content during fruit ripening [97]. Our transcriptome data showed that only the expression levels of *HDS*, *PSY*, *ZDS*, and *CRTISO* were more abundant in roots than leaves and stems (Figure 11b). However, trace expression of the *DWARF27* gene in leaves and low expression of the *CCD7* gene in roots, stems, and leaves suggested that they were two rate-limiting enzymes in azafrin biosynthesis. *DWARF27*, which exhibits increased tillers and reduced plant height, was first studied in rice [98]. It encodes an iron-containing protein localized in chloroplasts and is expressed mainly in vascular cells of shoots and roots [98]. Further studies indicated that DWARF27 is an all-trans/9-cis isomerase which can convert all-trans-β-carotene into 9-cis-β-carotene in vivo and in vitro [99]. Obviously, DWARF27 is vital for azafrin biosynthesis. What is interesting is that there is only one *CCD7* gene in our transcriptome which coincides with the all *CCD7* genes identified including maize, rice, sorghum, *Selaginella moellendorfii*, *Physcomitrella patens*, and *Chlamydomonas reinhardtii* and is a single copy [100]. The highest expression of the *CCD7* gene was found in roots among maize, *A. thaliana*, pea, and petunia [100]. However, the highest expression in our transcriptome is in stems.

In the future, studies related to catalpol, acteoside, and azafrin biosynthesis will focus on the following aspects: (1) to construct a transgenic system for *C. grandiflora* Benth according to the successive tissue culture technology for verification of gene function, to characterize the putative genes of three pathways, and to verify their functions by enzyme assays in vitro or to overexpress them in vivo; (2) to explore the correlation between the contents of active component and related gene expression levels, to clone the putative TFs, and to verify their functions in the biosynthesis of active components via chromatin immunoprecipitation and overexpression in vivo; and (3) to figure out the biosynthetic pathway using feeding experiments with suspension cells.

4. Materials and Methods

4.1. Plant Materials and RNA Isolation

The artificial, cultivated *Centranthera grandiflora* Benth was grown in fields with *Cyperus rotundus* in Yushancheng base, Yuxi Flyingbear Agricultural Development Company Limited, Yuxi, Yunnan Province, China (Figure 15a). Three healthy plants with the same growth potential were selected, and the fresh roots, stems, and leaves were collected from one-year-old *C. grandiflora* plants on May 7, 2018 (Figure 15b). Materials from three individual plants were collected using scissors to yield 1 g of root, stem, and leaf samples (BioSample accessions: SAMN12499651, SAMN12499652, SAMN12499653, SAMN12499654, SAMN12499655, SAMN12499656, SAMN12499657, SAMN12499658, and SAMN12499659). After wrapping with tinfoil and tagging, all samples were immediately frozen in liquid nitrogen and stored at −80 °C.

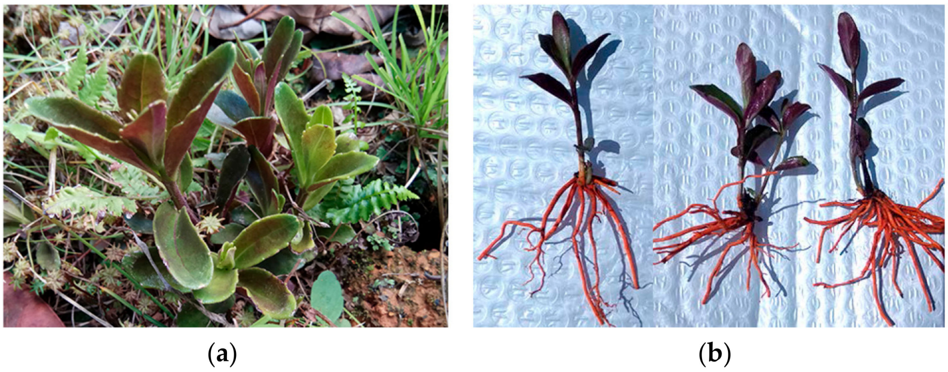


Figure 15. Plant materials of *C. grandiflora* Benth: (**a**) *C. grandiflora* Benth growing in the field with *Cyperus rotundus* and (**b**) Roots, stems, and leaves used in experiments for sequencing and qRT-PCR.

For total RNA extraction and quality control, refer to Zhang et al. [101].

4.2. Library Preparation for Transcriptome Sequencing

A total amount of 1.5 μg RNA per sample was used as input material for the RNA sample preparations. Sequencing libraries were generated using the NEBNext® UItra™ RNA Library Prep Kit for Illumina (NEB, Ipswich, MA, USA) following the manufacturer's recommendations. Index codes were added to attribute sequences to each sample. Briefly, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. Fragmentation was carried out using divalent cations under elevated temperature in a NEBNext First Strand Synthesis Reaction Buffer (5×). First strand cDNA was synthesized using random hexamer primer and M-MuLV Reverse Transcriptase (RNase H^{-}). Second-strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H. Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. After adenylation of 3′ ends of DNA fragments, NEBNext Adaptor with hairpin loop structure was ligated to prepare for hybridization. In order to select cDNA fragments preferentially 250–300 bp in length, the library fragments were purified with AMPure XP system (Beckman Coulter, Beverly, USA). Then, 3 μL USER Enzyme (NEB, Ipswich, MA, USA) was used with size-selected, adaptor-ligated cDNA at 37 °C for 15 min followed by 5 min at 95 °C before PCR. PCR was performed with Phusion High-Fidelity DNA polymerase, Universal PCR primers, and Index(X) Primer. At last, PCR products were purified (AMPure XP system) and library quality was assessed on the Agilent Bioanalyzer 2100 system (Agilent Technologies, Palo Alto, CA, USA).

4.3. Clustering and Sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v3-cBot-HS (Illumia, San Diego, CA, USA) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina Hiseq2000 platform and paired-end reads were generated.

4.4. Data Filtering and Transcriptome Assembly

The flow of bioinformatics analysis is showed in Figure 16. Before assembly, raw reads containing adaptors, more than 10% N bases, and more than 50% of low quality were removed to obtain clean reads. Meanwhile, Q20, Q30, and GC content were used to assess the data quality. All the subsequent analyses were based on these clean reads. As there are no reference genomes available for *C. grandiflora*, the clean reads of roots, stems, and leaves were assembled together. The paired-end reads of each sample were merged into one interleaved fastq file. All the nine pooled files were assembled using Trinity software (version: r20140413p1) (Cambridge, MA, USA) with min_k-mer_cov set to 2 and all other parameters settings as default [102]. After clustering and de-redundancy by Corset software (version: 1.07) (VIC, Austrilia) [103], the clean nonredundant unigenes was generated.

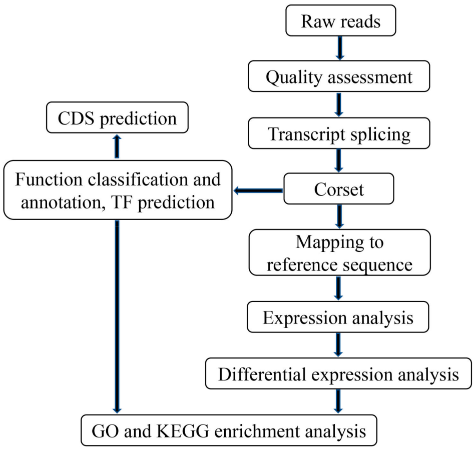

Raw reads
Quality assessment
Transcript splicing
Corset
Mapping to reference sequence
Expression analysis
Differential expression analysis
GO and KEGG enrichment analysis
Function classification and annotation, TF prediction
CDS prediction

Figure 16. Flow chart of transcriptome bioinformatics analysis for *C. grandiflora* Benth.

4.5. Gene Functional Annotation

Gene function was annotated based on the following databases: Nr (NCBI nonredundant protein sequences, diamond v0.8.22, *e*-value = 10^{-5}), NT (NCBI nucleotide sequences, NCBI blast 2.2.28+, *e*-value = 10^{-5}), PFAM (Protein family, HMMER 3.0 package, hmmscan *e*-value = 10^{-2}), SwissProt (a manually annotated and reviewed protein sequence database, diamond v0.8.22, *e*-value = 10^{-5}), KOG/COG (Clusters of Orthologous Groups of proteins/euKaryotic Ortholog Groups, diamond v0.8.22, *e*-value = 10^{-5}), KAAS (version: r140224, *e*-value = 10^{-10}), and GO (Blast2GO v2.5 and inhouse script, *e*-value = 10^{-6}). To figure out the TF families involved in the active ingredient biosynthesis, iTAK software (https://github.com/kentn/iTAK/) was used to predict the TF. Its basic principle is to identify TF by hmmscan using TF family and rules defined by classification in the database. For the identification and classification methods of TF, refer to Perez-Rodriguez et al. [104].

4.6. Differential Expression Analysis

The calculation of unigene expression was performed using the RPKM method, and gene expression levels were estimated by RSEM (version: 1.2.15, parameter for bowtie2: mismatch 0) for each sample [43]. Differential expression analysis of two organs was performed using the DESeq R package 1.10.1. DESeq provides statistical routines for determining differential expression in digital gene expression data using a model based on the negative binomial distribution. The resulting *p* values were adjusted using Benjamini and Hocberg's approach for controlling false discovery rates. Genes with a threshold of foldchange ≥ 2 and *p*-value < 0.05 found by DESeq were assigned as differentially expressed.

4.7. GO Enrichment and KEGG Pathway Enrichment Analysis

Gene ontology enrichment analysis of the differentially expressed genes (DEGs) was implemented by the GOseq (version: 1.10.0) and topGO (version: 2.10.0) R packages based Wallenius non-central hyper-genometric distribution, which can adjust for gene length bias in DEGs [105]. KOBAS software (Beijing, China) was used to test the statistical enrichment of DEGs in KEGG pathways [106]. The expressed genes (FPKM ≥ 1) were used as background with a corrected *p*-value ≤ 0.05 for both enrichment analyses.

4.8. qRT-PCR Analysis

Total RNA was extracted from roots, stems, and leaves of annual *C. grandiflora* Benth. The first strand of DNA was synthesized using reverse transcription kit PrimeScript RT Master Mix (Perfect Real Time) (Takara, Dalian, China). Specific primers were designed according to the selected gene sequences for expression analysis (Table S14). Using the *C. grandiflora* Benth *CgUbi* gene (Accession number: MK256646) as an internal reference, qPCR was performed using chimeric fluorescence detection kit TB Green Premix Ex Taq II (Takara, China). Each reaction was repeated three times. Reaction was amplified by LightCycler 480II fluorescent quantitative PCR (Roche, Basel, Switzerland). After the amplification, results were calibrated by internal reference gene and the relative gene expressions in roots, stems, and leaves were calculated automatically by the $2^{-\Delta\Delta Ct}$ method.

4.9. Data Submission

This Transcriptome Shotgun Assembly project (PRJNA558809) has been deposited at DDBJ/ENA/GenBank under the accession GHUX00000000. The version described in this paper is the first version, GHUX01000000.

5. Conclusions

Centranthera grandiflora Benth has been used to prevent and treat CVDs for a long time; however, the biosynthesis pathway of its active components including catalpol, acteoside, and azafrin remains undeciphered. Transcriptome sequencing technology is an effective way to discover the genes of this herb's biosynthesis pathways and its regulatory mechanisms. In this study, nine cDNA libraries were constructed from the roots, stems, and leaves of *C. grandiflora* Benth and sequenced by an Illumina Hiseq2000 platform. As a result, 438,112,930 clean reads were obtained and 153,198 unigenes were assembled. Among these genes, 557, 213, and 161 unigenes were annotated into catalpol, acteoside, and azafrin biosynthetic pathways, respectively. Azafrin can be synthesized through β-carotene, 9-cis-β-carotene, and 10′-apo-β-carotenal with the corresponding enzymes DWARF27, CCD7, ALDH, and CYP450. Also, the *PAL* gene silencing phenomenon is discovered and discussed. The candidate TF MYBs involved in the regulation of these pathways were proposed. Our results represent the first genomic resource for *C. grandiflora* Benth, which is a starting point for exploration of this valuable herb in molecular biology.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/20/23/6034/s1.

Author Contributions: X.Z. conceived and design of experiments. X.Z., C.L., and L.W. performed the experiments. X.Z., C.L., and W.Q. analyzed the data. Y.F. contributed materials. X.Z. wrote the manuscript, and W.Q. reviewed and edited the manuscript.

Funding: This research was funded by Yunnan Provincial Science and Technology Department (grant number 2017FH001-024, 2016FD113) and China Scholarship Council (grant number 201908530057).

Acknowledgments: We gratefully acknowledge Beijing Novogene Bioinformatics Technology Company of China for carrying out the sequencing of the transcriptomes. We gratefully acknowledge the anonymous reviewers for their constructive comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Clark, H. NCDs: A challenge to sustainable human development. *Lancet* **2013**, *381*, 510–511. [CrossRef]
2. Chen, W.; Gao, R.; Liu, L.; Zhu, M.; Wang, W.; Wang, Y.; Wu, Z.; Li, H.; Gu, D.; Yang, Y. China cardiovascular diseases report 2015: A summary. *J. Geriatr. Cardiol.* **2017**, *14*, 1–10. [CrossRef] [PubMed]
3. Hao, P.; Jiang, F.; Cheng, J.; Ma, L.; Zhang, Y.; Zhao, Y. Traditional Chinese medicine for cardiovascular disease: Evidence and potential mechanisms. *J. Am. Coll. Cardiol.* **2017**, *69*, 2952–2966. [CrossRef] [PubMed]
4. Editorial Committee of Flora of China. *Flora of China*, 1st ed.; Science Press: Beijing, China, 1979; p. 345.

5. Liang, D. A breakthrough in tissue culture and rapid propagation of rare and endangered medicinal *Centranthera grandiflora* Benth. *Yunnan Information Daily*, 14 May 2018.
6. Liao, L.; Zhang, Z.; Hu, Z.; Chou, G.; Wang, Z. Iridoid glycosides from *Centranthera grandiflora*. *Chin. Tradit. Herb. Drugs* **2012**, *43*, 2369–2371.
7. Liao, L. Investigation into the Bioactive Components and Chemical Constituents from the Roots of *Centranthera grandijlora* Benth. Ph.D. Thesis, Shanghai University of Traditional Chinese Medicine, Shanghai, China, 21 June 2014.
8. Liang, J.; Zhang, J.; Ma, X.; Wang, G.; Chen, Y.; Wen, Y.; Gan, L. Identification of chemical constituents from *Centranthera grandiflora*. *Chin. Bull. Bot.* **1984**, *2*, 47.
9. Editorial Committee of Chinese Materia Medica. *Chinese Materia Medica*, 1st ed.; Shanghai Science and Technology Press: Shanghai, China, 1996; p. 397.
10. Wang, Z.; Wang, Q.; Yan, J.; Cong, Y. Chemical constituents and activities research of *Centranthera grandiflora* Benth. In Proceedings of the Yunnan Pharmaceutical Conference in 2012, Yunnan Pharmaceutical Conference in 2012, Kunming, China, 20 September 2012; pp. 1–4.
11. Chen, M.; Ye, Z.; Xueying, Z.; Jianjun, Z.; Yongchun, Z.; Liqing, Y.; Liuyan, Y. A tissue culture method of Centranthera grandiflora Benth. China Patent ZL201710499456.0, 15 March 2019.
12. Hua, W.; Zheng, P.; He, Y.; Cui, L.; Kong, W.; Wang, Z. An insight into the genes involved in secoiridoid biosynthesis in *Gentiana macrophylla* by RNA-seq. *Mol. Biol. Rep.* **2014**, *41*, 4817–4825. [CrossRef]
13. Liu, Y.; Wang, Y.; Guo, F.; Zhan, L.; Mohr, T.; Cheng, P.; Huo, N.; Gu, R.; Pei, D.; Sun, J.; et al. Deep sequencing and transcriptome analyses to identify genes involved in secoiridoid biosynthesis in the Tibetan medicinal plant *Swertia mussotii*. *Sci. Rep.* **2017**, *7*, 43108. [CrossRef]
14. Miettinen, K.; Dong, L.; Navrot, N.; Schneider, T.; Burlat, V.; Pollier, J.; Woittiez, L.; van der Krol, S.; Lugan, R.; Ilc, T.; et al. The seco-iridoid pathway from *Catharanthus roseus*. *Nat. Commun.* **2014**, *5*, 3606–3616. [CrossRef]
15. Munkert, J.; Pollier, J.; Miettinen, K.; Van Moerkercke, A.; Payne, R.; Müller-Uri, F.; Burlat, V.; O'Connor, S.E.; Memelink, J.; Kreis, W. Iridoid synthase sctivity is common among the plant progesterone 5β-reductase family. *Mol. Plant* **2015**, *8*, 136–152. [CrossRef]
16. Huang, L.; Liu, C. *Molecular Pharmacognosy*, 3rd ed.; Science Press: Beijing, China, 2015; pp. 238–342.
17. Shitiz, K.; Sharma, N.; Pal, T.; Sood, H.; Chauhan, R.S. NGS transcriptomes and enzyme inhibitors unravel complexity of picrosides biosynthesis in *Picrorhiza kurroa* Royle ex. Benth. *Plos ONE* **2015**, *10*, e0144546. [CrossRef] [PubMed]
18. Zhi, J.; Li, Y.; Zhang, Z.; Yang, C.; Xie, C. Molecular regulation of catalpol and acteoside accumulation in radial striation and non-radial striation of *Rehmannia glutinosa* tuberous root. *Int. J. Mol. Sci.* **2018**, *19*, 3751. [CrossRef] [PubMed]
19. Xue, L.; He, Z.; Bi, X.; Xu, W.; Wei, T.; Wu, S.; Hu, S. Transcriptomic profiling reveals MEP pathway contributing to ginsenoside biosynthesis in *Panax ginseng*. *BMC Genom.* **2019**, *20*, 134. [CrossRef] [PubMed]
20. Dinda, B. Chemistry and Biosynthesis of Iridoids. In *Pharmacology and Applications of Naturally Occurring Iridoids*, 1st ed.; Dinda, B., Ed.; Springer International Publishing: Cham, Switzerland, 2019; pp. 119–143.
21. Wajid, W.B.; Niha, D.; Sumeer, R.; Satiander, R.; Rukmankesh, M.; Amit, N.; Rekha, S.D.; Nasheeman, A.; Ram, V.; Surrinder, K.L. Molecular characterization of UGT94F2 and UGT86C4, two glycosyltransferases from *Picrorhiza kurrooa*: Comparative structural insight and evaluation of substrate recognition. *PLoS ONE* **2013**, *8*, e73804. [CrossRef]
22. Kumar, V. OMICS-based approaches for elucidation of picrosides bsynthesis in *Picrorhiza kurroa*. In *OMICS-Based Approaches in Plant Biotechnology*, 1st ed.; Banerjee, R., Kumar, V., Kumar, J., Eds.; Scrivener Publishing LLC: Beverly, MA, USA, 2019; pp. 145–166.
23. Ye, P.; Liang, S.; Wang, X.; Duan, L.; Jiang-Yan, F.; Yang, J.; Zhan, R.; Ma, D. Transcriptome analysis and targeted metabolic profiling for pathway elucidation and identification of a geraniol synthase involved in iridoid biosynthesis from *Gardenia jasminoides*. *Ind. Crop. Prod.* **2019**, *132*, 48–58. [CrossRef]
24. Nagatoshi, M.; Terasaka, K.; Nagatsu, A.; Mizukami, H. Iridoid-specific Glucosyltransferase from *Gardenia jasminoides*. *J. Biol. Chem.* **2011**, *286*, 32866–32874. [CrossRef]
25. Fuji, Y.; Ohtsuki, T.; Matsufuji, H. Accumulation and subcellular localization of acteoside in sesame plants (*Sesamum indicum* L.). *ACS Omega* **2018**, *3*, 17287–17294. [CrossRef]
26. Saimaru, H.; Orihara, Y. Biosynthesis of acteoside in cultured cells of *Olea europaea*. *J. Nat. Med.* **2010**, *64*, 139–145. [CrossRef]

27. Alipieva, K.; Korkina, L.; Orhan, I.E.; Georgiev, M.I. Verbascoside—A review of its occurrence, (bio)synthesis and pharmacological significance. *Biotechnol. Adv.* **2014**, *32*, 1065–1076. [CrossRef]
28. Wang, F.; Zhi, J.; Zhang, Z.; Wang, L.; Suo, Y.; Xie, C.; Li, M.; Zhang, B.; Du, J.; Gu, L. Transcriptome analysis of salicylic acid treatment in *Rehmannia glutinosa* hairy roots using RNA-seq technique for identification of genes involved in acteoside biosynthesis. *Front. Plant Sci.* **2017**, *8*, 787. [CrossRef]
29. Cheimonidi, C.; Samara, P.; Polychronopoulos, P.; Tsakiri, E.N.; Nikou, T.; Myrianthopoulos, V.; Sakellaropoulos, T.; Zoumpourlis, V.; Mikros, E.; Papassideri, I. Selective cytotoxicity of the herbal substance acteoside against tumor cells and its mechanistic insights. *Redox Biol.* **2018**, *16*, 169–178. [CrossRef] [PubMed]
30. Deng, H.; Sun, M.; Chen, H.; Wang, X.; Qiong, W.; Chang, Q. Effect of acteoside on behavioral changes and endoplasmic reticulum stress in prefrontal cortex of depressive rats. *Chin. J. Pathophysiol.* **2018**, *34*, 101–106.
31. Khullar, M.; Sharma, A.; Wani, A.; Sharma, N.; Sharma, N.; Chandan, B.; Kumar, A.; Ahmed, Z. Acteoside ameliorates inflammatory responses through NFkB pathway in alcohol induced hepatic damage. *Int. Immunopharmacol.* **2019**, *69*, 109–117. [CrossRef]
32. Bai, Y.; Zhu, R.; Tian, Y.; Li, R.; Chen, B.; Zhang, H.; Xia, B.; Zhao, D.; Mo, F.; Zhang, D. Catalpol in diabetes and its complications: A review of pharmacology, pharmacokinetics, and safety. *Molecules* **2019**, *24*, 3302. [CrossRef] [PubMed]
33. Yang, S.; Chou, G.; Li, Q. Cardioprotective role of azafrin in against myocardial injury in rats via activation of the Nrf2-ARE pathway. *Phytomedicine* **2018**, *47*, 12–22. [CrossRef] [PubMed]
34. Ohmiya, A.; Kato, M.; Shimada, T.; Nashima, K.; Kishimoto, S.; Nagata, M. Molecular basis of carotenoid accumulation in horticultural crops. *Horticult. J.* **2019**, UTD-R003. [CrossRef]
35. Ampomah-Dwamena, C.; Thrimawithana, A.H.; Dejnoprat, S.; Lewis, D.; Espley, R.V.; Allan, A.C. A kiwifruit (*Actinidia deliciosa*) R2R3-MYB transcription factor modulates chlorophyll and carotenoid accumulation. *New Phytol.* **2019**, *221*, 309–325. [CrossRef]
36. Wang, Q.; Cao, T.; Zheng, H.; Zhou, C.; Wang, Z.; Wang, R.; Lu, S. Manipulation of carotenoid metabolic flux by lycopene cyclization in ripening red pepper (*Capsicum annuum* var. conoides) fruits. *J. Agric. Food Chem.* **2019**, *67*, 4300–4310. [CrossRef]
37. Kanehisa, M. Kegg Pathway Database. Available online: https://www.kegg.jp/kegg/pathway.html (accessed on 1 July 2019).
38. Schliemann, W.; Kolbe, B.; Schmidt, J.; Nimtz, M.; Wray, V. Accumulation of apocarotenoids in mycorrhizal roots of leek (*Allium porrum*). *Phytochemistry* **2008**, *69*, 1680–1688. [CrossRef]
39. Alder, A.; Jamil, M.; Marzorati, M.; Bruno, M.; Vermathen, M.; Bigler, P.; Ghisla, S.; Bouwmeester, H.; Beyer, P.; Al-Babili, S. The path from β-carotene to carlactone, a strigolactone-like plant hormone. *Science* **2012**, *335*, 1348–1351. [CrossRef]
40. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [CrossRef] [PubMed]
41. Sun, P.; Song, S.; Zhou, L.; Zhang, B.; Qi, J.; Li, X. Transcriptome analysis reveals putative genes involved in iridoid biosynthesis in *Rehmannia glutinosa*. *Int. J. Mol. Sci.* **2012**, *13*, 13748–13763. [CrossRef] [PubMed]
42. Cao, H.; Nuruzzaman, M.; Xiu, H.; Huang, J.; Wu, K.; Chen, X.; Li, J.; Wang, L.; Jeong, J.H.; Park, S.J.; et al. Transcriptome analysis of methyl jasmonate-elicited *Panax ginseng* adventitious roots to discover putative ginsenoside biosynthesis and transport genes. *Int. J. Mol. Sci.* **2015**, *16*, 3035–3057. [CrossRef] [PubMed]
43. Li, B.; Dewey, C. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [CrossRef] [PubMed]
44. Pankratov, I.; McQuinn, R.; Schwartz, J.; Bar, E.; Fei, Z.; Lewinsohn, E.; Zamir, D.; Giovannoni, J.J.; Hirschberg, J. Fruit carotenoid-deficient mutants in tomato reveal a function of the plastidial isopentenyl diphosphate isomerase (IDI1) in carotenoid biosynthesis. *Plant J.* **2016**, *88*, 82–94. [CrossRef] [PubMed]
45. Wieffering, J. Aucubinartige glucoside (pseudoindikane) und verwandte heteroside als systematische merkmale. *Phytochemistry* **1966**, *5*, 1053–1064. [CrossRef]
46. Oshio, H.; Inouye, H. Iridoid glycosides of *Rehmannia glutinosa*. *Phytochemistry* **1982**, *21*, 133–138. [CrossRef]
47. Damtoft, S. Biosynthesis of the iridoids aucubin antirrinoside from 8-epi-deoxyloganic acid. *Phytochemistry* **1983**, *22*, 1929–1930. [CrossRef]
48. Damtoft, S. Biosynthesis of catalpol. *Phytochemistry* **1994**, *35*, 1187–1189. [CrossRef]

49. He, J.; Hu, X.; Zeng, Y.; Li, Y.; Wu, H.; Qiu, R.; Ma, W.; Li, T.; Li, C.; He, Z. Advanced research on acteoside for chemistry and bioactivities. *J. Asian Nat. Prod. Res.* **2011**, *13*, 449–464. [CrossRef]
50. Berry, H.M.; Rickett, D.V.; Baxter, C.J.; Enfissi, E.M.A.; Fraser, P.D. Carotenoid biosynthesis and sequestration in red chilli pepper fruit and its impact on colour intensity traits. *J. Exp. Bot.* **2019**, *70*, 2637–2650. [CrossRef] [PubMed]
51. Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **2011**, *40*, D109–D114. [CrossRef] [PubMed]
52. Dela Seña, C.; Sun, J.; Narayanasamy, S.; Riedl, K.M.; Yuan, Y.; Curley, R.W.; Schwartz, S.J.; Harrison, E.H. Substrate specificity of purified recombinant chicken β-carotene 9′, 10′-oxygenase (BCO2). *J. Biol. Chem.* **2016**, *291*, 14609–14619. [CrossRef] [PubMed]
53. Rasool, S.; Mohamed, R. Plant cytochrome P450s: Nomenclature and involvement in natural product biosynthesis. *Protoplasma* **2016**, *253*, 1197–1209. [CrossRef] [PubMed]
54. Kirby, J.; Keasling, J.D. Biosynthesis of plant isoprenoids: Perspectives for microbial engineering. *Annu. Rev. Plant Biol.* **2009**, *60*, 335–355. [CrossRef]
55. Yan, T.; Chen, M.; Shen, Q.; Li, L.; Fu, X.; Pan, Q.; Tang, Y.; Shi, P.; Lv, Z.; Jiang, W. HOMEODOMAIN PROTEIN 1 is required for jasmonate-mediated glandular trichome initiation in *Artemisia annua*. *New Phytol.* **2017**, *213*, 1145–1155. [CrossRef]
56. Chen, M.; Yan, T.; Shen, Q.; Lu, X.; Pan, Q.; Huang, Y.; Tang, Y.; Fu, X.; Liu, M.; Jiang, W.; et al. GLANDULAR TRICHOME-SPECIFIC WRKY 1 promotes artemisinin biosynthesis in *Artemisia annua*. *New Phytol.* **2017**, *214*, 304–316. [CrossRef]
57. Tan, H.; Xiao, L.; Gao, S.; Li, Q.; Chen, J.; Xiao, Y.; Ji, Q.; Chen, R.; Chen, W.; Zhang, L. TRICHOME AND ARTEMISININ REGULATOR 1 is required for trichome development and artemisinin biosynthesis in *Artemisia annua*. *Mol. Plant* **2015**, *8*, 1396–1411. [CrossRef]
58. Yu, Z.; Li, J.; Yang, C.; Hu, W.; Wang, L.; Chen, X. The jasmonate-responsive AP2/ERF transcription factors AaERF1 and AaERF2 positively regulate artemisinin biosynthesis in *Artemisia annua* L. *Mol. Plant* **2012**, *5*, 353–365. [CrossRef]
59. Han, J.; Wang, H.; Lundgren, A.; Brodelius, P.E. Effects of overexpression of AaWRKY1 on artemisinin biosynthesis in transgenic *Artemisia annua* plants. *Phytochemistry* **2014**, *102*, 89–96. [CrossRef]
60. Lu, X.; Zhang, L.; Zhang, F.; Jiang, W.; Shen, Q.; Zhang, L.; Lv, Z.; Wang, G.; Tang, K. AaORA, a trichome-specific AP2/ERF transcription factor of *Artemisia annua*, is a positive regulator in the artemisinin biosynthetic pathway and in disease resistance to *Botrytis cinerea*. *New Phytol.* **2013**, *198*, 1191–1202. [CrossRef] [PubMed]
61. Ji, Y.; Xiao, J.; Shen, Y.; Ma, D.; Li, Z.; Pu, G.; Li, X.; Huang, L.; Liu, B.; Ye, H.; et al. Cloning and characterization of AabHLH1, a bHLH transcription factor that positively regulates artemisinin biosynthesis in *Artemisia annua*. *Plant Cell Physiol.* **2014**, *55*, 1592–1604. [CrossRef] [PubMed]
62. Zhang, F.; Fu, X.; Lv, Z.; Lu, X.; Shen, Q.; Zhang, L.; Zhu, M.; Wang, G.; Sun, X.; Liao, Z. A basic leucine zipper transcription factor, AabZIP1, connects abscisic acid signaling with artemisinin biosynthesis in *Artemisia annua*. *Mol. Plant* **2015**, *8*, 163–175. [CrossRef] [PubMed]
63. Majid, I.; Kumar, A.; Abbas, N. A basic helix loop helix transcription factor, AaMYC2-Like positively regulates artemisinin biosynthesis in *Artemisia annua* L. *Ind. Crop Prod.* **2019**, *128*, 115–125. [CrossRef]
64. Shen, Q.; Lu, X.; Yan, T.; Fu, X.; Lv, Z.; Zhang, F.; Pan, Q.; Wang, G.; Sun, X.; Tang, K. The jasmonate-responsive AaMYC2 transcription factor positively regulates artemisinin biosynthesis in *Artemisia annua*. *New Phytol.* **2016**, *210*, 1269–1281. [CrossRef]
65. Shen, Q.; Zhang, L.; Liao, Z.; Wang, S.; Yan, T.; Shi, P.; Liu, M.; Fu, X.; Pan, Q.; Wang, Y. The genome of *Artemisia annua* provides insight into the evolution of Asteraceae family and artemisinin biosynthesis. *Mol. Plant* **2018**, *11*, 776–788. [CrossRef]
66. Dubos, C.; Stracke, R.; Grotewold, E.; Weisshaar, B.; Martin, C.; Lepiniec, L. MYB transcription factors in Arabidopsis. *Trends Plant Sci.* **2010**, *15*, 573–581. [CrossRef]
67. Matías-Hernández, L.; Jiang, W.; Yang, K.; Tang, K.; Brodelius, P.E.; Pelaz, S. AaMYB1 and its orthologue AtMYB61 affect terpene metabolism and trichome development in *Artemisia annua* and *Arabidopsis thaliana*. *Plant J.* **2017**, *90*, 520–534. [CrossRef]

68. Zvi, M.M.B.; Shklarman, E.; Masci, T.; Kalev, H.; Debener, T.; Shafir, S.; Ovadis, M.; Vainstein, A. PAP1 transcription factor enhances production of phenylpropanoid and terpenoid scent compounds in rose flowers. *New Phytol.* **2012**, *195*, 335–345. [CrossRef]
69. Gonzalez, A.; Zhao, M.; Leavitt, J.M.; Lloyd, A.M. Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in Arabidopsis seedlings. *Plant J.* **2008**, *53*, 814–827. [CrossRef]
70. Roth, G.A.; Johnson, C.; Abajobir, A.; Abd-Allah, F.; Abera, S.F.; Abyu, G.; Ahmed, M.; Aksut, B.; Alam, T.; Alam, K. Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J. Am. Coll. Cardiol.* **2017**, *70*, 1–25. [CrossRef] [PubMed]
71. Yuan, Y.; Yu, M.; Jia, Z.; Song, X.; Liang, Y.; Zhang, J. Analysis of *Dendrobium huoshanense* transcriptome unveils putative genes associated with active ingredients synthesis. *BMC Genom.* **2018**, *19*, 978. [CrossRef] [PubMed]
72. Ge, Y.; Cheng, Z.; Si, X.; Ma, W.; Tan, L.; Zang, X.; Wu, B.; Xu, Z.; Wang, N.; Zhou, Z. Transcriptome profiling provides insight into the genes in carotenoid biosynthesis during the mesocarp and seed developmental stages of avocado (*Persea americana*). *Int. J. Mol. Sci.* **2019**, *20*, 4117. [CrossRef] [PubMed]
73. Jensen, S.R.; Franzyk, H.; Wallander, E. Chemotaxonomy of the Oleaceae: Iridoids as taxonomic markers. *Phytochemistry* **2002**, *60*, 213–231. [CrossRef]
74. Zhang, C.; Ma, X.; Zhu, C.; Chen, Z.; Li, W.; Zhao, X.; He, S.; Du, H. Analysis on the content of two chemical composition in different parts of cultivated and wild *Centranthera grandiflora*. *Guihaia* **2019**, 1–7. [CrossRef]
75. Zhang, F.; Liu, W.; Xia, J.; Zeng, J.; Xiang, L.; Zhu, S.; Zheng, Q.; Xie, H.; Yang, C.; Chen, M. Molecular characterization of the 1-deoxy-D-xylulose 5-phosphate synthase gene family in *Artemisia annua*. *Front. Plant Sci.* **2018**, *9*, 952. [CrossRef]
76. Palazón, J.; Cusidó, R.M.; Bonfill, M.; Morales, C.; Piñol, M.T. Inhibition of paclitaxel and baccatin III accumulation by mevinolin and fosmidomycin in suspension cultures of *Taxus baccata*. *J. Biotechnol.* **2003**, *101*, 157–163. [CrossRef]
77. Grąbkowska, R.; Mielicki, W.; Wielanek, M.; Wysokińska, H. Changes of phenylethanoid and iridoid glycoside distribution in various tissues of shoot cultures and regenerated plants of *Harpagophytum procumbens* (Burch.) DC. ex Meisn. *S. Afr. J. Bot.* **2014**, *95*, 159–164. [CrossRef]
78. Zhang, X.; Liu, C.J. Multifaceted regulations of gateway enzyme phenylalanine ammonia-lyase in the biosynthesis of phenylpropanoids. *Mol. Plant* **2015**, *8*, 17–27. [CrossRef]
79. Khakdan, F.; Alizadeh, H.; Ranjbar, M. Molecular cloning, functional characterization and expression of a drought inducible phenylalanine ammonia-lyase gene (*ObPAL*) from *Ocimum basilicum* L. *Plant Physiol. Biochem.* **2018**, *130*, 464–472. [CrossRef]
80. Xu, F.; Cai, R.; Cheng, S.; Du, H.; Wang, Y. Molecular cloning, characterization and expression of phenylalanine ammonia-lyase gene from *Ginkgo biloba*. *Afr. J. Biotechnol.* **2008**, *7*, 721–729. [CrossRef]
81. Song, J.; Wang, Z. Molecular cloning, expression and characterization of a phenylalanine ammonia-lyase gene (*SmPAL1*) from *Salvia miltiorrhiza*. *Mol. Biol. Rep.* **2009**, *36*, 939–952. [CrossRef] [PubMed]
82. Zhang, Y.; Fu, X.; Hao, X.; Zhang, L.; Wang, L.; Qian, H.; Zhao, J. Molecular cloning and promoter analysis of the specific salicylic acid biosynthetic pathway gene phenylalanine ammonia-lyase (*AaPAL1*) from *Artemisia annua*. *Biotechnol. Appl. Biochem.* **2016**, *63*, 514–524. [CrossRef] [PubMed]
83. Chang, A.; Lim, M.; Lee, S.; Robb, E.J.; Nazar, R.N. Tomato phenylalanine ammonia-lyase gene family, highly redundant but strongly underutilized. *J. Biol. Chem.* **2008**, *283*, 33591–33601. [CrossRef] [PubMed]
84. Kellogg, E.A. Different ways to be redundant. *Nat Genet* **2019**, *51*, 770–771. [CrossRef]
85. Lachowiec, J.; Mason, G.A.; Schultz, K.; Queitsch, C. Redundancy, feedback, and robustness in the *Arabidopsis thaliana BZR/BEH* gene family. *Front. Genet.* **2018**, *9*, 523. [CrossRef]
86. Whitaker, J.R. Polyphenol Oxidase. In *Food Enzymes: Structure and Mechanism*; Wong, D.W.S., Ed.; Springer: Boston, MA, USA, 1995; pp. 271–307. [CrossRef]
87. Taranto, F.; Pasqualone, A.; Mangini, G.; Tripodi, P.; Miazzi, M.; Pavan, S.; Montemurro, C. Polyphenol oxidases in crops: Biochemical, physiological and genetic aspects. *Int. J. Mol. Sci.* **2017**, *18*, 377. [CrossRef]
88. Tran, L.T.; Taylor, J.S.; Constabel, C.P. The polyphenol oxidase gene family in land plants: Lineage-specific duplication and expansion. *BMC Genom.* **2012**, *13*, 395. [CrossRef]
89. Vaughn, K.; Lax, A.; Duke, S. Polyphenol oxidase. *Handb. Plant Cytochem.* **1987**, *1*, 159–162.

90. Verma, R.; Tapwal, A.; Kumar, D.; Puri, S. Assessment of Antimicrobial Potential and Phytochemical Profiling of Ethnomedicinal Plant *Bergenia ciliata* (haw.) sternb. in Western Himalaya. *J. Microbiol. Biotechnol Food Sci.* **2019**, *9*, 15–20. [CrossRef]
91. Sharma, S.K.; Patil, A.; Agnihotri, A.K.; Mehrotra, S. In vitro conservation of *Alectra chitrakutensis*: A critically endangered root parasitic plant of high medicinal importance. *Acta Physiol. Plant.* **2018**, *40*, 29. [CrossRef]
92. Michael, E.; Mohan, V.R. Determination of bioactive components of *Caralluma umbellata* haw. (apocynaceae) by gas chromatography and mass spectroscopy analysis. *Asian J. Pharm. Clin. Res.* **2018**, *11*, 194–199. [CrossRef]
93. Agrawal, P.; Laddha, K.; Tiwari, A. Isolation and HPLC method development of azafrin from *Alectra parasitica* var. chitrakutensis. *Nat. Prod. Res.* **2014**, *28*, 940–944. [CrossRef] [PubMed]
94. Paul, M. *Medicinal Natural Products: A Biosynthetic Approach*, 3rd ed.; John Wiley and Sons, Ltd.: West Sussex, UK, 2009; p. 303. [CrossRef]
95. Brocker, C.; Vasiliou, M.; Carpenter, S.; Carpenter, C.; Zhang, Y.; Wang, X.; Kotchoni, S.O.; Wood, A.J.; Kirch, H.; Kopečný, D. Aldehyde dehydrogenase (ALDH) superfamily in plants: Gene nomenclature and comparative genomics. *Planta* **2013**, *237*, 189–210. [CrossRef] [PubMed]
96. Cheng, X.; Zhao, X.; Huang, C.; Zhang, X.; Lyu, Y. Lutein content in petals and leaves of marigold and analysis of lutein synthesis gene expression. *Acta Physiol. Plant.* **2019**, *41*, 128. [CrossRef]
97. Hyun, T.K.; Rim, Y.; Jang, H.; Kim, C.H.; Park, J.; Kumar, R.; Lee, S.; Kim, B.C.; Bhak, J.; Nguyen-Quoc, B. De novo transcriptome sequencing of *Momordica cochinchinensis* to identify genes involved in the carotenoid biosynthesis. *Plant Mol. Biol.* **2012**, *79*, 413–427. [CrossRef]
98. Lin, H.; Wang, R.; Qian, Q.; Yan, M.; Meng, X.; Fu, Z.; Yan, C.; Jiang, B.; Su, Z.; Li, J. DWARF27, an iron-containing protein required for the biosynthesis of strigolactones, regulates rice tiller bud outgrowth. *Plant Cell* **2009**, *21*, 1512–1525. [CrossRef]
99. Abuauf, H.; Haider, I.; Jia, K.; Ablazov, A.; Mi, J.; Blilou, I.; Al-Babili, S. The Arabidopsis *DWARF27* gene encodes an all-trans-/9-cis-β-carotene isomerase and is induced by auxin, abscisic acid and phosphate deficiency. *Plant Sci.* **2018**, *277*, 33–42. [CrossRef]
100. Vallabhaneni, R.; Bradbury, L.M.; Wurtzel, E.T. The carotenoid dioxygenase gene family in maize, sorghum, and rice. *Arch. Biochem. Biophys.* **2010**, *504*, 104–111. [CrossRef]
101. Zhang, X.; Allan, A.; Li, C.; Wang, Y.; Yao, Q. *De novo* assembly and characterization of the transcriptome of the Chinese medicinal herb, *Gentiana rigescens*. *Int. J. Mol. Sci.* **2015**, *16*, 11550–11573. [CrossRef]
102. Haas, B.J.; Papanicolaou, A.; Yassour, M.; Grabherr, M.; Blood, P.D.; Bowden, J.; Couger, M.B.; Eccles, D.; Li, B.; Lieber, M.; et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **2013**, *8*, 1494–1512. [CrossRef] [PubMed]
103. Davidson, N.M.; Oshlack, A. Corset: Enabling differential gene expression analysis for de novoassembled transcriptomes. *Genome Biol.* **2014**, *15*, 410. [CrossRef] [PubMed]
104. Paulino, P.R.; Diego Mauricio, R.O.P.; Corrêa, L.G.G.; Rensing, S.A.; Birgit, K.; Bernd, M.R. PlnTFDB: Updated content and new features of the plant transcription factor database. *Nucleic Acids Res.* **2010**, *38*, D822–D827. [CrossRef]
105. Young, M.D.; Wakefield, M.J.; Smyth, G.K.; Oshlack, A. Method Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biol.* **2010**, *11*, R14. [CrossRef] [PubMed]
106. Mao, X.; Cai, T.; Olyarchuk, J.G.; Wei, L. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* **2005**, *21*, 3787–3793. [CrossRef] [PubMed]

International Journal of
Molecular Sciences

MDPI

Article

Species Identification of Oaks (*Quercus* L., Fagaceae) from Gene to Genome

Xinbo Pang [1,2,3,4], Hongshan Liu [2], Suran Wu [2], Yangchen Yuan [2], Haijun Li [2], Junsheng Dong [2], Zhaohua Liu [2], Chuanzhi An [2], Zhihai Su [2] and Bin Li [1,2,3,4,*]

[1] Research Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, China; pxb15633717296@163.com
[2] Administration Bureau of Hongyashan State Owned Forest Farm in Yixian County, Yixian 074200, China; lhs13803120634@163.com (H.L.); wsr15830855155@163.com (S.W.); 18730272192@163.com (Y.Y.); lhj13831238335@163.com (H.L.); 13930286026@163.com (J.D.); yyc16603261128@163.com (Z.L.); Anchuanzhi2002@163.com (C.A.); szh13833020580@163.com (Z.S.)
[3] State Key Laboratory of Tree Genetics and Breeding, Chinese Academy of Forestry, Beijing 100091, China
[4] Key Laboratory of Tree Breeding and Cultivation of State Forestry Administration, Chinese Academy of Forestry, Beijing 100091, China
* Correspondence: libin1200@163.com

Received: 24 September 2019; Accepted: 8 November 2019; Published: 26 November 2019

Abstract: Species identification of oaks (*Quercus*) is always a challenge because many species exhibit variable phenotypes that overlap with other species. Oaks are notorious for interspecific hybridization and introgression, and complex speciation patterns involving incomplete lineage sorting. Therefore, accurately identifying *Quercus* species barcodes has been unsuccessful. In this study, we used chloroplast genome sequence data to identify molecular markers for oak species identification. Using next generation sequencing methods, we sequenced 14 chloroplast genomes of *Quercus* species in this study and added 10 additional chloroplast genome sequences from GenBank to develop a DNA barcode for oaks. Chloroplast genome sequence divergence was low. We identified four mutation hotspots as candidate *Quercus* DNA barcodes; two intergenic regions (*matK-trnK-rps16* and *trnR-atpA*) were located in the large single copy region, and two coding regions (*ndhF* and *ycf1b*) were located in the small single copy region. The standard plant DNA barcode (*rbcL* and *matK*) had lower variability than that of the newly identified markers. Our data provide complete chloroplast genome sequences that improve the phylogenetic resolution and species level discrimination of *Quercus*. This study demonstrates that the complete chloroplast genome can substantially increase species discriminatory power and resolve phylogenetic relationships in plants.

Keywords: oak species identification; chloroplast genome; *Quercus*; mutation hotspots

1. Introduction

DNA barcoding has recently emerged as a new molecular tool for species identification [1]. A DNA barcode is a short, standardized DNA region normally employed for species identification. The mitochondrial gene cytochrome oxidase 1 (*COI*) is an effective and reliable standard animal DNA barcode for species identification [1]. Over the past 10 years, plant DNA barcode researchers have been evaluating the proposed barcode segments of plants. Previously proposed barcode segments exist primarily in chloroplast genomes that are relatively stable, single-copy, and easy to amplify. These proposed barcodes are *matK*, *rbcL*, *ropC1*, and *rpoB* in the coding region, and *atpF-H*, *trnL-F*, *trnH-psbA*, and *psbK-I* in the non-coding region [2]. At the third DNA barcode conference held in Mexico City in 2009, the majority of the Consortium for the Barcode of Life (CBOL) Plant Working Group preferred to recommend a core-barcode combination consisting of portions of two plastid coding regions, *rbcL* and *matK*, which are supplemented with additional markers (such as *trnH-psbA* and

Int. J. Mol. Sci. **2019**, *20*, 5940; doi:10.3390/ijms20235940 www.mdpi.com/journal/ijms

internal transcribed spacers [ITS]) as required. In 2011, the China Plant BOL Group suggested using ITS as the plant DNA barcode [3]. However, increasing numbers of studies show that core-barcodes remain problematic, especially in recently diverged and rapidly radiated taxa [4–6].

With the development of next-generation sequencing (NGS), the number of sequenced chloroplast genomes has increased rapidly, making it possible to generate chloroplast genome data to extend the concept of DNA barcoding for plant species identification [6–9]. The DNA barcoding approaches for species identification has extended from gene to genome, promptly extending phylogeny analysis from gene-based phylogenetics to phylogenomics. Chloroplast genome sequences are a primary source of data for inferring plant phylogenies and DNA barcoding because of their conserved gene content and genome structure, low nucleotide substitution mutation rates, usually uni-parental inheritance, and the low cost of generating whole chloroplast genomes with high throughput sequencing. Using chloroplast genome data, longstanding controversies at various taxonomic levels have been resolved [10–12], suggesting its power in resolving evolutionary relationships. However, challenges still exist in establishing phylogeny relationships and discrimination of closely related, recently divergent, hybridized, or introgressed lineages such as the oak group.

Oaks (*Quercus* L., Fagaceae) comprise approximately 400–500 species that are widespread throughout the temperate zones of the Northern Hemisphere; they are dominant, diverse forest and savannah angiosperm trees and shrubs belonging to a taxonomically complex group. The taxonomy of oak species remains controversial and incomplete, owing to the overlapping variation of individuals and population produced by ecological adaptation and differential reproductive isolation. A series of phylogenetic and DNA barcoding studies have mainly used several chloroplast DNA markers [13,14] such as *rbcL*, *rpoC1*, *trnH-psbA*, *matK*, *ycf3-trnS*, *ycf1*, and the nuclear ribosomal DNA ITS [4,15–17]. These studies focused only on regional flora, and those markers revealed low sequence divergence leading to lower discrimination success [4,18]. Yang et al. [13] compared two closely related species (*Quercus rubra* and *Castanea mollissima*) by exploring nine highly variable chloroplast DNA markers for species identification. However, the results showed a very low discrimination success rate using a single marker and all their combinations. On the other hand, oaks are notorious for interspecific hybridization and introgression, as well as complex speciation patterns involving incomplete lineage sorting [19–21], which have possible negative effects for barcoding and phylogeny of the species-rich *Quercus* genus [4].

In this study, we sequenced the complete chloroplast genome of 14 *Quercus* species and combined the previously reported chloroplast genomes of 10 other *Quercus* species in order to provide a comparative analysis. The study aimed to (1) investigate the genome structure, gene order, and gene content of the whole chloroplast genome of multiple *Quercus* species; (2) test whether chloroplast genome data yielded sufficient variation to construct a well-supported phylogeny of *Quercus* species; and (3) determine if multiple variable markers or whole chloroplast genome data can be successfully used for oaks species identification.

2. Results

2.1. General Features of the Quercus Chloroplast Genome

Using the Illumina HiSeq X Ten system, 14 *Quercus* species were sequenced to produce 9,910,273–16,862,000 paired-end raw reads (150 bp average read length), with an average sequencing depth of 162× to 480× (Table S1). To validate the accuracy of the assembled chloroplast genome, we carried out Sanger sequencing of PCR amplicons spanning the junction regions (LSC/IRA, LSC/IRB, SSC/IRA, and SSC/IRB). The 14 *Quercus* chloroplast genome sequences were deposited in GenBank (accession numbers MK105451–MK105453, MK105456–MK105464, and MK105466-MK105467).

The total chloroplast genome sequence lengths of 14 *Quercus* species ranged from 161,132 bp (*Q. phillyraeoides*) to 161,366 bp (*Q. rubra*). These genomes displayed typical circular quadripartite structure consisting of a pair of IR regions (25,817–25,870 bp) separated by an LSC region

(90,363–90,624 bp) and an SSC region (18,946–19,073 bp) (Figure 1). The overall GC content was absolutely identical (36.8%; Table 1) across all plastomes, but was clearly higher in the IR region (42.8%) than in the other regions (LSC 34.7%; SSC 30.9%), possibly because of the high GC content of the rRNA that was located in the IR regions. All plastomes possessed 113 unique genes, including 79 protein-coding genes, 30 tRNA genes, and 4 rRNA genes. Among the unique genes, 15 genes contained one intron, and two genes contained two introns.

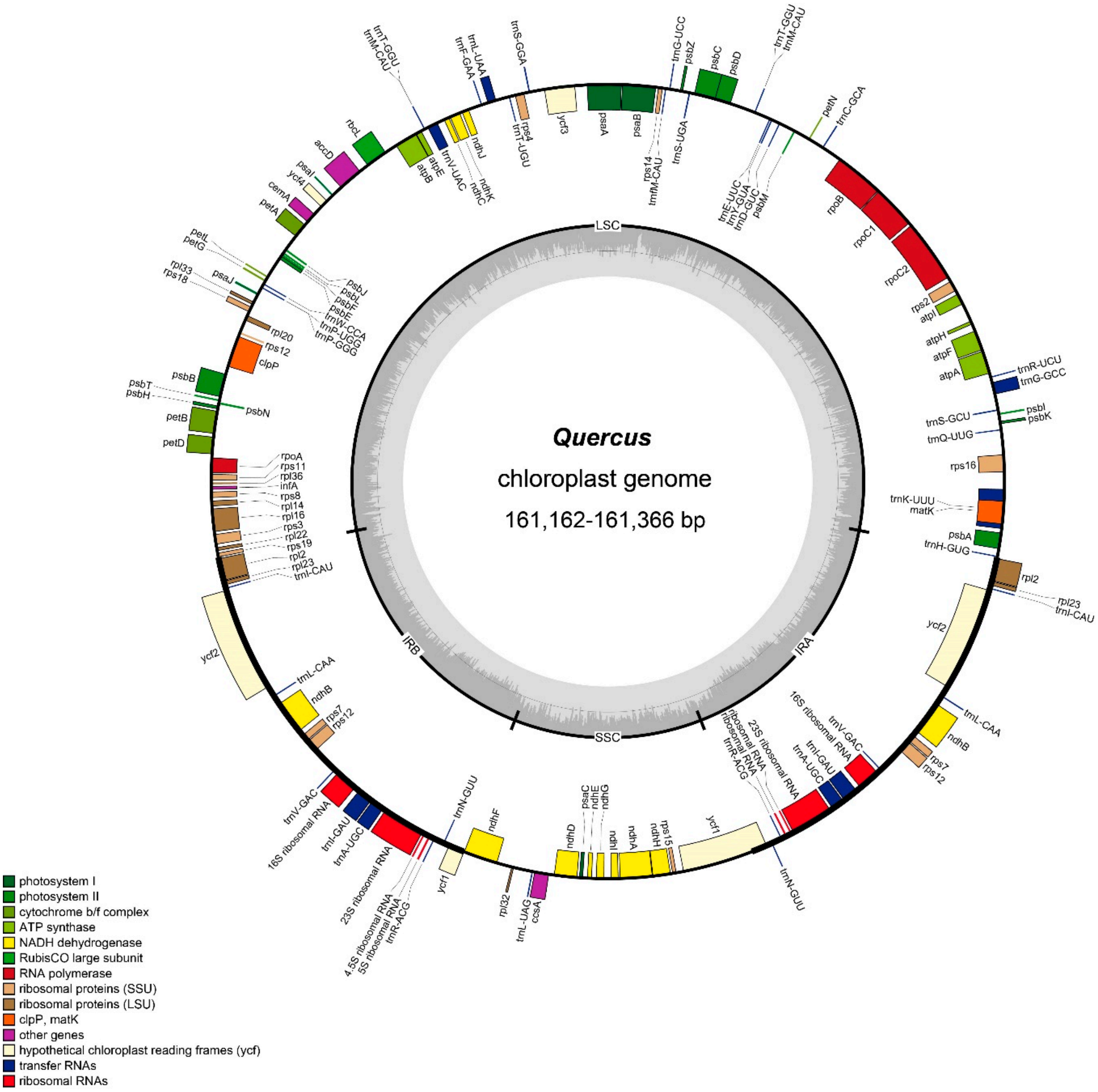


Figure 1. Gene map of *Quercus* chloroplast genome. Genes drawn within the circle are transcribed clockwise; genes drawn outside are transcribed counterclockwise. Genes in different functional groups are shown in different colors. Dark bold lines indicate the extent of the inverted repeats (IRa and IRb) that separate the genomes into small single-copy (SSC) and large single-copy (LSC) regions.

The chloroplast genome results showed that all 14 *Quercus* plastomes were remarkably similar in terms of size, genes, and genome structures. The LSC/IR and IR/SSC boundaries were conserved. *Rps19* was located in the LSC near the LSC/IRb, and *trnH-GUG* was located in the LSC near the IRa/LSC border. Additionally, the location of the SSC/IRa junction was within the coding region of the *ycf1* gene.

Table 1. Summary statistics for the assembly of 14 *Quercus* species chloroplast genomes.

Species	LSC	IR	SSC	Total Size (bp)	Number of Genes	Protein Coding Genes	tRNA	rRNA	Accession Number in Genbank
Q. macrocarpa	90594	25848	18946	161236	113	79	30	4	MK105459
Q. gambelii	90570	25848	18947	161213	113	79	30	4	MK105457
Q. stellata	90562	25848	18956	161214	113	79	30	4	MK105467
Q. palustris	90624	25852	18956	161284	113	79	30	4	MK105461
Q. aliena var. *acuteserrata*	90532	25837	18988	161194	113	79	30	4	MK105452
Q. phillyraeoides	90363	25866	19037	161132	113	79	30	4	MK105462
Q. glandulifera var. *brevipetiolata*	90534	25826	19038	161224	113	79	30	4	MK105458
Q. wutaishanica	90520	25825	19041	161211	113	79	30	4	MK105466
Q. mongolica	90504	25820	19047	161191	113	79	30	4	MK105460
Q. dentata	90593	25826	19055	161300	113	79	30	4	MK105453
Q. fabri	90557	25832	19064	161285	113	79	30	4	MK105456
Q. serrata	90447	25817	19065	161146	113	79	30	4	MK105464
Q. variabilis	90464	25817	19070	161168	113	79	30	4	MK105451
Q. rubra	90553	25870	19073	161366	113	79	30	4	MK105463

2.2. Phylogenetic Analyses

The matrix of whole chloroplast genome sequences was used to reconstruct the *Quercus* phylogenetic tree (Figure 2). Both maximum likelihood and Bayesian analyses produced similar topologies for the 24 species and were highly branch supported. All the sampled *Quercus* species were clustered into one clade with 100% bootstrap value (BS) or Bayesian posterior probability (PP). However, backbone branch supports were relatively poor, as were some internal branches. Moreover, six major clades were identified in *Quercus* and the analyses obtained high support for all six of the nodes.

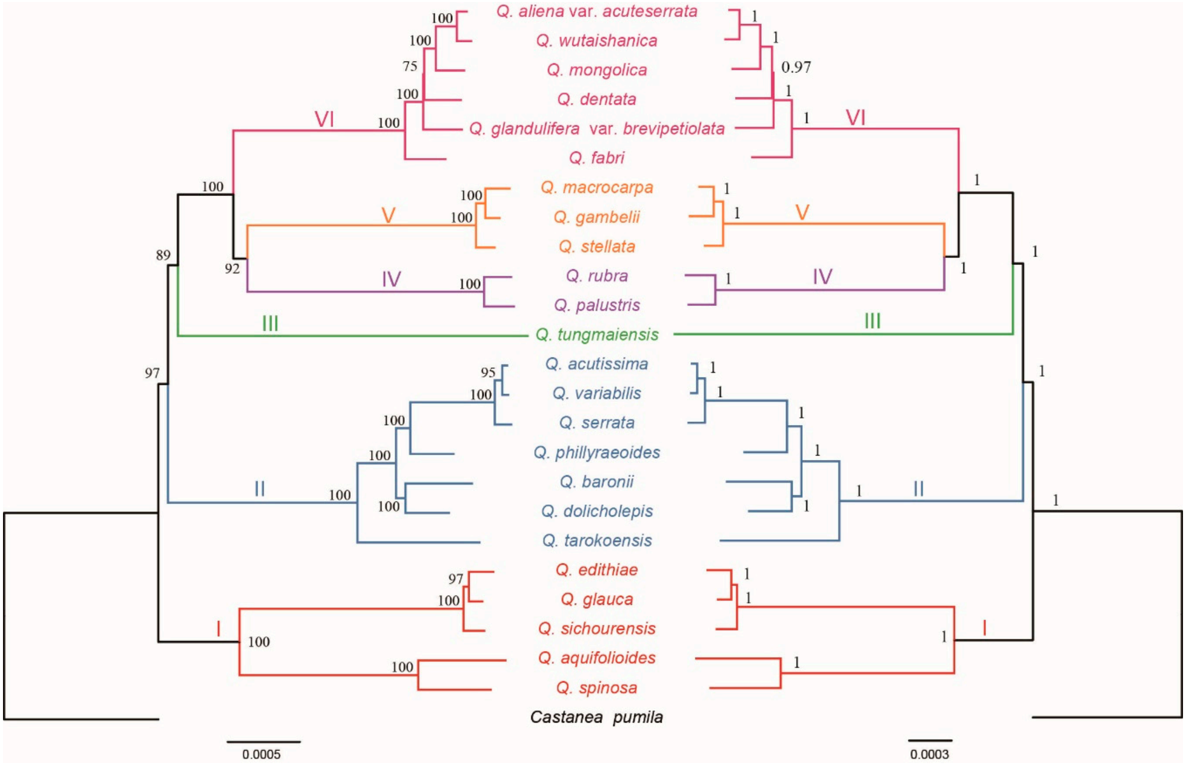


Figure 2. Phylogenetic tree inferred from the 25 chloroplast genomes. Left: Maximum likelihood tree with maximum likelihood (ML) bootstrap values; right: Bayesian tree with posterior probabilities.

Clade I on the base of the tree (BS = 100% and PP = 1) comprised *Q. edithiae*, *Q. gambelii*, *Q. sichourensis*, *Q. aquifolioides*, and *Q. spinosa* being the earliest diverging lineages. Clade II (BS = 100% and PP = 1) contained seven species: *Q. acutissima*, *Q. variabilis*, *Q. serrata*, *Q. phillyraeoides*, *Q. dolicholepis*, *Q. baronii*, and *Q. tarokoensis*. Clade III only contained *Q. tungmaiensis*. *Q. rubra* and *Q. palustris* formed clade IV, which was identified as a sister to clade V with high support value (BS = 92% and PP = 1). Clade V included three species, *Q. macrocarpa*, *Q. glauca*, and *Q. stellata*. The last clade (BS = 100% and PP = 1) was made up of *Q. aliena* var. *acuteserrata*, *Q. wutaishanica*, *Q. mongolica*, *Q. fabri*, *Q. glandulifera* var. *brevipetiolata*, and *Q. dentata*.

2.3. Analyses of the Standard DNA Barcodes

The *trnH-psbA* intergenic spacer region ranged from 412 bp to 474 bp with 27 variable sites, 16 informative sites, and nine indels of 3–20 bp within 574 aligned bp. A small 32 bp inversion occurred at 454 bp. *RbcL* and *matK* genes, both without indels, were 698 bp with eight variable and five informative sites, and 744 bp with 21 variable and 11 informative sites, respectively (Table 2). The mean interspecific genetic distances of the 24 oaks species with K2P were 0.0026 for *rbcL*, 0.0048 for *matK*, and 0.0125 for *trnH-psbA*. Based on the distance method, the universal DNA barcode had less discriminatory power; *rbcL*, *matK*, and *trnH-psbA* had only a 12.50%, 25.00%, and 37.50% success rate, respectively. With the two core DNA barcodes (*rbcL* and *matK*) combined, success was only 29.17%. Combined analyses of *rbcL*, *matK*, and *trnH-psbA* or *rbcL* and *matK* generated lower branch supported trees (Figure 3).

Table 2. The variability of the four new markers, chloroplast genome, and the universal chloroplast DNA barcodes in *Quercus*.

Markers	Length	Variable Sites		Information Sites		Discrimination Success (%) Based on Distance Method
		Numbers	%	Numbers	%	
rbcL	698	8	1.15%	5	0.72%	12.50%
matK	744	21	2.82%	11	1.48%	25.00%
trnH-psbA	574	27	4.70%	16	2.79%	37.50%
rbcL + *matK*	1442	29	2.01%	16	1.11%	29.17%
rbcL + *matK* + *trnH-psbA*	2016	56	2.78%	32	1.59%	50.00%
matK-trnK-rps16	2311	93	4.02%	59	2.55%	79.17%
trnR-atpA	1309	57	4.35%	35	2.67%	66.67%
ndhF	1536	74	4.82%	45	2.93%	83.33%
ycf1b	1765	94	5.33%	59	3.34%	70.83%
ndhF+*ycf1b*	3301	168	5.09%	104	3.15%	91.67%
matK-trnK-rps16 + *trnR-atpA* + *ndhF* + *ycf1b*	6921	318	4.59%	198	2.86%	100.00%

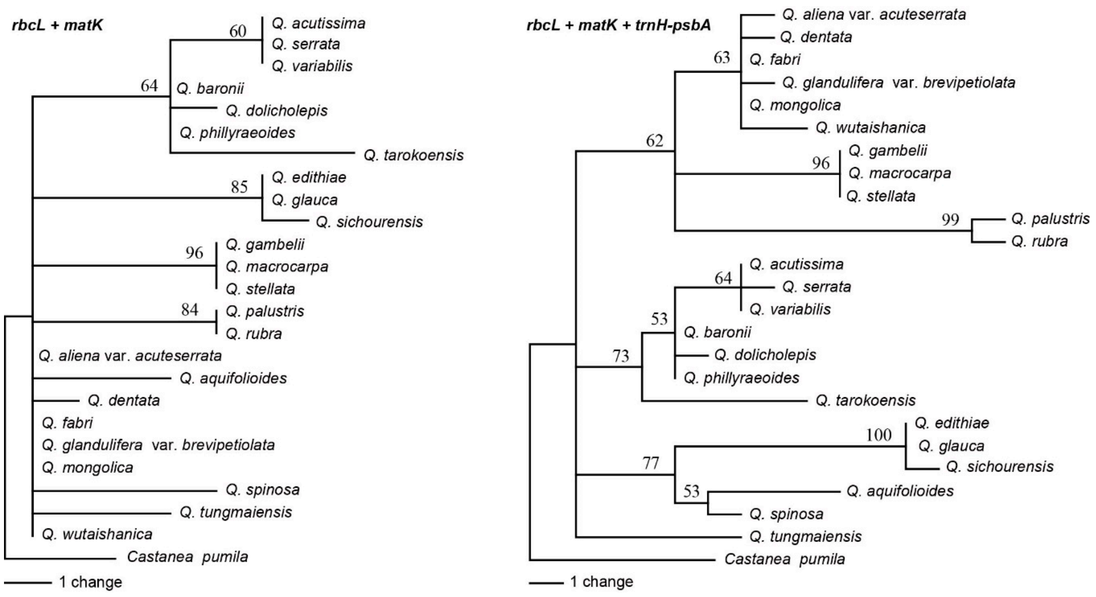


Figure 3. Neighbor joining trees for *Quercus* using *rbcL* + *matK*, *rbcL* + *matK*, and *trnH-psbA* combinations.

2.4. Analyses of Specific Barcodes

To identify closely related species, it is imperative to identify rapidly evolving markers. We used DNAsp and SPIDER to discover the variable mutation regions of the *Quercus* chloroplast genome (Figure 4). The nucleotide diversity (pi) value ranged from 0 to 0.01766 in the 800 bp window size, while the K2P-distance ranged from 0 to 0.0179. We found four relatively variable regions: *matK-trnK-rps16*, *trnR-atpA*, *ndhF*, and *ycf1b*. Two intergenic regions (*matK-trnK-rps16* and *trnR-atpA*) were located in the LSC region, and two coding regions (*ndhF* and *ycf1b*) in the SSC region. We designed new primers for four variable regions (Table S3).

The *ycf1b* marker possessed the highest variability (5.33%), followed by the *ndhF* (4.82%), *trnR-atpA* (4.35%), and *matK-trnK-rps16* (4.02%) regions. Of the four variable makers, *ndhF* had the highest rate of correct identifications (83.33%), followed by *matK-trnK-rps16* (79.17%) and *ycf1b* (70.83%). Combining the four variable markers produced the most correct identifications (100%). The NJ tree-based method generated a graphical representation of the results and they were the same as those of the distance-based method (Figure 5).

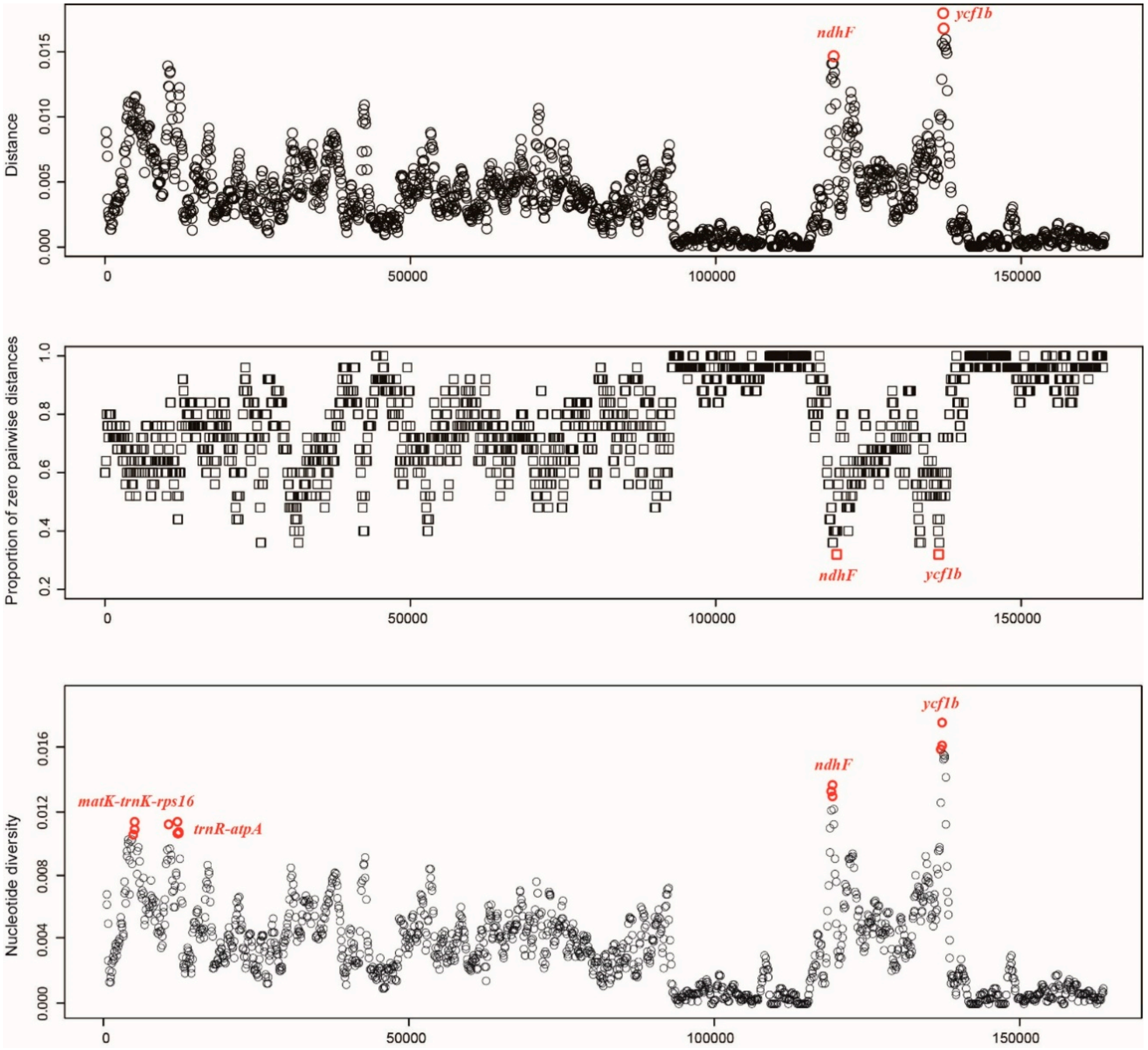


Figure 4. Specific DNA barcode development. (**A**) Mean distance of each window; (**B**) proportion of zero pairwise distances for each species; (**C**) nucleotide diversity (pi) of each window. Window length: 800 bp; Step size: 100 bp; X-axis: position of the midpoint of a window.

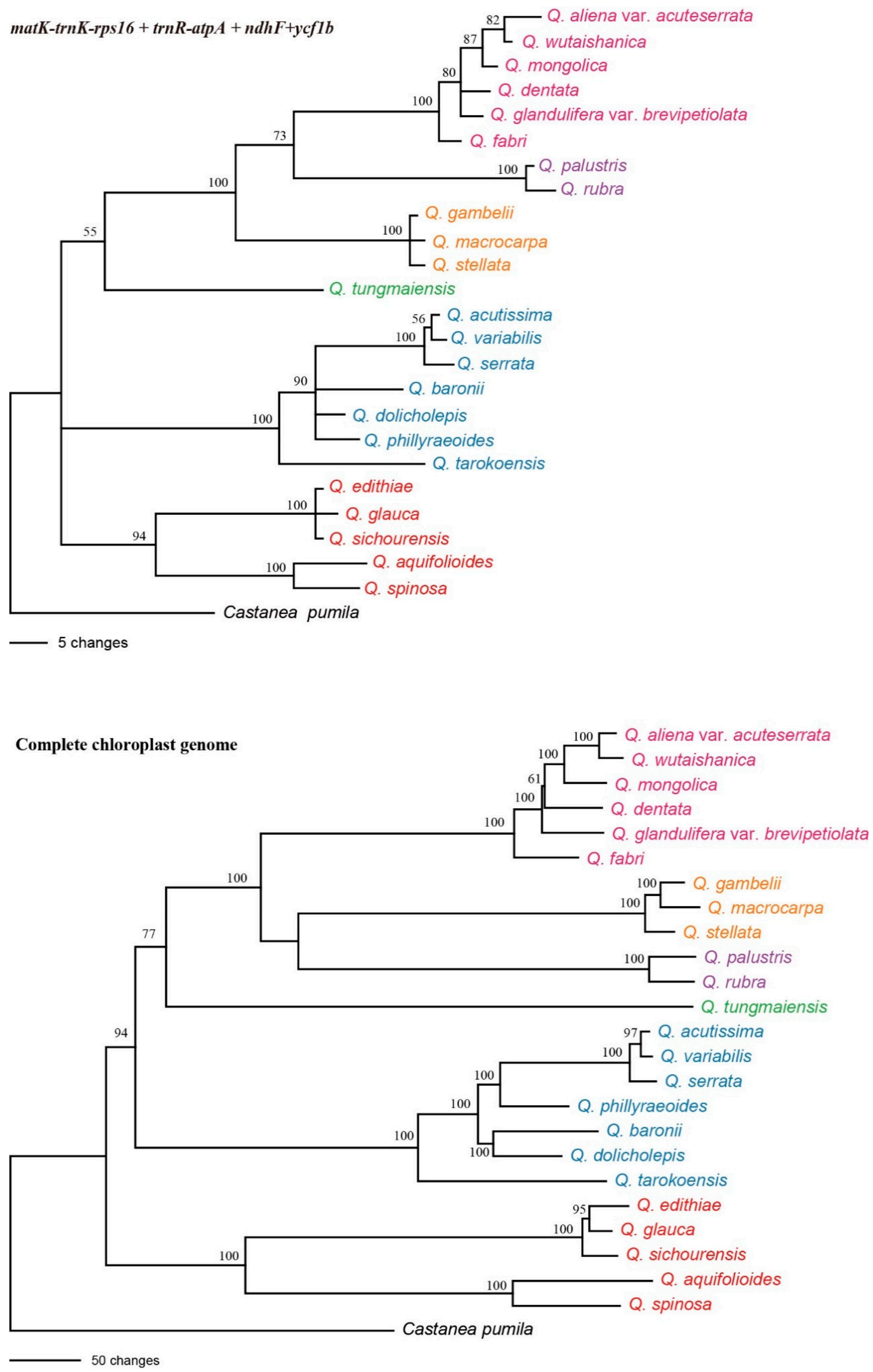


Figure 5. Neighbor joining tree for *Quercus* using the four highly variable markers and complete chloroplast genome data.

2.5. Super-Barcode

The 24 *Quercus* chloroplast genomes were fully aligned, and an alignment matrix of 164,156 bp was obtained (Table 3). We identified 2778 variable sites (1.69%), including 1727 parsimony-informative sites (1.05%), in the total chloroplast genome. The average Pi value for the 24 *Quercus* chloroplast genomes was 0.00335. Among these regions, IR exhibited the least nucleotide diversity (0.00073) and SSC exhibited high divergence (0.00624).

Table 3. Variable site analyses in *Quercus* chloroplast genomes.

	Number of Sites	Variable Sites		Information Sites		Nucleotide Diversity
		Numbers	%	Numbers	%	
LSC	92,888	2009	2.16%	1257	1.35%	0.0043
SSC	19,535	593	3.04%	368	1.88%	0.00624
IR	25,879	91	0.35%	54	0.21%	0.00073
Complete chloroplast genome	164,156	2778	1.69%	1727	1.05%	0.00335

To estimate the genetic divergence among *Quercus* chloroplast genomes, nucleotide substitutions and p-distance were calculated using MEGA. The overall sequence divergence estimated by p-distance among the 24 chloroplast genome sequences was only 0.0036. The number of nucleotide substitutions among the 24 species ranged from 14 to 734, and the p-distance ranged from 0.0001 to 0.0046. *Q. tungmaiensis* and *Q. serrata* had the largest sequence divergence. *Q. variabilis* had only 14 nucleotide substitutions with *Q. acutissima*.

The discriminatory power of the complete chloroplast genome as a DNA barcode was assessed using distance and tree-based methods. Compared to the standard DNA barcode or the four newly identified markers (specific barcodes), the complete chloroplast genome had the highest discriminatory power (Table 2 and Figure 5).

3. Discussion

Species delimitation remains one of the most controversial topics in biology. However, the accurate discrimination of material using only morphological characteristics is difficult. DNA barcoding is a widely used and effective tool that has enabled rapid and accurate identification of plant species since its development in 2003 [1]. Though DNA barcoding technology has developed significantly, no barcode can achieve the goal of sophisticated plant species identification [2]. In plants, the determination of a standardized barcode has been more complex. At present, increasing amounts of practical research tend to use chloroplast markers, such as *atpB-rbcL*, *atpF-H*, *matK*, *rbcL*, *psbK-I*, *rpoB*, *rpoC1*, *trnH-psbA*, and *trnL-F*, to identify species because of their relatively low evolutionary rates compared to those of nuclear loci and universal PCR primers [22–25]. The CBOL Working Group recently recommended a two-locus combination of *matK* + *rbcL* as the core plant barcode, with the recommendation to complement these using *trnH-psbA* and the ITS of the nuclear ribosomal DNA. However, because of the lower variability in standard DNA barcodes, discrimination power was low in plants [26]. In this study, the combination of *rbcL*, *matK*, and *trnH-psbA* had poor resolution (less than 50%) within *Quercus* (Table 2). Using the universal DNA barcode, the 12 Italian oak species revealed extremely low discrimination success (0%) [4]. Combined five chloroplast genome markers (*psbA-trnH, matK-trnK, ycf3-trnS, matK*, and *ycf1*), the species identification powers were only less than 20% [13]. Thus, there is an ongoing drive to develop additional oak barcodes.

With sequencing method development, greater numbers of DNA sequences were easily acquired. Identification of specific barcodes was an effective strategy for barcoding complex groups. Most studies showed that chloroplast genome mutations were clustered into hotspots, and those hotspots were defined as DNA barcodes [27–30]. The strategy of searching the complete chloroplast genome has been successfully applied to *Oryza* [30], *Panax* [28], *Diospyros* [31], and *Dioscorea* [32]. By comparing

24 *Quercus* chloroplast genomes in the present study, we identified four oak-specific barcodes including *matK-trnK-rps16*, *trnR-atpA*, *ndhF,* and *ycf1b* (Figure 4). The *ycf1* gene was more variable than the *matK* and *rbcL* genes in most plant lineages, and recently has been the focus of a DNA barcoding and plant phylogeny study [14]. Furthermore, *ycf1* has previously provided a higher species resolution in *Quercus* [13,14]. The *ndhF* gene has been widely used in plant phylogeny and is considered a variable coding gene in the chloroplast genome [27,33–35]. *MatK-trnK-rps16* and *trnR-atpA* are two interspace regions less commonly used as DNA barcode. Combined with the four highly variable markers, all 24 *Quercus* species were successfully identified using the distance method (Table 2).

Although the four specific barcodes had the highest discriminatory power, it was necessary to develop additional markers for *Quercus* because of its complex evolutionary history. With the advent of the next-generation DNA sequencing technologies, genomic data have extended the concept of DNA barcoding for species identification [6,8,36–38]. The DNA barcode has extended from gene or genes to the entire genome, and the extended DNA barcoding approach has been referred to as "ultra-barcoding" [39], "super-barcoding" [7], or "plant barcoding 2.0" [40]. Compared to the nuclear and mitochondrial genomes, the chloroplast genome is easily sequenced and may be the best-suited genome for plant species super-barcoding [36,41].

4. Materials and Methods

4.1. Taxon Sampling

The collection and GenBank accession information for taxa sampled in the present study are listed in Table 1 and Table S1. Ten species with previously sequenced chloroplast genomes used for analysis in this study are listed in Table S2. *Castanea pumila*, the sister group of *Quercus*, was used as the out-group.

4.2. DNA Extraction and Sequencing

We used an Illumina HiSeq X Ten platform to produce chloroplast genome sequences. *Quercus* species total DNA was extracted from silica-dried leaflets using the mCTAB protocol [42]. After extraction, total DNA was quantified with a Nanodrop 1000 Spectrophotometer. Fragmented samples of 350 bp were used to prepare paired-end libraries using a NEBNext®Ultra™DNA Library Prep Kit following the manufacturer's protocol. Each library that passed the first quality control step was tested with an Agilent 2100 Bio-147 analyzer (Agilent Technologies, Santa Clara, CA, USA) to ensure the libraries had the required size distributions. Real-time quantitative PCR was carried out to precisely measure library concentrations to balance the amounts used in multiplexed reactions. Paired-end sequencing (2 × 150 bp) was conducted on an Illumina HiSeq X Ten platform. For each species, approximately 5 Gb of raw data were generated.

4.3. Genome Assembly and Genome Annotation

A five-step approach was used to assemble the chloroplast genome. First, raw sequence reads were filtered for high quality reads by removing duplicate reads, as well as adapter-contaminated reads and reads with more than five Ns using the NGS QC Tool Kit [43]. Second, the SPAdes 3.6.1 program [44] was used for de novo assemblies. Third, chloroplast genome sequence contigs were selected from the SPAdes software by performing a BLAST search using the *Quercus variabilis* chloroplast genome sequence as a reference. Fourth, the Sequencher 5.4.5 program (Gene Codes Corp., Ann Arbor, Michigan, USA) was used to merge the selected contigs. Finally, small gaps or ambiguous nucleotides were bridged with specific primers designed for PCR based on their flanking sequences by Sanger sequencing. The four junctional regions between the IRs and small single copy (SSC) and large single copy (LSC) regions in the chloroplast genome sequences were further checked by PCR amplification and Sanger sequencing with specific primers as previously described [45].

Chloroplast genome annotation was performed with Plann [46] using the *Quercus variabilis* reference sequence. The chloroplast genome map was drawn using OGdraw online [47].

4.4. Phylogenetic Analyses

Multiple sequence alignment was performed using MAFFT v7 [48]. We estimated phylogenetic trees on the nucleotide substitution matrix using maximum likelihood (ML) and Bayesian inference (BI). ML analyses were performed using RAxML v.8.1.24 [49].

The RAxML analyses included 1000 bootstrap replicates in addition to a search for the best-scoring ML tree. BI was conducted with Mrbayes v3.2 [50]. The Metropolis-coupled Markov chain Monte Carlo (MCMC) algorithm was run for 50,000,000 generations with one cold and three heated chains, starting with a random tree and sampling one tree every 2000 generations. The first 25% of the trees were discarded as burn-in, and the remaining trees were used to build a 50% majority-rule consensus tree. Stationarity was considered reached when the average standard deviation of split frequencies remained below 0.01.

4.5. Sequence Divergence and Hotspot Identification

We analyzed the aligned sequences and counted the sequence divergence among *Quercus* chloroplast genomes to evaluate *Quercus* species divergence. Variable, parsimony-informative base sites, p-distances across the complete chloroplast genomes, and LSC, SSC, and inverted repeat (IR) regions of the 14 taxa were calculated using MEGA 6.0 software [51].

We used two methods to identify the hypervariable chloroplast genome regions. The first (nucleotide variability) was conducted using DnaSP version 5.1 software with the sliding window method. The second (genetic distance) was conducted using the *slideAnalyses* function of SPIDER [52] version 1.2-0 software. This function extracts all passable windows of a chosen size in a DNA alignment and performs pairwise distance (K2P) analyses of each window. The proportion of zero pairwise distances for each species and mean distance were considered for the definition of hypervariable regions. The step size was set to 100 bp with an 800 bp window length.

4.6. DNA Barcoding Analysis

To access the effectiveness of marker discriminatory performance, we used two methods to assess the barcoding resolution. The distance method used the *nearNeighbour* function of SPIDER software [52]. The distance method was used to analyze the barcode performances of newly identified highly variable regions.

Tree building analyses provide a convenient and visualized method for evaluating discriminatory performance by calculating the proportion of monophyletic species. A neighbor joining (NJ) tree was constructed for each hypervariable marker and different marker combinations using PAUP* 4.0 software [53]. Relative support for the NJ tree branches was assessed via 200 bootstrap replicates.

5. Conclusions

In this study, we sequenced and compared the chloroplast genomes of 24 *Quercus* species. The structure, size, and gene content of the *Quercus* chloroplast genomes were found to be well conserved, and comparative analyses revealed low levels of sequence variability. Four higher variable regions were identified, which were suitable as DNA barcodes for *Quercus* species identification. We also evaluated the resolution of the complete chloroplast genome in phylogenetic reconstruction and species discrimination in *Quercus.* The complete chloroplast genome sequence data produced strongly supported and highly resolved phylogenies in this taxonomically complex group despite the extensive hybridization and introgression in *Quercus*. Compared to standard plant DNA barcodes and the specific barcodes, analyses of the complete chloroplast genome sequences improved species identification resolution.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/20/23/5940/s1. Table S1. Sampling and assembly information for the 14 *Quercus* species; Table S2. A list of the 10 taxa sampled from GenBank in this study. Table S3. Primers for amplifying four highly variable loci.

Author Contributions: B.L. and X.P. designed the experiment; B.L., X.P., H.L., S.W., Y.Y., H.L., J.D., Z.L., C.A., Z.S. and P.H. collected samples and performed the experiment; B.L. and X.P. analyzed the data and wrote the manuscript; All of the authors have read and approved the final manuscript.

Funding: The study was funded by by "national forest germplasm resources bank of Quercus mongolica and Quercus variabilis in Hongyashan of Hebei (2017, 2018, 2019)".

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

LSC	Large single copy
SSC	Small single copy
IR	Inverted repeat

References

1. Hebert, P.D.N.; Cywinska, A.; Ball, S.L.; DeWaard, J.R. Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. B* **2003**, *270*, 313–321. [CrossRef]
2. Hollingsworth, P.M.; Graham, S.W.; Little, D.P. Choosing and using a plant DNA barcode. *PLoS ONE* **2011**, *6*, e19254. [CrossRef]
3. Groups, C.P.B.; Li, D.Z.; Gao, L.M.; Li, H.T.; Wang, H.; Ge, X.J.; Liu, J.Q.; Chen, Z.D.; Zhou, S.L.; Chen, S.L.; et al. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proc. Nat. Acad. Sci. USA* **2011**. [CrossRef]
4. Simeone, M.C.; Piredda, R.; Attimonelli, M.; Bellarosa, R.; Schirone, B. Prospects of barcoding the Italian wild dendroflora: Oaks reveal severe limitations to tracking species identity. *Mol. Ecol. Resour.* **2011**, *11*, 72–83. [CrossRef]
5. von Crautlein, M.; Korpelainen, H.; Pietilainen, M.; Rikkinen, J. DNA barcoding: A tool for improved taxon identification and detection of species diversity. *Biodivers. Conserv.* **2011**, *20*, 373–389. [CrossRef]
6. Coissac, E.; Hollingsworth, P.M.; Lavergne, S.; Taberlet, P. From barcodes to genomes: Extending the concept of DNA barcoding. *Mol. Ecol.* **2016**. [CrossRef]
7. Li, X.; Yang, Y.; Henry, R.J.; Rossetto, M.; Wang, Y.; Chen, S. Plant DNA barcoding: From gene to genome. *Biol. Rev.* **2015**, *90*, 157–166. [CrossRef]
8. Ruhsam, M.; Rai, H.S.; Mathews, S.; Ross, T.G.; Graham, S.W.; Raubeson, L.A.; Mei, W.; Thomas, P.I.; Gardner, M.F.; Ennos, R.A.; et al. Does complete plastid genome sequencing improve species discrimination and phylogenetic resolution in Araucaria? *Mol. Ecol. Resour.* **2015**. [CrossRef]
9. Fabiana, F.; Rizzo, Z.A.; Weismann, G.J.; Souza, O.R.; Lohmann, L.G.; Marie-Anne, V.S. Complete chloroplast genome sequences contribute to plant species delimitation: A case study of the Anemopaegma species complex. *Am. J. Bot.* **2017**, *104*, 1493–1509. [CrossRef]
10. Wu, C.S.; Chaw, S.M.; Huang, Y.Y. Chloroplast phylogenomics indicates that Ginkgo biloba is sister to cycads. *Genome Biol. Evol.* **2013**, *5*, 243–254. [CrossRef]
11. Cox, C.J.; Li, B.; Foster, P.G.; Embley, T.M.; Civan, P. Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. *Syst. Biol.* **2014**, *63*, 272–279. [CrossRef]
12. Carbonell-Caballero, J.; Alonso, R.; Ibañez, V.; Terol, J.; Talon, M.; Dopazo, J. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus *Citrus*. *Mol. Biol. Evol.* **2015**, *32*, 2015–2035. [CrossRef]
13. Yang, J.; Vazquez, L.; Chen, X.; Li, H.; Zhang, H.; Liu, Z.; Zhao, G. Development of Chloroplast and Nuclear DNA Markers for Chinese Oaks (Quercus Subgenus Quercus) and Assessment of Their Utility as DNA Barcodes. *Front. Plant Sci.* **2017**, *8*, 816. [CrossRef]
14. Dong, W.; Xu, C.; Li, C.; Sun, J.; Zuo, Y.; Shi, S.; Cheng, T.; Guo, J.; Zhou, S. ycf1, the most promising plastid DNA barcode of land plants. *Sci. Rep.* **2015**, *5*, 8348. [CrossRef]
15. Mayol, M.; Rossello, J.A. Why nuclear ribosomal DNA spacers (ITS) tell different stories in Quercus. *Mol. Phylogenet. Evol.* **2001**, *19*, 167–176. [CrossRef]

16. Bellarosa, R.; Simeone, M.C.; Papini, A.; Schirone, B. Utility of ITS sequence data for phylogenetic reconstruction of *Italian Quercus* spp. *Mol. Phylogenet. Evol.* **2005**, *34*, 355–370. [CrossRef]
17. Simeone, M.C.; Piredda, R.; Papini, A.; Vessella, F.; Schirone, B. Application of plastid and nuclear markers to DNA barcoding of Euro-Mediterranean oaks (Quercus, Fagaceae): Problems, prospects and phylogenetic implications. *Bot. J. Linn. Soc.* **2013**, *172*, 478–499. [CrossRef]
18. Fineschi, S.; Taurchini, D.; Grossoni, P.; Petit, R.J.; Vendramin, G.G. Chloroplast DNA variation of white oaks in Italy. *For. Ecol. Manage.* **2002**, *156*, 103–114. [CrossRef]
19. Lumaret, R.; Jabbour-Zahab, R. Ancient and current gene flow between two distantly related Mediterranean oak species, Quercus suber and Q. ilex. *Ann. Bot.* **2009**, *104*, 725–736. [CrossRef]
20. McVay, J.D.; Hipp, A.L.; Manos, P.S. A genetic legacy of introgression confounds phylogeny and biogeography in oaks. *Proc. Biol. Sci.* **2017**, *284*. [CrossRef]
21. Eaton, D.A.R.; Hipp, A.L.; González-Rodríguez, A.; Cavender-Bares, J. Historical introgression among the American live oaks and the comparative nature of tests for introgression. *Evolution* **2015**, *69*, 2587–2601. [CrossRef]
22. de Groot, G.A.; During, H.J.; Maas, J.W.; Schneider, H.; Vogel, J.C.; Erkens, R.H. Use of rbcL and trnL-F as a two-locus DNA barcode for identification of NW-European ferns: An ecological perspective. *PLoS ONE* **2011**, *6*, e16371. [CrossRef]
23. Pang, X.; Liu, C.; Shi, L.; Liu, R.; Liang, D.; Li, H.; Cherny, S.S.; Chen, S. Utility of the *trnH-psbA* intergenic spacer region and its combinations as plant DNA barcodes: A meta-analysis. *PLoS ONE* **2012**, *7*, e48833. [CrossRef]
24. Saarela, J.M.; Sokoloff, P.C.; Gillespie, L.J.; Consaul, L.L.; Bull, R.D. DNA barcoding the Canadian Arctic flora: Core plastid barcodes (*rbcL* + *matK*) for 490 vascular plant species. *PLoS ONE* **2013**, *8*, e77982. [CrossRef]
25. Krawczyk, K.; Szczecińska, M.; Sawicki, J. Evaluation of 11 single-locus and seven multilocus DNA barcodes in *Lamium* L. (*Lamiaceae*). *Mol. Ecol. Resour.* **2014**, *14*, 272–285. [CrossRef]
26. Group, C.P.W. A DNA barcode for land plants. *Proc. Nat. Acad. Sci. USA* **2009**, *106*, 12794–12797. [CrossRef]
27. Dong, W.; Liu, J.; Yu, J.; Wang, L.; Zhou, S. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLoS ONE* **2012**, *7*, e35071. [CrossRef]
28. Dong, W.; Liu, H.; Xu, C.; Zuo, Y.; Chen, Z.; Zhou, S. A chloroplast genomic strategy for designing taxon specific DNA mini-barcodes: A case study on ginsengs. *BMC Genet.* **2014**, *15*, 138. [CrossRef]
29. Xu, C.; Dong, W.; Li, W.; Lu, Y.; Xie, X.; Jin, X.; Shi, J.; He, K.; Suo, Z. Comparative Analysis of Six *Lagerstroemia* Complete Chloroplast Genomes. *Front. Plant Sci.* **2017**, *8*, 15. [CrossRef]
30. Song, Y.; Wang, S.; Ding, Y.; Xu, J.; Li, M.F.; Zhu, S.; Chen, N. Chloroplast Genomic Resource of *Paris* for Species Discrimination. *Sci. Rep.* **2017**, *7*, 3427. [CrossRef]
31. Li, W.; Liu, Y.; Yang, Y.; Xie, X.; Lu, Y.; Yang, Z.; Jin, X.; Dong, W.; Suo, Z. Interspecific chloroplast genome sequence diversity and genomic resources in *Diospyros*. *BMC Plant Biol.* **2018**, *18*, 210. [CrossRef]
32. Zhao, Z.; Wang, X.; Yu, Y.; Yuan, S.; Jiang, D.; Zhang, Y.; Zhang, T.; Zhong, W.; Yuan, Q.; Huang, L. Complete chloroplast genome sequences of Dioscorea: Characterization, genomic resources, and phylogenetic analyses. *PeerJ* **2018**, *6*, e6032. [CrossRef] [PubMed]
33. Kim, K.J.; Jansen, R.K. NdhF sequence evolution and the major clades in the sunflower Family. *Proc. Nat. Acad. Sci. USA* **1995**, *92*, 10379–10383. [CrossRef] [PubMed]
34. Li, J.H. Phylogeny of *Catalpa* (Bignoniaceae) inferred from sequences of chloroplast *ndhF* and nuclear ribosomal DNA. *J. Syst. Evol.* **2008**, *46*, 341–348. [CrossRef]
35. Park, S.J.; Korompai, E.J.; Francisco-Ortega, J.; Santos-Guerra, A.; Jansen, R.K. Phylogenetic relationships of Tolpis (*Asteraceae*: *Lactuceae*) based on ndhF sequence data. *Plant Syst. Evol.* **2001**, *226*, 23–33. [CrossRef]
36. Ji, Y.; Liu, C.; Yang, Z.; Yang, L.; He, Z.; Wang, H.; Yang, J.; Yi, T. Testing and using complete plastomes and ribosomal DNA sequences as the next generation DNA barcodes in Panax (Araliaceae). *Mol. Ecol. Resour.* **2019**, *19*, 1333–1345. [CrossRef] [PubMed]
37. Kim, Y.; Choi, H.; Shin, J.; Jo, A.; Lee, K.-E.; Cho, S.-S.; Hwang, Y.-P.; Choi, C. Molecular Discrimination of Cynanchum wilfordii and Cynanchum auriculatum by InDel Markers of Chloroplast DNA. *Molecules* **2018**, *23*, 1337. [CrossRef]
38. Wang, A.; Wu, H.; Zhu, X.; Lin, J. Species Identification of Conyza bonariensis Assisted by Chloroplast Genome Sequencing. *Front. Genet.* **2018**, *9*. [CrossRef]

39. Kane, N.; Sveinsson, S.; Dempewolf, H.; Yang, J.Y.; Zhang, D.; Engels, J.M.; Cronk, Q. Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *Am. J. Bot.* **2012**. [CrossRef]
40. Hollingsworth, P.M.; Li, D.Z.; van der Bank, M.; Twyford, A.D. Telling plant species apart with DNA: From barcodes to genomes. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **2016**, *371*. [CrossRef]
41. Fu, C.N.; Wu, C.S.; Ye, L.J.; Mo, Z.Q.; Liu, J.; Chang, Y.W.; Li, D.Z.; Chaw, S.M.; Gao, L.M. Prevalence of isomeric plastomes and effectiveness of plastome super-barcodes in yews (Taxus) worldwide. *Sci. Rep.* **2019**, *9*, 2773. [CrossRef] [PubMed]
42. Li, J.; Wang, S.; Jing, Y.; Wang, L.; Zhou, S. A modified CTAB protocol for plant DNA extraction. *Chin. Bull. Bot.* **2013**, *48*, 72–78.
43. Patel, R.K.; Jain, M. NGS QC Toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE* **2012**, *7*, e30619. [CrossRef] [PubMed]
44. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D.; et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [CrossRef]
45. Dong, W.; Xu, C.; Cheng, T.; Lin, K.; Zhou, S. Sequencing angiosperm plastid genomes made easy: A complete set of universal primers and a case study on the phylogeny of Saxifragales. *Genome Biol. Evol.* **2013**, *5*, 989–997. [CrossRef]
46. Huang, D.I.; Cronk, Q.C.B. Plann: A command-line application for annotating plastome sequences. *Appl. Plant Sci.* **2015**, *3*, 1500026. [CrossRef]
47. Greiner, S.; Lehwark, P.; Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* **2019**, *47*, W59–W64. [CrossRef]
48. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef]
49. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef]
50. Ronquist, F.; Teslenko, M.; van der Mark, P.; Ayres, D.L.; Darling, A.; Hohna, S.; Larget, B.; Liu, L.; Suchard, M.A.; Huelsenbeck, J.P. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **2012**, *61*, 539–542. [CrossRef]
51. Tamura, K.; Stecher, G.; Peterson, D.; Filipski, A.; Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **2013**, *30*, 2725–2729. [CrossRef] [PubMed]
52. Brown, S.D.; Collins, R.A.; Boyer, S.; Lefort, M.C.; Malumbres-Olarte, J.; Vink, C.J.; Cruickshank, R.H. Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Mol. Ecol. Resour.* **2012**, *12*, 562–565. [CrossRef] [PubMed]
53. Swofford, D. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*; Version 4.0 Beta; Sinauer Associates: Sunderland, MA, USA, 2002.

International Journal of
Molecular Sciences

MDPI

Article

Genome-Wide Analysis of the YABBY Transcription Factor Family in Pineapple and Functional Identification of *AcYABBY4* Involvement in Salt Stress

Zeyun Li [1,†], Gang Li [2,†], Mingxing Cai [1], Samaranayaka V.G.N. Priyadarshani [1], Mohammad Aslam [1,3], Qiao Zhou [1], Xiaoyi Huang [1], Xiaomei Wang [4], Yeqiang Liu [3] and Yuan Qin [1,3,*]

[1] Key Laboratory of Genetics, Breeding and Multiple Utilization of Crops, Ministry of Education, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology; State Key Laboratory of Ecological Pest Control for Fujian and Taiwan Crops, College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China; lizeyun0514@163.com (Z.L.); caimingxing510704@163.com (M.C.); niro323@yahoo.com (S.V.G.N.P.); aslampmb1@gmail.com (M.A.); zhouqiao0606@163.com (Q.Z.); xiaoyi0922@163.com (X.H.)

[2] College of Agriculture, Fujian Agriculture and Forestry University, Fuzhou 350002, China; whu_ligang@whu.edu.cn

[3] State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, Guangxi Key Lab of Sugarcane Biology, College of Agriculture, Guangxi University, Nanning 530004, China; lyq91158@163.com

[4] Horticulture Research Institute, Guangxi Academy of Agricultural Sciences, Nanning Investigation Station of South Subtropical Fruit Trees, Ministry of Agriculture, Nanning 530007, China; wangxiaomei159@163.com

[*] Correspondence: yuanqin@fafu.edu.cn

[†] These authors contributed equally to this work.

Received: 29 September 2019; Accepted: 19 November 2019; Published: 22 November 2019

Abstract: The plant-specific transcription factor gene family, YABBY, belongs to the subfamily of zinc finger protein superfamily and plays an essential regulatory role in lateral organ development. In this study, nine *YABBY* genes were identified in the pineapple genome. Seven of them were located on seven different chromosomes and the remaining two were located on scaffold 1235. Through protein structure prediction and protein multiple sequence alignment, we found that *AcYABBY3*, *AcYABBY5 and AcYABBY7* lack a C2 structure in their N-terminal C2C2 zinc finger protein structure. Analysis of the *cis*-acting element indicated that all the seven pineapple *YABBY* genes contain multiple MYB and MYC elements. Further, the expression patterns analysis using the RNA-seq data of different pineapple tissues indicated that different *AcYABBYs* are preferentially expressed in various tissues. RT-qPCR showed that the expression of *AcYABBY2*, *AcYABBY3*, *AcYABBY6 and AcYABBY7* were highly sensitive to abiotic stresses. Subcellular localization in pineapple protoplasts, tobacco leaves and *Arabidopsis* roots showed that all the seven pineapple YABBY proteins were nucleus localized. Overexpression of *AcYABBY4* in *Arabidopsis* resulted in short root under NaCl treatment, indicating a negative regulatory role of *AcYABBY4* in plant resistance to salt stress. This study provides valuable information for the classification of pineapple *AcYABBY* genes and established a basis for further research on the functions of AcYABBY proteins in plant development and environmental stress response.

Keywords: YABBY; pineapple; expression pattern; subcellular localization; abiotic stress

1. Introduction

Plants are often exposed to extreme environments during their development and growth. Abiotic stresses such as salt, drought, high temperature and cold lead to adverse effect on growth and

Int. J. Mol. Sci. **2019**, *20*, 5863; doi:10.3390/ijms20235863 www.mdpi.com/journal/ijms

development of plants, resulting in yield loss. In plants, a series of defense systems play a vital role in survival under extreme external environmental changes. Transcription factors play a crucial role in plant defense system regulating gene expression, some of which are associated with the abiotic stress response [1].

A number of transcription factors are plant-specific. Structurally, transcription factors are divided into four functional regions, namely DNA binding domain, oligomeric site, transcriptional regulatory domain and nuclear localization signal region. These functional domains determine the characteristics, function, regulation and nuclear localization of transcription factors. The YABBY transcription factor family is widely present in plants and is a subfamily of the zinc finger protein superfamily. YABBY transcription factor possess two conserved domains, the N-terminal zinc finger domain and the C-terminal YABBY domain. The amino acid residues in these two domains are highly conserved and these domains are involved in the specific binding of DNA [2].

The evolutionary history of *YABBY* gene family is consistent with the origin of the leaves of seed plants. These transcription factors are specific to seed plants [3] and play important regulatory roles in the development of plants lateral organ. There are five YABBY subfamilies in angiosperms, namely *INO*, *CRC*, *YABBY2*, *FIL*/*YABBY3* and *YABBY5* [4]. The *YABBY* genes are well studied in *Arabidopsis thaliana*, where the six members of *YABBY* genes revealed overlapping functions [2,5]. All *AtYABBYs* promote the differentiation of the abaxial surface cells of lateral organs and participate in the establishment of dorsal-ventral polarity, leaf expansion and flower organ development [6,7]. Although they display similar functions, yet different *AtYABBY* genes have diversified expression pattern and function. For example, *AtYABBY2*, *AtYABBY3* and *AtFIL* genes are specifically expressed in the distal region of the aerial part of plants and *Arabidopsis* overexpression of *AtYABBY3* and *AtFIL* exhibited leaf-rolling [8,9]. *AtYABBY2*, *AtYABBY3*, *AtFIL* and *AtYABBY5* show functional redundancy during leaf development [10]. While *AtCRC* regulates the development of carpel and nectary [3,10]. *AtINO* mainly regulates the development of ovules [11]. In rice, *DL* and *OsYABBY1*, the homologs of *AtYAB2* and *AtCRC* are not expressed in a polar manner in the lateral organs and their functions are not associated with polarity regulation of lateral organ development [12,13]. *OsDL* mainly regulates the carpel identity and the formation of main veins of the leaves by promoting cell proliferation in the central region of rice leaves [14,15] whereas, *OsYABBY1* was found to be involved in the feedback regulation of gibberellin (GA) biosynthesis and metabolism resulting semi-dwarf phenotype in overexpression lines. [16,17]. Although *OsYABBY3* is involved in leaf development, unlike its function in *Arabidopsis*, it does not affect the establishment of leaf polarity [18]. Other than above developmental processes, the YABBY family *Shattering1* gene is responsible for seed shattering in cereals including sorghum, rice and maize [19].

Pineapple (*Ananas comosus* L.), a tropical edible fruit, is a perennial monocotyledon belonging to the Bromeliaceae. The completed assembly of the whole genome of pineapple [20] provides an opportunity for the systematic study of the pineapple YABBY family. Here, we identified 9 *YABBY* genes in pineapple and analyzed the gene structure, motif pattern of AcYABBYs, phylogenetic relationship of YABBYs between pineapple, *Arabidopsis* and rice. We found 7 *AcYABBY* genes located on seven different chromosomes. Through protein structure prediction and protein multiple sequence alignment analysis, we discovered that AcYABBY3, AcYABBY5 and AcYABBY7 lack a C2 at the N-terminus and thus could not constitute a C2C2 zinc finger domain. We further performed subcellular localization analysis for the seven YABBY proteins in pineapple protoplasts, tobacco leaves and *Arabidopsis* roots and found that the AcYABBY proteins were mainly localized in the nucleus. Moreover, we also analyzed the promoter *cis*-acting element and found that all the seven *AcYABBYs* have MYB and MYC domains, those are involved in drought, low temperature, salt and ABA stress responses. We further performed RT-qPCR analysis and showed that the expression levels of the seven *AcYABBYs* were changed under different abiotic stresses, including salt, drought, cold, hot, ABA and ethephon stresses at different time points. Finally, we found that over-expression of *AcYABBY4* in *Arabidopsis* resulted in increased susceptibility to salt stress. This study provides comprehensive information about pineapple AcYABBY proteins and a basis for studying the function of *AcYABBY* family members in plant development and response to environmental stresses.

2. Results

2.1. Identification and Characterization of the Pineapple YABBY Transcription Factors

To identify *AcYABBY* genes, BLAST and Hidden Markov Model searches were used to search the pineapple genome with *YABBY* sequences from *Arabidopsis* as query. A total of 9 *AcYABBY* genes were identified from pineapple genome and named *AcYABBY1* to *AcYABBY9*. The protein lengths of these genes ranged from 49 aa (*AcYABBY8*) to 226 aa (*AcYABBY2*) with the corresponding molecular weight ranging from 5383.98 to 24706.13 Da. The additional information about *AcYABBYs* transcript ID, gene name, proteins size, protein isoelectric point and exons are listed in Table 1. There are two genes, *AcYABBY8* and *AcYABBY9*, with incomplete N-terminal sequence. We found that the CDSs of *AcYABBY8* and *AcYABBY9* lack ATG, which codes the initiation codon, suggesting we did not obtain the full length genes. Unfortunately, the conserved YABBY domain was located at the N terminal of YABBY proteins according to the reported studies and our conversation assay. To obtain more exact results, we considered the remaining seven genes for further studies. According to the mapping results, seven *AcYABBY* genes were localized on seven different chromosomes and two genes were located on scaffold1235 (Figure 1).

Table 1. Protein information of pineapple *YABBYs*, including sequenced ID, chromosome locations, isoelectric point (pI), molecular weight (MW), protein length, CDS length and exon number.

Name	Gene Locus	Chromosome Location	PI	MW (Da)	Protein Length(aa)	CDS Length(bp)	Exon
AcYABBY1	*Aco007606*	8	9.45	20,042.95	182	549	5
AcYABBY2	*Aco016279*	3	6.88	24,706.13	226	681	7
AcYABBY3	*Aco005138*	7	9.25	19,962.60	178	537	6
AcYABBY4	*Aco008751*	9	8.64	20,161.95	178	537	6
AcYABBY5	*Aco028479*	scaffold1235	9.05	15,073.06	136	411	5
AcYABBY6	*Aco002202*	4	7.71	20,872.43	188	567	6
AcYABBY7	*Aco003917*	15	9.05	21,453.75	190	573	7
AcYABBY8	*Aco026269*	12	7.01	5383.98	49	150	2
AcYABBY9	*Aco028478*	scaffold1235	5.82	5988.82	53	162	2

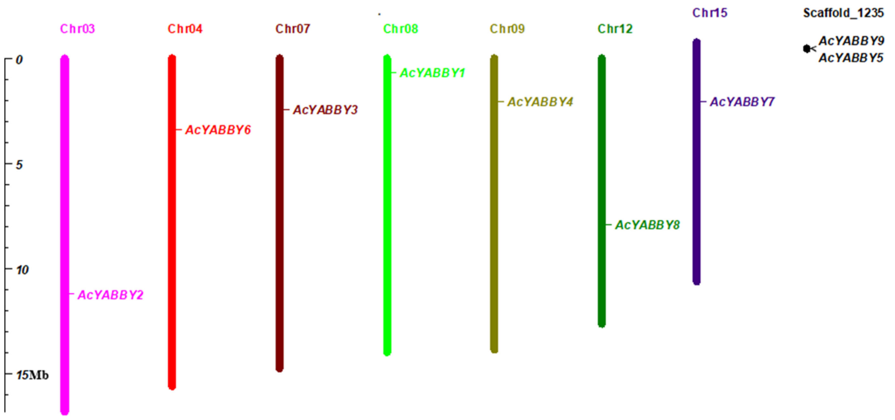


Figure 1. Distribution of YABBY genes in pineapple genome. Pineapple chromosomes and scaffolds having YABBY genes are shown here. The length of bar represents the size of the chromosome.

2.2. Phylogenetic Analysis of YABBY Family Genes

A phylogenetic tree was constructed to determine the phylogenetic relationships of *YABBY* genes in pineapple, *Arabidopsis* and rice. Phylogenetic analysis of 7 pineapple *YABBY* genes, 6 *Arabidopsis*

YABBY genes and 8 rice *YABBY* genes was performed by generating a neighbor-joining phylogenetic tree (Figure 2). The result showed that the *YABBY* genes of these three species could be divided into five subfamilies—YAB2, CRC, YAB5, FIL/YAB3 and INO. However, the *AcYABBYs* were divided to three subfamilies—FIL/YAB3, CRC and YAB2 (Figure 3a). Among these subfamilies, YAB2 had the largest members with four pineapple genes, one *Arabidopsis* gene and three rice genes. FIL/YAB3 was the second largest subgroup, containing two pineapple genes, two *Arabidopsis* genes and three rice genes. CRC contained three genes, including one pineapple gene (*AcYABBY7*), one *Arabidopsis thaliana* gene (*AtCRC*) and one rice gene (*OsDL*). However, INO contained one *Arabidopsis* gene (*AtYABBY4*) and one rice gene (*OsYABBY7*), yet no pineapple gene was categorized into this group. The smallest subgroup, YAB5, only had one *Arabidopsis* gene (*AtYABBY5*), suggesting that YAB5 may have a particular function in the *Arabidopsis thaliana*.

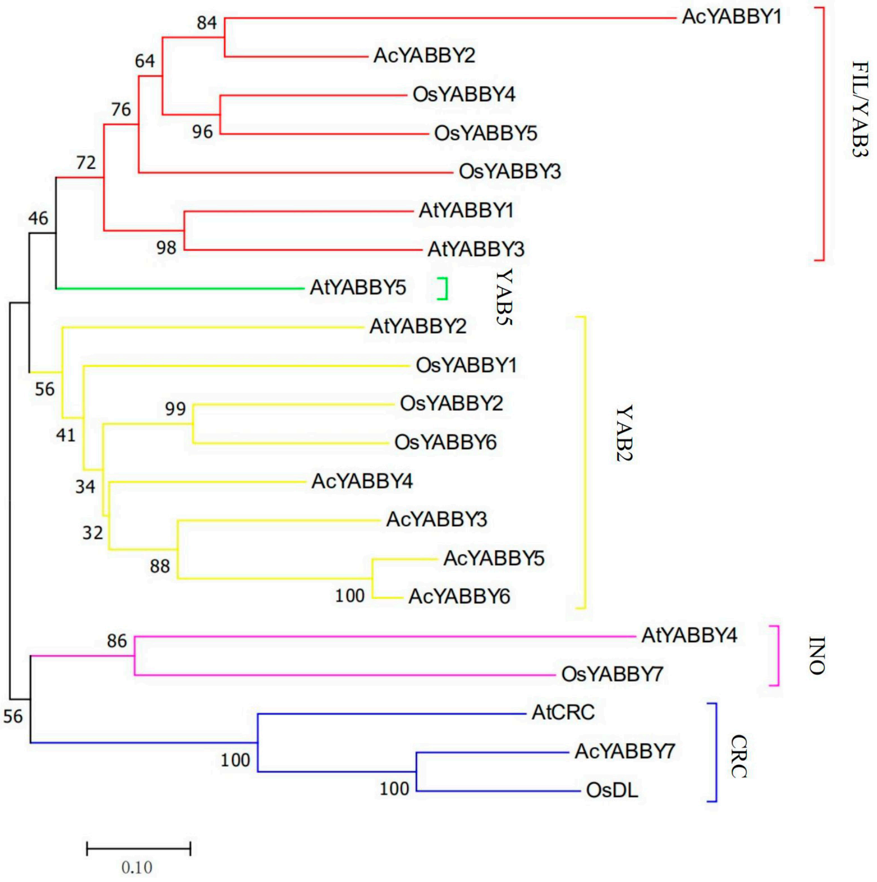


Figure 2. Phylogenetic relationship of the *YABBYs* proteins of Ac (Pineapple), At (*Arabidopsis thaliana*), Os (Oryza sativa). The phylogenetic tree was made using neighbor-joining with a bootstrap values of 1000. Phylogenetic tree divides *YABBYs* into five subfamilies (*YAB2*, *CRC*, *YAB5*, *FIL/YAB3 and INO*). Each subfamily is represented with different color.

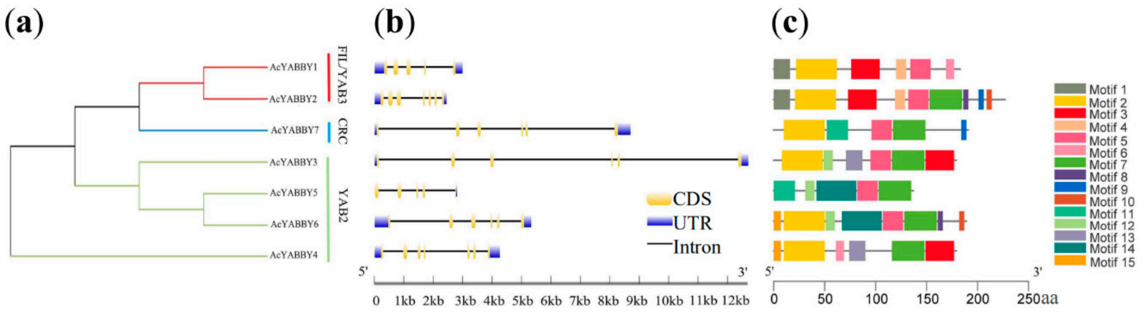


Figure 3. Phylogenetic relationships, gene structure and architecture of conserved protein motifs in pineapple YABBY genes. (**a**) The phylogenetic tree was constructed based on the full-length sequences of pineapple YABBY proteins using MEGA 7 software. Details of clusters are shown in different colors. (**b**) Exon-intron structure of pineapple YABBY genes. Blue boxes represent untranslated 5′- and 3′-regions; yellow boxes represent exons; black lines represent introns. (**c**) The motif composition of pineapple YABBY proteins. The motifs, numbers 1–15, are displayed with different colored boxes.

2.3. Gene Structure Analysis and Identification of Conserved Motifs

To investigate the structure diversity, we used Gene Structure Display Server [21]. The results showed that the exon number of the *AcYABBY* genes ranged from a minimum of 2 and a maximum of 7. *AcYABBY2* and *AcYABBY7* had the maximum exons numbers (Figure 3b). The schematic diagram representing the structure of AcYABBY proteins was constructed from the results of the MEME motif analysis (Figure 3c). The number of motifs ranged from 1 to 15 [22]. *AcYABBY5* and *AcYABBY7* contained five motifs. *AcYABBY1*, *AcYABBY3* and *AcYABBY4* had six motifs, while *AcYABBY6* had eight motifs. *AcYABBY2* had the largest number of motifs, containing nine motifs. The similar motif arrangements among AcYABBY proteins indicate that the protein architecture is conserved within a particular subfamily. The functions of these conserved motifs remain to be elucidated.

2.4. AcYABBY Protein Homology Modeling and Sequence Alignment

The YABBY family possesses a C2C2 zinc finger domain at the N-terminus and a YABBY domain at the C-terminus. To study AcYABBY protein conformation, seven pineapple YABBY protein were analyzed by SWISS-MODEL to identify the best template with known structures and similar sequence. All the seven AcYABBY proteins were predicted with a YABBY domain. Whereas, only four AcYABBY proteins contained a zinc-finger domain (Figure 4a). The zinc-finger domain was not detected in the three *AcYABBY* genes (*AcYABBY3*, *AcYABBY5*, *AcYABBY7*). To analyze further, we used DNAMAN for domain analysis of protein sequence alignment (Figure 4b). We found that one C2 structure was missing from the C2C2 structure in the protein sequences of the these three *AcYABBY* genes. Therefore, we concluded that zinc finger structure was incomplete and unpredictable in *AcYABBY3*, *AcYABBY5*, *AcYABBY7*. The specific cause of this phenomenon is still unclear and may be related to the evolution of the pineapple *YABBY* genes.

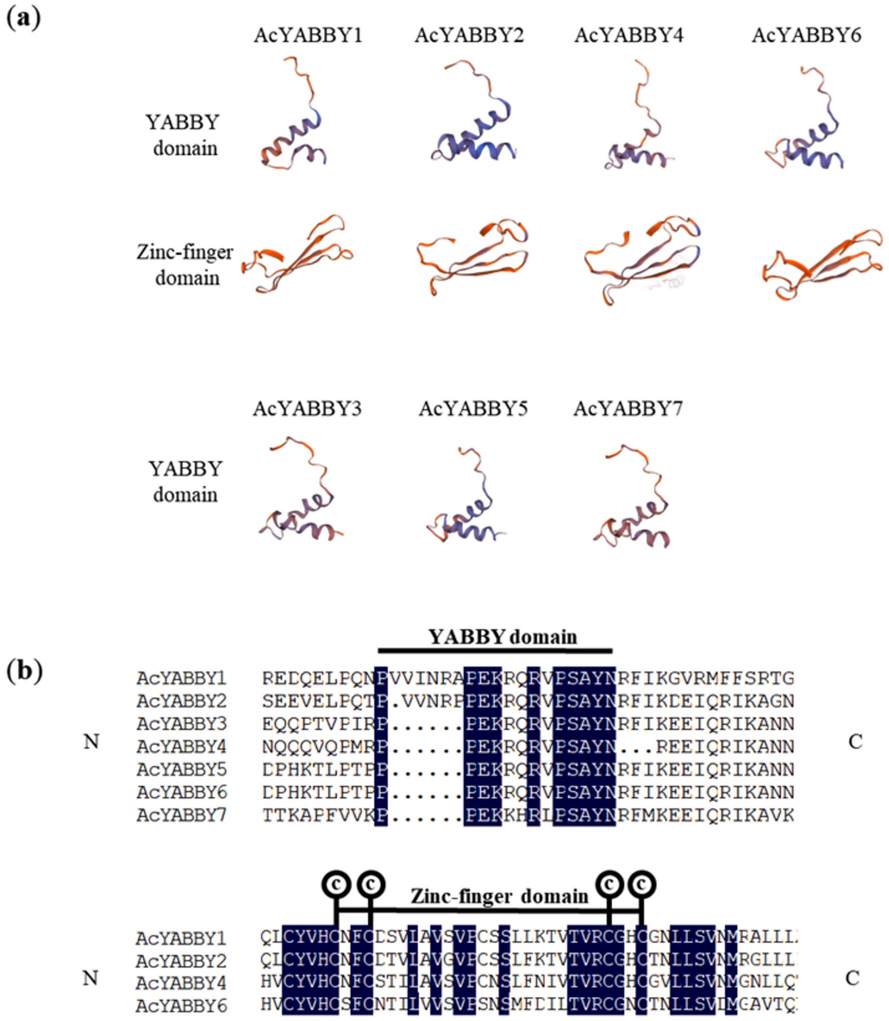

(a)
AcYABBY1
AcYABBY2
AcYABBY4
AcYABBY6
YABBY domain
Zinc-finger domain
AcYABBY3
AcYABBY5
AcYABBY7
YABBY domain
(b)
YABBY domain
N
C
AcYABBY1 REDQELPQNPVVINRAPEKRQRVPSAYNRFIKGVRMFFSRTG
AcYABBY2 SEEVELPQTP.VVNRPPEKRQRVPSAYNRFIKDEIQRIKAGN
AcYABBY3 EQQPTVPIRP......PEKRQRVPSAYNRFIKEEIQRIKANN
AcYABBY4 NQQQVQPMRP......PEKRQRVPSAYN...REEIQRIKANN
AcYABBY5 DPHKTLPTPP......PEKRQRVPSAYNRFIKEEIQRIKANN
AcYABBY6 DPHKTLPTPP......PEKRQRVPSAYNRFIKEEIQRIKANN
AcYABBY7 TTKAPFVVKP......PEKKHRLPSAYNRFMKEEIQRIKAVK
C C Zinc-finger domain C C
N
C
AcYABBY1 QLCYVHCNFCDSVLAVSVPCSSLLKTVTVRCGHCGNLLSVNMFALLL
AcYABBY2 QLCYVHCNFCDTVLAVGVPCSSLFKTVTVRCGHCTNLLSVNMRGLLL
AcYABBY4 HVCYVHCNFCSTILAVSVPCNSLFNIVTVRCGHCGVLLSVNMGNLLQ
AcYABBY6 HVCYVHCSFCNTILVVSVPSNSMFDILTVRCGNCTNLLSVIMGAVTQ

Figure 4. Predicated structures and multiple sequence alignment of pineapple YABBY proteins. (**a**) Predicted protein structures of AcYABBY proteins. (**b**) Multiple sequence alignment of AcYABBYs, zinc-finger domain and YABBY domain amino. 'N' and 'C' indicate the N-terminal and C-terminal. The four cysteine residues putatively responsible of the zinc-finger structure are also indicated. Identical amino acids are highlighted in blue.

2.5. Expression Profiling and Subcellular Localization of AcYABBYs

Transcriptome data from different developmental stages of pineapple tissue were used to study the expression patterns of seven *AcYABBY* genes to understand the transcriptional diversity of the *AcYABBY* genes. The expression levels of *AcYABBY* genes were analyzed by FPKM values from RNA sequence data. Hierarchical clusters and expression patterns of each gene were generated based on the average log values of each gene in each tissue. According to Heat-map analysis, the *AcYABBY3* expressed ubiquitously in most organs. The *AcYABBY3* was highly and specifically expressed at all stages of development of the organs (sepal Se1-4, gynoecium Gy1-7, petal Pe1-3, fruit S1-7, flower and leaf). Conversely, the *AcYABBY5* and *AcYABBY6* showed relatively low expression levels in almost all organs. Besides, some genes were highly expressed in specific organs. For example, *AcYABBY1* was highly expressed in sepal (Se2 stage) and fruit (S4-7 stages). *AcYABBY2* was highly expressed in sepal (Se1-4 stages), pistil (Gy1/2/4/5 stages) and petal (Pe1-2 stages). *AcYABBY4* was highly expressed in sepal (Se1-4 stages) and petal (Pe3 stage). *AcYABBY7* was highly expressed in gynoecium (Gy1-7 stages) (Figure 5a). The reliability of transcriptome data was further verified by RT-qPCR experiment, which was carried out on eight representative samples for seven *YABBY* genes. The results revealed that expression patterns of the *YABBY* genes detected by RT-qPCR were partially consistent with the results of RNA-seq analysis (Figure 5b). The differences between RT-qPCR and RNA-seq may be caused by sampling and it is difficult to keep the samples' stage exactly at the same as RNA-seq sequencing samples. To understand molecular characteristics of *AcYABBYs*, seven pineapple *YABBY* genes (*YABBY1-7*) were selected for subcellular localization. In pineapple protoplasts, *AcYABBYs-GFP* were co-localized with a DAPI signal, suggesting that the AcYABBY proteins were mainly localized in the nucleus (Figure 6). Also, the *AcYABBYs-GFP* fusion proteins were also localized in the nucleus of tobacco leaves (Figure S1) and *Arabidopsis* roots (Figure S2). These localization results were consistent with each other, suggesting that AcYABBY proteins were localized to the nucleus.

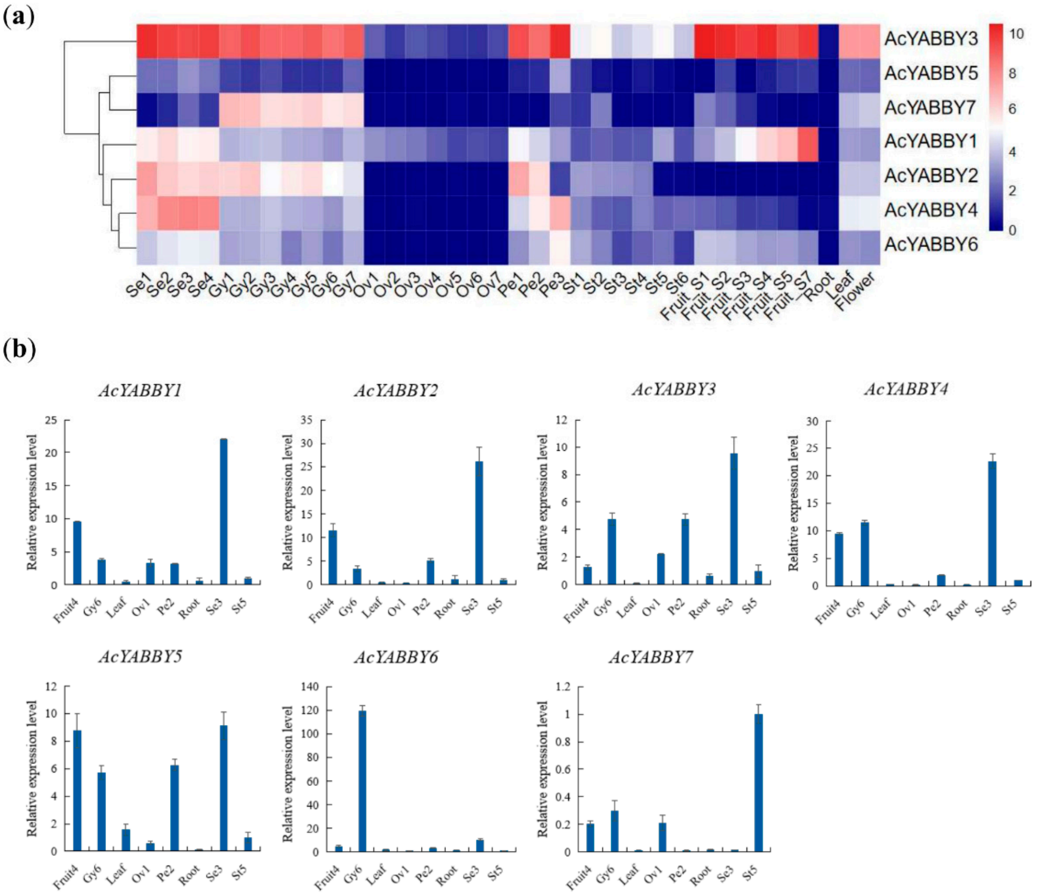


Figure 5. Expression profiles of the pineapple *YABBY* genes. (**a**) Heat-map of *AcYABBY* genes expression profiles in different tissue generated from RNA-seq data. The *AcYABBYs* were clustered according to their expression patterns. Red color indicates high levels of transcript abundance and blue indicates low transcript abundance. The color scale is shown on the right. Samples are mentioned at the bottom of each lane: Se (sepal) Se1-Se4, Gy (gynoecium) Gy1-7 Ov (ovule) Ov1-Ov7, Pe (petal) Pe1-Pe3, St (stamen) St1-St5, Fruit S1-S7, Root, Leaf, Flower. (**b**) Expression analysis of 7 pineapple *YABBY* genes in eight representative samples by RT-qPCR. RT-qPCR data were normalized using pineapple *PP2A* gene and vertical bars indicate standard deviation.

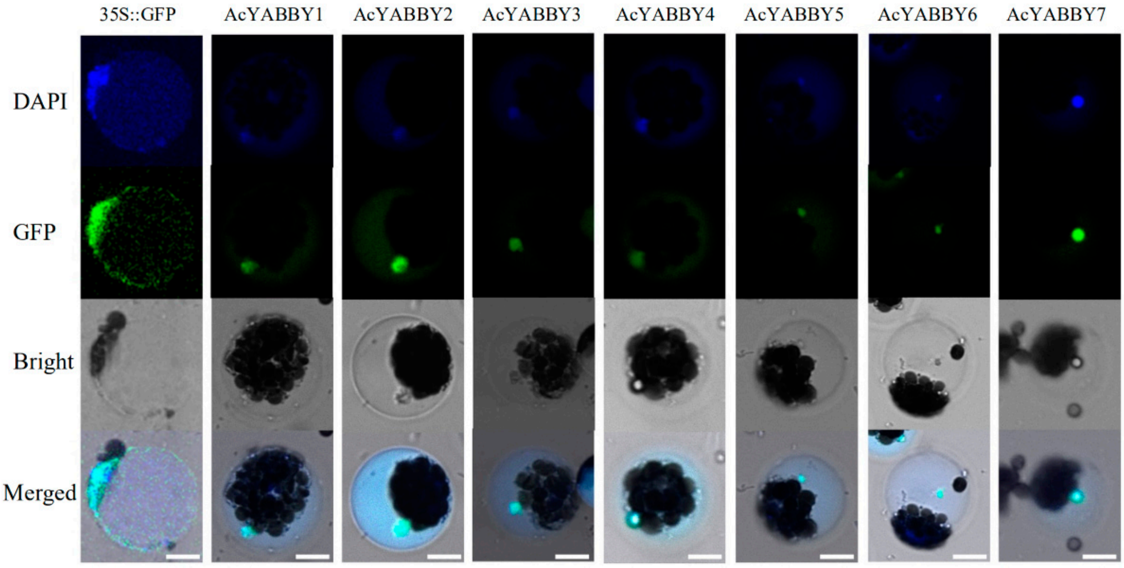


Figure 6. Subcellular localization of pineapple YABBY proteins in pineapple protoplasts. The *35S::AcYABBYs-GFP* and *35S::GFP* control vectors were transiently expressed in pineapple protoplasts. Results were visualized by a confocal microscope after 16 h transformation. Bar = 10 μm.

2.6. Cis-Acting Elements and RT-qPCR Analysis of AcYABBY Ggenes

The study of the *cis*-element indicated that each of the pineapple genes comprised the MYB and MYC elements in their promoter regions. MYB elements have been reported to be associated with drought, low temperature, salt and ABA stress responses and MYC elements are associated with drought and salt stress response [1]. Besides, the majority of the pineapple *YABBYs* except *AcYABBT3, AcYABBT4 and AcYABBT7* also contained at least an ERE element, which responds to ethylene. The identified Motif of the *cis*-acting elements and the sequences are listed in Table 2. To investigate the expression pattern of *AcYABBY* genes in response to different abiotic stresses, RT-qPCR was carried out using pineapple leaves exposed to various abiotic stresses for different time intervals such as 0, 6, 12, 24 and 48 h (h). The *AcYABBY* genes expressed diversely under the six types of abiotic stresses. Under salt stress (150 mM NaCl), the transcription level of *AcYABBYs* increased gradually. Among them, *AcYABBY1*, *AcYABBY2*, *AcYABBY4*, *AcYABBY5* and *AcYABBY7* transcript levels were highly up-regulated with the highest expression at 12 h. Notably, *AcYABBY2* increased to a peak (~250-fold) at 12 h and then rapidly declined to a level similar to the control (Figure 7a). Under drought stress, *AcYABBY2* increased to a peak of ~15-fold at 12 h. *AcYABBY4* increased to a maximum of ~10-fold at 12 h. *AcYABBY5* increased to a maximum of ~8-fold at 12 h. After 12 h, the expression level of *AcYABBY2*, *AcYABBY4 and AcYABBY5* rapidly declined to a level similar to the control when plants were subjected to drought stress (Figure 7b). Expression level of *AcYABBY*3 increased slowly to a peak of ~3-fold at 12 h and then declined to a level similar to the control under cold stress (Figure 7c). Under heat stress, *AcYABBY7* increased to a peak of ~30-fold at 12 h and then rapidly declined to a level similar to the control. *AcYABBY2* rose to a maximum of ~15-fold at 48 h (Figure 7d). Under ABA stress, *AcYABBY6* and *AcYABBY7* transcript levels were highly up-regulated at 48h. *AcYABBY6* increased to a peak of ~100-fold and *AcYABBY7* increased to a maximum of ~60-fold at 48h (Figure 7e). Under ethephon stress, *AcYABBY6* increased to a maximum of ~8-fold at 48h and *AcYABBY7* increased to a maximum of ~17-fold at 48h (Figure 7f). These results together suggested that *AcYABBYs* are involved in plant response to abiotic stresses.

Table 2. Distribution of MYB, MYC and ERE *cis*-acting element in pineapple YABBY promoters.

Gene	Motif Sequence	AcYABBY1	AcYABBY2	AcYABBY3	AcYABBY4	AcYABBY5	AcYABBY6	AcYABBY7
MYB	TAACCA CAACAG CAACCA	6	5	6	9	2	4	2
MYC	CATTTG CATGTG CAATTG	4	3	4	7	8	5	5
ERE	ATTTCATA ATTTTAAA	3	2	0	0	1	3	0

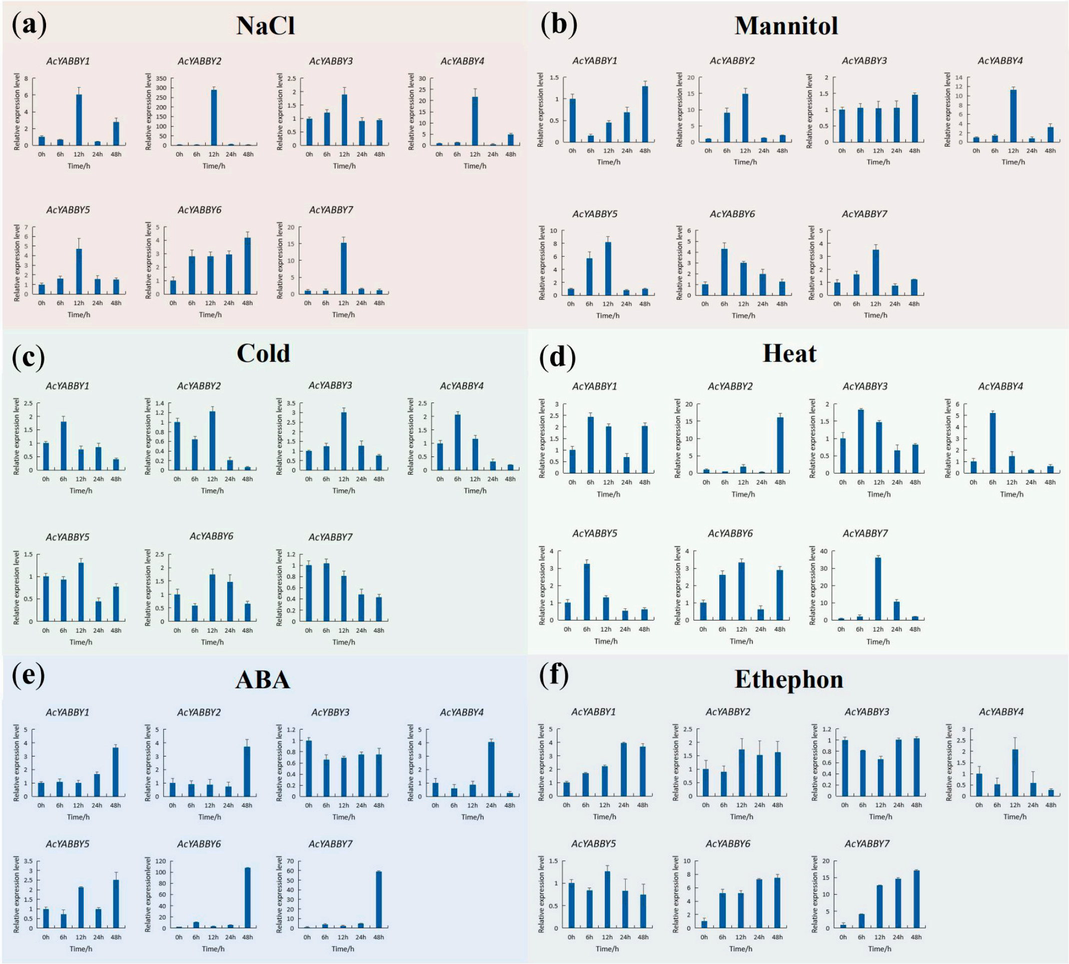


Figure 7. Expression profiles of 7 selected *AcYABBY* genes in response to various abiotic stress treatments. (**a**) NaCl treatment. (**b**) Mannitol treatment. (**c**) Cold treatment. (**d**) Heat treatment. (**e**) ABA treatment. (**f**) Ethephon treatment. RT-qPCR data were normalized using pineapple *PP2A* gene as reference gene. Error bars indicate Standard Deviation.

2.7. *AcYABBY4 Negatively Regulates the High Salinity Tolerance in Arabidopsis*

To further investigate the function of *AcYABBYs* upon abiotic stress, we overexpressed *AcYABBY4* driven by 35S promoter in *Arabidopsis* plants and compared the growth phenotype of *AcYABBY4-overexpression (AcYABBY4-OE)* lines with the wild-type plants under optimum and salinity condition. For the phenotype comparison under salt stress, seedlings were cultured vertically on 1/2 MS medium for three days and then transferred to 1/2 MS medium supplemented with or without NaCl and allowed to grow for additional seven days. The root length of all lines showed a similar phenotype when grown on 1/2 MS medium without NaCl (Figure 8a,b). Under the 100 mM

NaCl treatment, the root length of *AcYABBY4-OE* seedlings were also comparable to that of wild-type (Figure 8c). Whereas, under 150 mM NaCl treatment, the root length of *AcYABBY4-OE* seedlings were significantly reduced compared to wild-type seedlings (Figure 8d). Collectively, the results indicate that *AcYABBY4-OE* plants are susceptible to high salinity stress, suggesting that *AcYABBY4* may be a negative regulator in plant response to high salinity stress.

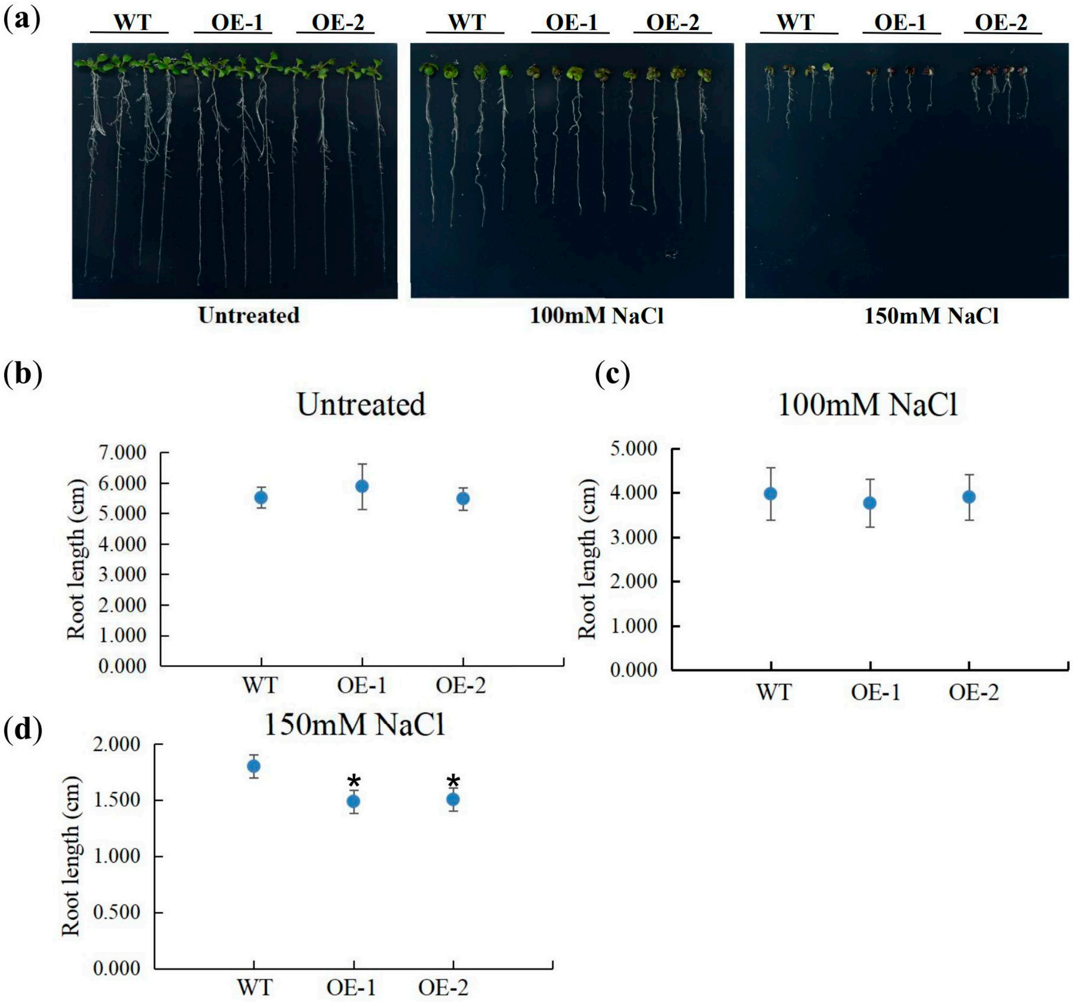


Figure 8. Phenotypic comparison of wild-type (WT) and *AcYABBY-over-exression* (*AcYABBY-OE*) *Arabidopsis* seedling under salt treatment. (**a**) Seedlings grown on 1/2 MS medium for three days were transferred to 1/2 MS medium supplemented with NaCl or without NaCl and allowed to grow for additional seven days. (**b**) Comparative root lengths of WT and transgenic lines in control conditions. (**c**) Comparative root length between WT and transgenic lines under 100mM NaCl treatment. (**d**) Comparative root lengths of WT and transgenic lines under 150 mM NaCl treatment. * denotes the significant difference (* $p < 0.05$) of t-tests between the transgenic lines and wild type.

3. Discussion

3.1. Diversity of YABBY Transcription Factors in Plants

The plant-specific YABBY transcription factor family is involved in early embryonic development and lateral organ development. In particular, it plays an important role in the establishment of the near-distal axis polarity of leaves and is also involved in biological processes, such as plant development and stress response [23]. The *YABBY* genes are well studied in dicotyledonous plants such as *Arabidopsis*. However, relatively less information is available about monocot *YABBY* genes though it is reported in rice. In *Arabidopsis*, there are six members in YABBY family, namely *AtCRC*, *AtYABBY1*, *AtFIL*, *AtYABBY3*, *AtINO/AtYABBY4*, *AtABBY5* [2,5], with their unique and overlapping functions [17,24]. Various evidences suggest that they promote the differentiation of the distal axis of lateral organ cells [6,7] and have important effect on leaf expansion and floral organ development. In rice, there are eight *YABBY* members, *OsDL*, *OsYABBY1*, *OsYABBY2*, *OsYABBY3*, *OsYABBY4*, *OsYABBY5*, *OsYABBY6* and *OsYABBY7* [10]. They play important roles in regulating the development of lateral organs such as leaves and flowers [25]. Pineapple (*Ananas comosus* L.) is the third most economically important tropical fruit crop in the world after bananas [26]. Scientists are trying to improve its quality by improving resistance to environmental changes, increase the yield and improve its taste [27]. Therefore, the functional study of *YABBY* genes in regulation of plant development and stress response is important for breeding programs and agricultural production [25]. However, there are no reports on the characterization of pineapple YABBY proteins until now.

Here, we identified nine *YABBY* genes in pineapple genome and named them from *AcYABBY1* to *AcYABBY9*. According to phylogenetic analysis, pineapple *YABBY* genes were more closely related to rice *YABBY* genes, because pineapple as a perennial monocot is evolutionarily more related to grasses, including corn, rice and wheat. The YABBY transcription factor family is a subfamily of the zinc finger protein superfamily, with a zinc finger domain at the N-terminus and a "helix-loop-helix" YABBY domain at the C-terminus similar to HMG-box, which have been confirmed to be associated with specific binding of DNA [2]. Some of the pineapple YABBY proteins are highly conserved with *Arabidopsis* and rice YABBY proteins. We found that some pineapple *YABBY* genes, including *AcYABBY3*, *AcYABBY5* and *AcYABBY7*, lack a N-terminal C2C2 zinc finger domain. It could be due to a technical issue in the process of genome assembly. To have more exact results, we only selected seven *YABBY* genes for further studies with *YABBY* gene features.

3.2. AcYABBY Gene Expression Profiles Analysis

The completion of the pineapple genome sequence has provided us an opportunity to explore the specific genes responsible for specific traits [20]. As shown in Figure 5, hierarchical clusters and expression patterns shows that the *AcYABBY* genes have distinct expression pattern. For example, *AcYABBY7* is highly expressed in pistil. It was reported that *OsDL* is involved in the regulation of floral meristem [13]. Phylogenetic analysis indicate that *AcYABBY7* and *OsDL* are in the same subcluster, implying that *AcYABBY7* may also be involved in the tissue development of pineapple pistil. Interestingly, some *AcYABBYs* are preferentially expressed in floral organs. For example, *AcYABBY3* and *AcYABBY2* are enriched in sepals, pistils and petals. The specific enrichments of *AcYABBY4* in sepals and petals suggests its probable role in floral development of pineapple. *OsYABBY1* was found to be involved in the feedback regulation of gibberellin (GA) metabolism in rice [16]. Phylogenetic analysis shows that *AcYABBY3* and *AcYABBY4* and *OsYABBY1* are clustered together. It will be interesting to test whether *AcYABBY3* and *AcYABBY4* are also involved in the feedback regulation of gibberellin metabolism. Besides, we found that *AcYABBY3* and *AcYABBY1* are highly expressed in the fruit. It will be worth to investigate their potential role in the pineapple fruit development. The expression pattern of *AcYABBYs* and its corresponding function of rice homologous genes provide us a clue for understanding the function of *AcYABBYs* in pineapples.

3.3. AcYABBY4 Inhibits Root Growth of Seedlings under Salt Stress

Pineapple plant growth is affected by many factors during its growth and development, such as the uneven distribution of hormones, drought, salt and other adverse environmental conditions. In soybean, *YABBY* genes are involved in abiotic stress response and *GmYABBY10* act as a negative regulator of salt and drought stress in *Arabidopsis* [1]. However, the function of *AcYABBY* in plant response to abiotic stresses remains unknown. We analyzed the *cis*-acting elements of the pineapple YABBY family and found that the *AcYABBY* genes also comprise *cis*-acting elements such as MYB, MYC and ERE in their promoters. These *cis*-acting elements are involved in abiotic stress response in plants such as MYB elements are involved in salt, drought, low temperature and response to

ABA [28,29]. MYC elements are involved in salt, drought and ABA stress response and ERE is related to ethylene response. Plant abiotic stress-responsive transcription factors could bind to these *cis*-acting elements [1]. The presence of these *cis*-elements in the promoter of *AcYABBYs* indicate that *AcYABBYs* are associated with stress response and may also be involved in plant adaptation to the environmental changes. Studying the function of the *YABBY* genes in plant response to abiotic stresses may help us to improve the agriculture production of pineapple. Using RT-qPCR we analyzed the response of the AcYABBY gene family under abiotic stress conditions such as NaCl, drought, cold, heat, ABA and ethephon treatments. The results showed that the expression patterns of pineapple *YABBY* genes were different under six stress conditions. For example, the expression of *AcYABBY2*, *AcYABBY3*, *AcYABBY4*, *AcYABBY6* and *AcYABBY7* was induced by NaCl, cold, drought, ABA and heat stress, respectively (Figure 7). These results suggest that *AcYABBYs* may play an important role in response to abiotic stress. The response of plants to abiotic stress is a complex process that is regulated by different molecular and cellular pathways. Here, the response of the *YABBY* genes to six different stresses laid the foundation for further functional study of the pineapple *YABBY* genes. To investigate the *AcYABBYs'* functions, we constructed all the *YABBYs'* over expression vectors and transformed it into *Arabidopsis*. We noted that the *AcYABBY4* over expression lines showed obvious phenotype with vegetative growth and development. Considering that the *AcYABBY4* basically had an early response than *AcYABBY2* during different stress conditions and the *AcYABBY4* indeed have a significant response to NaCl stress, so we chose *AcYABBY4* as our target gene for the functional analysis. However, *Arabidopsis* overexpressing *AcYABBY4* showed delayed growth and small seedlings size than wild-type (WT). We also found that the root growth of *AcYABBY4* overexpression lines were like WT and shows a normal phenotype in early stages of growth. Therefore, to avoid the effect of developmental changes on interpretation of our results of salt stress, we selected root growth as the parameter to study the salt stress. We found that the *Arabidopsis* overexpressing *AcYABBY4* is more sensitive to salt stress, indicating that *AcYABBY4* facilitates plants to perceive changes in the external salt environment more quickly and then inhibits plant growth through other complex regulatory mechanisms. Therefore, *AcYABBY4* may be a negative regulator of salt tolerance in transgenic *Arabidopsis* (Figure 8), providing a reliable basis for understanding *YABBY* participation in the biological process of plant stress response.

There is a close relationship between hormonal signal transduction and stress response. The response elements related to ABA and ethylene have important functions in plant abiotic stress and disease resistance. These observations also suggests that the *YABBY* genes may play an important role in response to ABA and ethylene accumulation of pineapple plants. The regulation mechanism of AcYABBY proteins under biological stress and in hormone signal transduction still remains elusive and could be a key part of further study of AcYABBYs.

4. Materials and Methods

4.1. Identification of YABBY Transcription Factors in Pineapple

The pineapple YABBY protein and genome sequences were downloaded from Phytozome 12 (https://phytozome.jgi.doe.gov/pz/portal.html) and PlantTFDB (http://planttfdb.cbi.pku.edu.cn/) (Table S4). To identify the pineapple *YABBY* genes, HMMER3.02 (http://hmmer.wustl.edu/) with default parameters settings were used to search for the PFAM YABBY domain (PF04690) (http://pfam.sanger.ac.uk/). Also, BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgiwas) was used to search for potential pineapple genes containing the YABBY transcription factor. To achieve the accuracy of the analysis, we further analyzed the conserved sequence using NCBI-CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) with an E-value threshold of 0.013 and abandoned any sequences lacking the YABBY annotation [30]. The isoelectric point (pI) and molecular weight (MW) of the AcYABBY proteins were predicted using ExPASy-Compute pI/MW (https://web.expasy.org/compute_pi/).

4.2. Chromosome Locations of Pineapple YABBY Genes

The information about the chromosome location of the *AcYABBY* genes were obtained from Phytozome. Based on the starting position of the gene and the length of the relevant chromosome, Mapchart [31] was used to visualize the localization of the pineapple *YABBY* gene on chromosomes and scaffolds.

4.3. Construction of Phylogenetic Tree

To generate phylogenetic trees, amino acid sequences of YABBYs in *Arabidopsis*, rice and pineapple were compared and analyzed by MEGA7 [32]. The protein sequences were aligned by MUSCLE and phylogenetic tree was constructed using Neighbor-Joining method with a bootstrap value of 1000.

4.4. Gene Structure Analysis and Identification of Conserved Motifs

The exon-intron substructure map of the pineapple *YABBY* genes were analyzed using online tool GSDS (http://gsds.cbi.pku.edu.cn/). The motifs of the AcYABBY proteins were determined using the MEME Suite 5.0.5 with classic mode, zoops for selecting the site distribution and 15 motifs. (http://meme-suite.org/).

4.5. Modeling of Protein 3D Structures and Sequence Alignment of AcYABBY Proteins

SWISS-MODEL (https://www.swissmodel.expasy.org) was used to predict the structures of seven YABBY proteins. The templates used in the study are listed in Table S1. The alignment of the AcYABBY protein sequences were carried out using DNAMAN [33].

4.6. RNA-Seq and RT-qPCR Data Analysis

To understand the expression pattern of *YABBY* genes in different development stages of pineapple, sepal, pistil, ovule, petal, stamen, fruit, root, leaf and flower RNA-Seq data (SRA315090) was downloaded from the NCBI database [20]. The trimmed pair-end reads of all tissues were aligned with the pineapple genome using TopHat v2.1.1 [34] with default parameter settings. The FPKM values were derived from Cuffdiff v2.2.1 and then a heatmap of *YABBY* genes expression was generated by the Rmap software package. To determine the relative transcript level of the selected *AcYABBY* genes, total RNA was extracted from different tissues (sepal, pistil, ovule, petal, stamen, fruit, root, leaf) of pineapple according to the procedure in the description of the RNA Plant Extraction Kit (OMEGA, Guangzhou, China). RNA quality was tested using gel electrophoresis and NanoDrop2000c (Thermo Fisher Scientific, Fujian, China). One µg of purified total RNA was reverse transcribed into cDNA in a 20 µL reaction volume using AMV reverse transcriptase (Takara, Beijing, China) according to the supplier's instructions. Quantitative primers were designed by IDT (http://www.idtdna.com/pages/products/gene-expression/primetime-qpcr-assays-and-primers), with G+C content between 45%–55%, primer length between 17–25 bases, Tm value between 58–62 and quantitative product size between 80–150 bp. The reaction volume of RT-qPCR was 20 µL (10 µL 2×TransStart Top Green qPCR SuperMix, 8 µL nuclease-free water, 0.5 µL forward primer, 0.5 µL reverse primer and 1ul cDNA) according to the supplier's instructions (TransGen Biotech, Beijing, China). The RT-qPCR parameters were as follows: 95 °C for 30 s; 95 °C for 5 s and 60 °C for 40 s for 40 cycles; 95 °C for 15 s. In each case, three technical replicates and at least three independent biological replicates were used. The primers used in this study are listed in Table S2. For RT-qPCR data analysis, the quantification cycle (Cq)value was automatically calculated by the Bio-Rad CFX Manager 3.1 system software and the delta-delta Cq method was used to calculate the relative expression levels.

4.7. Vector Construction and Subcellular Localization

The full-length of coding sequences of the *AcYABBY* genes were amplified using the primers listed in Table S3. The PCR fragments were cloned into the pENTR/D-TOPO vectors (Invitrogen) and sequenced separately. The positive clones were recombined into the destination vector pGWB505 using LR reaction. The vectors harboring the *AcYABBY* genes were transformed into *Agrobacterium tumefaciens* (GV3101) and infiltrated to tobacco leaves. Subcellular localizations in tobacco leaves were observed with confocal microscope using GFP and DAPI stain. The *A. tumefaciens* (GV3101) with *AcYABBY* genes were also used to transform wild-type *Arabidopsis* by a floral dip procedure [35]. Transgenic *Arabidopsis* lines were selected on 1/2 MS medium supplemented with 50 mg/L hygromycin. All the experiments were carried out in the T_2 generation. The roots of the positive transgenic plants were stained by PI solution and subcellular localization was observed with confocal microscope. All *AcYABBY* genes were also used to study the protein localization in the protoplast of pineapple [36].

4.8. Cis-Acting Elements in the Pineapple YABBY Genes Promoter region

For each pineapple *YABBY* gene, 2000 bp upstream of the translational start codon was selected from Phytozome 12 as promoter region and analyzed for presence of *cis*-acting elements using PlantCARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/). The main *cis*-acting elements are listed in Table 2.

4.9. Stress Treatment

One month old tissue culture raised seedlings of pineapple (MD2 variety) were maintained under 3000 Lux light intensity and a day/night cycle of 16/8 h at 25 ± 2 °C in a controlled environment [36]. For different treatments, the seedlings were transferred to solutions containing 150 mM NaCl for salt stress, 350 mM Mannitol for osmotic stress, 100 μM ABA, 100 μM Ethephon. Cold and heat stress were performed by placing seedlings in the 4 °C and 45 °C chambers. Samples were collected at 0, 6, 12, 24 and 48h [37] after each treatment and immediately stored in liquid nitrogen.

4.10. Tolerance Assays under Stress Conditions

To assess phenotype under salt stress, the seeds of wild type and transgenic *Arabidopsis* lines were surface sterilized and seeded on 1/2 MS plates, then kept at 4 °C for 48 h in the dark before germination. About 50 seeds of each transgenic line were seeded on 1/2 MS medium for three days were transferred to 1/2 MS medium supplemented with NaCl or without NaCl and kept at 22 ± 2 °C under 16 h light/8 h dark for seven days and root length was measured.

5. Conclusions

In this study, nine pineapple *YABBY* genes were identified those are preferentially expressed in different tissues. Transient expression analysis in pineapple protoplasts, tobacco leaves and *Arabidopsis* roots showed that pineapple YABBY protein were localized in the nucleus. RT-qPCR results demonstrate that *AcYABBY2*, *AcYABBY4 and AcYABBY7* are regulated by drought, NaCl, cold and heat stress. Functional analysis of *AcYABBY4* suggests that *AcYABBY4* is a negative regulator of salt stress.

Supplementary Materials: The following are available online at http://www.mdpi.com/1422-0067/20/23/5863/s1.

Author Contributions: Z.L. and G.L. designed the experiment and wrote the manuscript. M.C. helped in photography and figure arrangement. S.V.G.N.P. provided the pineapple materials and helped in tissue culture. M.A. helped with article modification. Q.Z. helped in protoplast isolation. X.H. helped in plant culturing. X.W. and Y.L. provided reagents. Y.Q. conceived the study and revised the manuscript. All authors read and approved the final manuscript.

Funding: This research was funded by NSFC (U1605212, 31761130074, 31970333 to Y.Q.), a Guangxi Distinguished Experts Fellowship and a Newton Advanced Fellowship (NA160391) to Y.Q., Science and Technology Major Project of Guangxi–Research and application of ecological and high efficient cultivation techniques for dominant and characteristic fruits (AA17204097-6) and Project of Guangxi featured fruit innovation team on pineapple breeding and cultivation post under national modern agricultural industry technology system (nycytxgxcxtd-17-05).

Acknowledgments: We would like to thank the reviewers for their helpful comments on the original manuscript.

Conflicts of Interest: The authors declare that they have no competing interests.

Abbreviations

ABA	Abscisic acid;
GSDS	Gene Structure Display Server
HMM	Hidden Markov Models
MEGE	Molecular Evolutionary Genetics Analysis
MEME	Multiple Em for Motif Elicitation
RT-qPCR	Real-time Quantitative PCR
RNA-Seq	RNA sequencing
SMART	Simple Modular Architecture Research Tool
MS	Murashige and Skoog
PEG	Polyethylene glycol
WT	Wild type
GFP	Green fluorescent protein
AtYABBY	*Arabidopsis thaliana* YABBY
OsYABBY	rice (*Oryza sativa*) YABBY
AcYABBY	pineapple (*Ananas comosus*) YABBY
CDS	Coding sequence
ATG	Starting codon
FPKM	Fragments Per Kilobase Million
DAPI	4′,6-Diamidino-2-Phenylindole
MS	Murashige and Skoog
HMG	High mobility group
MYB	V-myb Myeloblastosis Viral Oncogene Homolog
MYC	Myelocytomatosis oncogenes
ERE	Ethylene Responsive Element

References

1. Zhao, S.P.; Lu, D.; Yu, T.F.; Ji, Y.J.; Zheng, W.J.; Zhang, S.X.; Chai, S.C.; Chen, Z.Y.; Cui, X.Y. Genome-wide analysis of the YABBY family in soybean and functional identification of GmYABBY10 involvement in high salt and drought stresses. *Plant Physiol. Biochem.* **2017**, *119*, 132–146. [CrossRef] [PubMed]
2. Sawa, S.; Watanabe, K.; Goto, K.; Liu, Y.G.; Shibata, D.; Kanaya, E.; Morita, E.H.; Okada, K. FILAMENTOUS FLOWER, a meristem and organ identity gene of Arabidopsis, encodes a protein with a zinc finger and HMG-related domains. *Genes Dev.* **1999**, *13*, 1079–1088. [CrossRef] [PubMed]
3. Sarojam, R.; Sappl, P.G.; Goldshmidt, A.; Efroni, I.; Floyd, S.K.; Eshed, Y.; Bowman, J.L. Differentiating Arabidopsis shoots from leaves by combined YABBY activities. *Plant Cell* **2010**, *22*, 2113–2130. [CrossRef]
4. Yamada, T.; Ito, M.; Kato, M. Expression pattern of INNER NO OUTER homologue in Nymphaea (water lily family, Nymphaeaceae). *Dev. Genes Evol.* **2003**, *213*, 510–513. [CrossRef] [PubMed]
5. Siegfried, K.R.; Eshed, Y.; Baum, S.F.; Otsuga, D.; Drews, G.N.; Bowman, J.L. Members of the YABBY gene family specify abaxial cell fate in Arabidopsis. *Development* **1999**, *126*, 4117–4128. [PubMed]
6. Balasubramanian, S.; Schneitz, K. NOZZLE links proximal-distal and adaxial-abaxial pattern formation during ovule development in Arabidopsis thaliana. *Development* **2002**, *129*, 4291–4300. [PubMed]
7. Eshed, Y.; Baum, S.F.; Bowman, J.L. Distinct mechanisms promote polarity establishment in carpels of Arabidopsis. *Cell* **1999**, *99*, 199–209. [CrossRef]
8. Kerstetter, R.A.; Bollman, K.; Taylor, R.A.; Bomblies, K.; Poethig, R.S. KANADI regulates organ polarity in Arabidopsis. *Nature* **2001**, *411*, 706–709. [CrossRef]
9. Yamada, T.; Yokota, S.; Hirayama, Y.; Imaichi, R.; Kato, M.; Gasser, C.S. Ancestral expression patterns and evolutionary diversification of YABBY genes in angiosperms. *Plant J.* **2011**, *67*, 26–36. [CrossRef]
10. Eckardt, N.A. YABBY genes and the development and origin of seed plant leaves. *Plant Cell* **2010**, *22*, 2103. [CrossRef]

11. Stahle, M.I.; Kuehlich, J.; Staron, L.; von Arnim, A.G.; Golz, J.F. YABBYs and the transcriptional corepressors LEUNIG and LEUNIG_HOMOLOG maintain leaf polarity and meristem activity in Arabidopsis. *Plant Cell* **2009**, *21*, 3105–3118. [CrossRef] [PubMed]
12. Jang, S.; Hur, J.; Kim, S.J.; Han, M.J.; Kim, S.R.; An, G. Ectopic expression of OsYAB1 causes extra stamens and carpels in rice. *Plant Mol. Biol.* **2004**, *56*, 133–143. [CrossRef] [PubMed]
13. Yamaguchi, T.; Nagasawa, N.; Kawasaki, S.; Matsuoka, M.; Nagato, Y.; Hirano, H.Y. The YABBY gene DROOPING LEAF regulates carpel specification and midrib development in Oryza sativa. *Plant Cell* **2004**, *16*, 500–509. [CrossRef] [PubMed]
14. Ha, C.M.; Jun, J.H.; Fletcher, J.C. Control of Arabidopsis leaf morphogenesis through regulation of the YABBY and KNOX families of transcription factors. *Genetics* **2010**, *186*, 197–206. [CrossRef]
15. Zhang, X.L.; Yang, Z.P.; Zhang, J.; Zhang, L.G. Ectopic expression of BraYAB1–702, a member of YABBY gene family in Chinese cabbage, causes leaf curling, inhibition of development of shoot apical meristem and flowering stage delaying in Arabidopsis thaliana. *Int. J. Mol. Sci.* **2013**, *14*, 14872–14891. [CrossRef]
16. Toriba, T.; Harada, K.; Takamura, A.; Nakamura, H.; Ichikawa, H.; Suzaki, T.; Hirano, H.Y. Molecular characterization the YABBY gene family in Oryza sativa and expression analysis of OsYABBY1. *Mol. Genet Genom.* **2007**, *277*, 457–468. [CrossRef]
17. Villanueva, J.M.; Broadhvest, J.; Hauser, B.A.; Meister, R.J.; Schneitz, K.; Gasser, C.S. INNER NO OUTER regulates abaxial- adaxial patterning in Arabidopsis ovules. *Genes Dev.* **1999**, *13*, 3160–3169. [CrossRef]
18. Ohmori, Y.; Toriba, T.; Nakamura, H.; Ichikawa, H.; Hirano, H.Y. Temporal and spatial regulation of DROOPING LEAF gene expression that promotes midrib formation in rice. *Plant J.* **2011**, *65*, 77–86. [CrossRef]
19. Lin, Z.; Li, X.; Shannon, L.M.; Yeh, C.T.; Wang, M.L.; Bai, G.; Peng, Z.; Li, J.; Trick, H.N.; Clemente, T.E.; et al. Parallel domestication of the Shattering1 genes in cereals. *Nat. Genet* **2012**, *44*, 720–724. [CrossRef]
20. Ming, R.; VanBuren, R.; Wai, C.M.; Tang, H.; Schatz, M.C.; Bowers, J.E.; Lyons, E.; Wang, M.L.; Chen, J.; Biggers, E.; et al. The pineapple genome and the evolution of CAM photosynthesis. *Nat. Genet* **2015**, *47*, 1435–1442. [CrossRef]
21. Guo, A.Y.; Zhu, Q.H.; Chen, X.; Luo, J.C. GSDS: A gene structure display server. *Yi Chuan* **2007**, *29*, 1023–1026. [CrossRef] [PubMed]
22. Bailey, T.L.; Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1994**, *2*, 28–36. [PubMed]
23. Bowman, J.L. The YABBY gene family and abaxial cell fate. *Curr Opin Plant Biol* **2000**, *3*, 17–22. [CrossRef]
24. Finet, C.; Floyd, S.K.; Conway, S.J.; Zhong, B.; Scutt, C.P.; Bowman, J.L. Evolution of the YABBY gene family in seed plants. *Evol. Dev.* **2016**, *18*, 116–126. [CrossRef]
25. Dai, M.; Hu, Y.; Zhao, Y.; Liu, H.; Zhou, D.X. A WUSCHEL-LIKE HOMEOBOX gene represses a YABBY gene expression required for rice leaf development. *Plant Physiol.* **2007**, *144*, 380–390. [CrossRef]
26. Zhang, J.; Liu, J.; Ming, R. Genomic analyses of the CAM plant pineapple. *J. Exp. Bot.* **2014**, *65*, 3395–3404. [CrossRef]
27. Su, Z.X.; Wang, L.L.; Li, W.M.; Zhao, L.H.; Huang, X.Y.; Azam, S.M.; Qin, Y. Genome-Wide Identification of Auxin Response Factor (ARF) Genes Family and its Tissue-Specific Prominent Expression in Pineapple (*Ananas comosus*). *Trop Plant Biol.* **2017**, *10*, 86–96. [CrossRef]
28. He, Y.A.; Li, W.; Lv, J.; Jia, Y.B.; Wang, M.C.; Xia, G.M. Ectopic expression of a wheat MYB transcription factor gene, TaMYB73, improves salinity stress tolerance in Arabidopsis thaliana. *J. Exp. Bot.* **2012**, *63*, 1511–1522. [CrossRef]
29. Yang, A.; Dai, X.Y.; Zhang, W.H. A R2R3-type MYB gene, OsMYB2, is involved in salt, cold and dehydration tolerance in rice. *J. Exp. Bot.* **2012**, *63*, 2541–2556. [CrossRef]
30. Zhang, C.; Wang, D.; Yang, C.; Kong, N.; Shi, Z.; Zhao, P.; Nan, Y.; Nie, T.; Wang, R.; Ma, H.; et al. Genome-wide identification of the potato WRKY transcription factor family. *PLoS ONE* **2017**, *12*, e0181573. [CrossRef]
31. Li, W.; Yan, M.; Hu, B.; Priyadarshani, S.V.G.N.; Hou, Z.; Ojolo, S.P.; Xiong, J.; Zhao, H.; Qin, Y. Characterization and the Expression Analysis of Nitrate Transporter (NRT) Gene Family in Pineapple. *Trop. Plant Biol.* **2018**, *11*, 177–191. [CrossRef]

32. Zhang, M.; Liu, Y.; Shi, H.; Guo, M.; Chai, M.; He, Q.; Yan, M.; Cao, D.; Zhao, L.; Cai, H.; et al. Evolutionary and expression analyses of soybean basic Leucine zipper transcription factor family. *BMC Genomics* **2018**, *19*, 159. [CrossRef] [PubMed]
33. Wang, Y.; Hua, X.; Xu, J.; Chen, Z.; Fan, T.; Zeng, Z.; Wang, H.; Hour, A.L.; Yu, Q.; Ming, R.; et al. Comparative genomics revealed the gene evolution and functional divergence of magnesium transporter families in Saccharum. *BMC Genomics* **2019**, *20*, 1471–2164. [CrossRef] [PubMed]
34. Trapnell, C.; Roberts, A.; Goff, L.; Pertea, G.; Kim, D.; Kelley, D.R.; Pimentel, H.; Salzberg, S.L.; Rinn, J.L.; Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **2012**, *7*, 562–578. [CrossRef] [PubMed]
35. Clough, S.J.; Bent, A.F. Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant J.* **1998**, *16*, 735–743. [CrossRef] [PubMed]
36. Priyadarshani, S.; Hu, B.; Li, W.; Ali, H.; Jia, H.; Zhao, L.; Ojolo, S.P.; Azam, S.M.; Xiong, J.; Yan, M.; et al. Simple protoplast isolation system for gene expression and protein interaction studies in pineapple (*Ananas comosus* L.). *Plant Methods* **2018**, *14*, 95. [CrossRef]
37. Xie, T.; Chen, C.; Li, C.; Liu, J.; Liu, C.; He, Y. Genome-wide investigation of WRKY gene family in pineapple: evolution and expression profiles during development and stress. *BMC Genomics* **2018**, *19*, 490. [CrossRef]

International Journal of
Molecular Sciences

MDPI

Article

Exploring the Molecular Mechanism underlying the Stable Purple-Red Leaf Phenotype in *Lagerstroemia indica* cv. Ebony Embers

Zhongquan Qiao †, Sisi Liu †, Huijie Zeng, Yongxin Li, Xiangying Wang, Yi Chen, Xiaoming Wang * and Neng Cai *

Hunan Academy of Forestry, 658 South Shaoshan Road, Changsha 410004, China;
qiaozhongquan110@163.com (Z.Q.); liusisi274@126.com (S.L.); run507@163.com (H.Z.); yh5403@sohu.com (Y.L.); S13617494939@163.com (X.W.); lotofa@163.com (Y.C.)

* Correspondence: lky2090@163.com (X.W.); pengjiqing17@csuft.edu.cn (N.C.)

† These authors contributed equally to this work.

Received: 18 October 2019; Accepted: 8 November 2019; Published: 11 November 2019

Abstract: *Lagerstroemia indica* is an important ornamental tree worldwide. The development of cultivars with colorful leaves and increased ornamental value represents one of the current main research topics. We investigated the anthocyanin profiles in two contrasting cultivars for leaf color phenotypes and explored the underlying molecular basis. Both cultivars display purple-red young leaves (Stage 1), and when the leaves mature (Stage 2), they turn green in HD (Lagerstroemia Dynamite) but remain unchanged in ZD (Lagerstroemia Ebony Embers). Seven different anthocyanins were detected, and globally, the leaves of ZD contained higher levels of anthocyanins than those of HD at the two stages with the most pronounced difference observed at Stage 2. Transcriptome sequencing revealed that in contrast to HD, ZD tends to keep a higher activity level of key genes involved in the flavonoid–anthocyanin biosynthesis pathways throughout the leaf developmental stages in order to maintain the synthesis, accumulation, and modification of anthocyanins. By applying gene co-expression analysis, we detected 19 key MYB regulators were co-expressed with the flavonoid–anthocyanin biosynthetic genes and were found strongly down-regulated in HD. This study lays the foundation for the artificial manipulation of the anthocyanin biosynthesis in order to create new *L. indica* cultivars with colorful leaves and increased ornamental value.

Keywords: *Lagerstroemia indica*; gene expression; ornamental value; anthocyanins; leaf coloration; directional improvement

1. Introduction

Lagerstroemia indica L. is a deciduous shrub and small tree of the genus Lagerstroemia with a great ornamental value thanks to its attractive blossom, long-lasting flowering period, and vase-shaped features [1]. It originated in China and has long been used in landscaping in major cities, including Anyang, Fuyang, and Jincheng [2]. *L. indica* cultivars have a wide range of flower colors (white, red, purple, and their combined variants), which contrast with a dark green foliage. However, the few existing cultivars with both colorful flowers and leaves have attracted a great interest and are the prime choice on the market [3]. Therefore, the development of new cultivars with colorful leaves and increased ornamental value has become one of the key research directions in breeding programs. In line with this, the United States Department of Agriculture has released the cultivar 'Lagerstroemia Ebony Embers', which has stable purple-red leaves throughout its leaf development [4]. So far, efforts to develop new *L. indica* cultivars have been mainly based on traditional breeding techniques [5–8]. Hence, it is still tedious to achieve the directional improvement of leaf color in *L. indica*. It is expected that modern molecular techniques will considerably facilitate and accelerate the improvement of leaf color in *L. indica* [9]. However, this requires a detailed understanding of the molecular mechanism of color formation in leaves of *L. indica*.

Color formation is one the most investigated and fascinating research questions in ornamental plants. Flavonoids, particularly anthocyanins, have been reported as the main coloring pigments in plants [10]. Anthocyanins provide a large spectrum of colors ranging from orange/red to violet/blue. Over the past decades, numerous works have clarified the biosynthetic pathway of anthocyanins, which is a very well-conserved network in plant species [11,12]. The key structural genes that catalyze the early and late steps of anthocyanin biosynthesis have been revealed and include phenylalanine ammonia-lyase (PAL), chalcone synthase (CHS), chalcone isomerase (CHI), flavonone 3-hydroxylase (F3H), flavonoid 3′-monooxygenase (F3′H), dihydroflavonol 4-reductase (DFR), anthocyanin synthase (ANS), and UDP-glucose-flavonoid 3-*O*-glucosyltrasnferase (UFGT) [13]. The specific variation in the expression levels of these structural genes through various and complex regulation mechanisms results in quantitative and qualitative variations of anthocyanins, underlying the difference of colorations observed between species, genotypes, organs, or even between various positions on the same plant tissue. Transcription factors (TF) such as MYB, basic helix loop-helix, and WD40 genes were reported to be the key modulators of the anthocyanin biosynthetic structural genes [14–16], but other regulators belonging to the TF families of WRKY and NAC have also been discovered [17–19]. Moreover, recent studies have demonstrated that genetic mutations and microRNAs represent other forms of regulation of the anthocyanin biosynthetic genes [20,21]. The species-specific peculiarity of anthocyanin regulation mechanisms justifies the numerous studies on color formation in plants.

The overall goal of this study is to clarify the molecular mechanism of color formation in leaves of *L. indica*. To achieve this objective, we explored the key anthocyanins conferring the purple-red color in leaves of 'Lagerstroemia Ebony Embers' compared to the cultivar 'Lagerstroemia Dynamite', which features green-colored mature leaves. In addition, we investigated the competition mechanism between different branches of anthocyanin biosynthesis and the TFs regulating the anthocyanin biosynthetic genes. The findings from this study will guide the artificial manipulation of the anthocyanin biosynthesis in order to create new cultivars with colorful leaves and increased ornamental value.

2. Results

2.1. *Anthocyanin Analysis in the Leaves of the Two* Lagerstroemia indica *Cutlivars*

Two cultivars of *Lagerstroemia indica* with different leaf color phenotypes were studied. The two cultivars display purple-red young leaves (Stage 1), and when the leaves mature (Stage 2), they turn into green color in HD (Lagerstroemia Dynamite) but remain unchanged in ZD (Lagerstroemia Ebony Embers) (Figure 1A–D). Anthocyanins are known to be the major coloring pigments in plants [10]. We characterized the anthocyanin contents in the leaf samples of the two cultivars at the two stages of development. Seven anthocyanins, including peonidin O-hexoside, rosinidin O-hexoside, cyanidin O-syringic acid, cyanidin 3-O-glucoside (kuromanin), delphinidin 3-O-glucoside (mirtillin), cyanidin 3,5-O-diglucoside (cyanin), and cyanidin were detected (Table S1). Quantitative profiles showed that cyanidin was only detected in the leaves of HD, while the six other anthocyanins were present at different concentrations in the two cultivars. Globally, the leaves of ZD contained higher levels of anthocyanins than those of HD at the two stages with the most pronounced difference observed at Stage 2 (Figure 1E). In addition, the total anthocyanin content decreased from Stage 1 to Stage 2 in both cultivars, but the decrease was more conspicuous in HD (Figure 1E). This suggests that the leaf color change observed at Stage 2 in HD is associated with a significant decrease of total anthocyanins. Next, the concentrations of each metabolite were compared between the two cultivars (ZD-1_vs_HD-1 and ZD-2_vs_HD-2) and between the two developmental stages (HD-1_vs_HD-2 and ZD-1_vs_ZD-2) in order to identify the differentially accumulated metabolites (DAM) with the following parameters: variable importance in projection ≥1 and fold change ≥2 or fold change ≤0.5 [22]. In total, we found five, two, four, and seven DAM for HD-1_vs_HD-2, ZD-1_vs_ZD-2, ZD-1_vs_HD-1, and ZD-2_vs_HD-2, respectively (Table 1). This result further supports the premise that a strategy toward maintaining a higher content of all detected anthocyanins (except cyanidin) in ZD underpins the stable leaf coloration observed throughout the developmental stages.

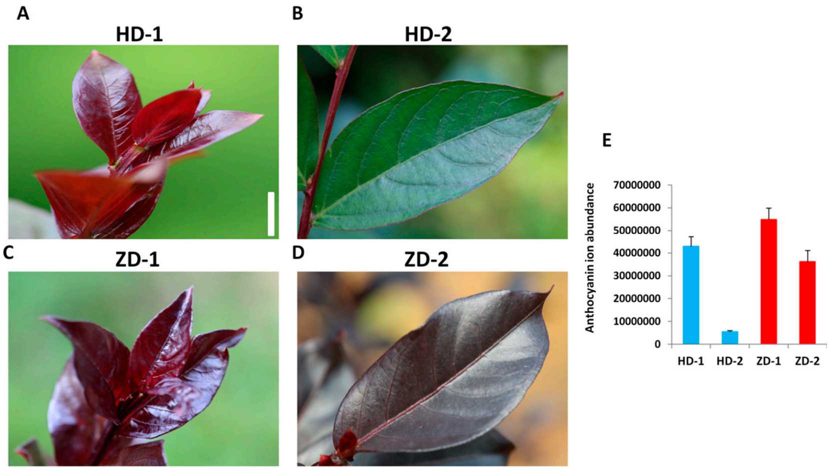


Figure 1. The phenotypes of young (**A**,**C**) and mature (**B**,**D**) leaves of HD and ZD. The bar represents 1.2 cm. (**E**) Total anthocyanin content measured in leaves of HD and ZD at young (Stage 1, S1) and mature (Stage 2, S2) stages. HD represents the cultivar Lagerstroemia Dynamite, while ZD represent the cultivar Lagerstroemia Ebony Embers. Data are from three replicate samples and represent the ion abundance of the anthocyanin compounds.

Table 1. Concentration of the anthocyanins detected in the two *L. indica* cutlivars and their log2 fold-change values. Bold values are those significantly changed between compared samples.

Compounds	**HD-1**	**ZD-1**	**HD-2**	**ZD-2**	**Log2 Fold Change**			
					HD-1_vs_HD-2	**ZD-1_vs_ZD-2**	**ZD-1_vs_HD-1**	**ZD-2_vs_HD-2**
Peonidin O-hexoside	24257333.3	13935333.3	2037100	12121666.7	**−3.57**	−0.20	0.80	**−2.57**
Rosinidin O-hexoside	89211.3333	144180	0	43131	**−13.28**	**−1.74**	−0.69	**−12.23**
Cyanidin O-syringic acid	2205133.33	4702233.33	144803.333	2828266.67	**−3.93**	−0.73	**−1.09**	**−4.29**
Cyanidin 3-O-glucoside	7512833.33	17313666.7	443633.333	8068766.67	**−4.08**	**−1.10**	**−1.20**	**−4.18**
Delphinidin 3-O-glucoside	3606100	5734733.33	2009400	5926300	−0.84	0.05	−0.67	**−1.56**
Cyanidin 3,5-O-diglucoside	4405800	13112000	394216.667	7217400	**−3.48**	−0.86	**−1.57**	**−4.19**
Cyanidin	801780	0	484516.667	0	−0.73	0.00	**16.44**	**15.72**

2.2. *De Novo Transcriptome Assembly and Gene Expression Profiles in the Two* L. indica *Cutlivars at Different Leaf Developmental Stages*

In order to decode the genes involved in the differential leaf color phenotype in HD and ZD, we de novo sequenced and assembled the transcriptome from leaf samples of the two cultivars at the two stages and in triplicate. In total, 12 RNA-seq were generated, yielding a total of 283 millions reads and 84 Gb of clean data with 94% of bases scoring Q30 and above (Table 2). Using the Trinity software, 45,925 unigenes were assembled. To predict the functions of these genes, various databases were searched, including COG (18475), GO (33922), KEGG (17473), KOG (26429), Pfam (36768), Swiss-Prot (32110), eggNOG (42527), and NR (43088), resulting in a total of 43,208 functionally annotated genes (Figure 2A–D). NR database homologous species distribution analysis showed that *L. indica* (Lythraceae) shares 40% of its genes with *Eucalyptus grandis* (Myrtaceae), both species belonging to the same order:

Myrtales (Figure 2E). We searched for genes encoding transcription factors (TF) and obtained a total of 2504 TFs classified into various families (Table S2).

Table 2. Overview of the transcriptome sequencing dataset and quality check.

SampleID.	Raw Read Number	Base Number	GC (%)	Q20 (%)	Q30 (%)
HD-1_1	29963667	8930322034	51.21	97.91	94.18
HD-2_1	29373060	8751104486	51.26	97.85	94.09
HD-2_2	20168830	6023966824	50.7	97.82	93.98
HD-2_3	23631048	7060390512	50.81	97.83	94
HD-1_2	23919615	7129091888	51.23	98	94.41
HD-1_2	23467265	7003557242	50.74	97.88	94.23
ZD-1_1	24273602	7232878206	51.32	97.69	93.8
ZD-1_2	21891679	6530271554	51.24	97.97	94.22
ZD-1_3	22505602	6715319880	51.25	97.88	94.12
ZD-2_1	21339681	6361383236	51.24	97.85	93.99
ZD-2_2	21255491	6336179806	51.32	97.8	93.93
ZD-2_3	22210402	6625694066	51.27	97.88	94.04

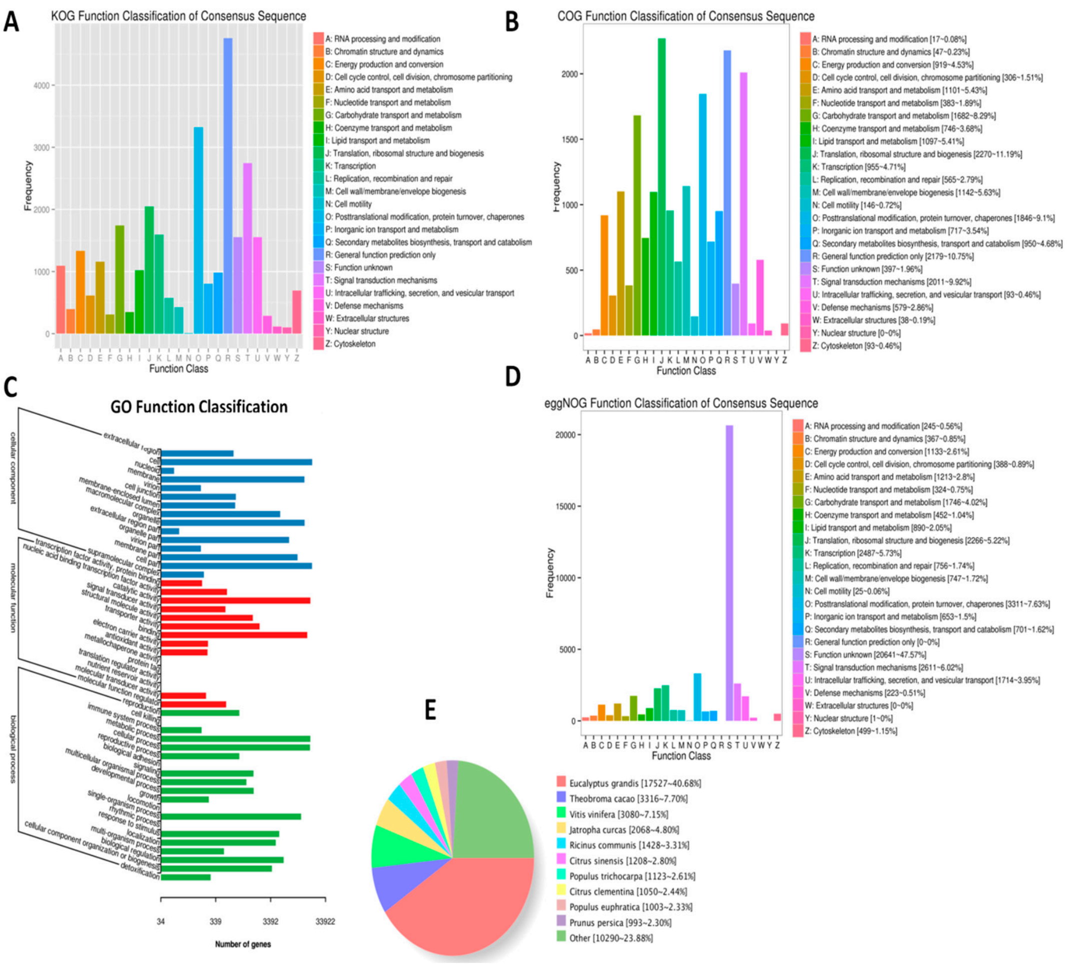


Figure 2. Functional annotation of the unigenes detected in *L. indica* based on orthologs from various databases. (**A**) KOG; (**B**) COG; (**C**) GO; (**D**) EggNOG; and (**E**) NR database homologous species distribution analysis.

Gene expression levels were estimated with the fragments per kilobase of exon per million fragments mapped (FPKM) values ranging from 0.01 to 786,889 (Figure 3A). To assess the quality of the replicate samples, we performed hierarchical clustering analysis based on FPKM data. The result showed that all the biological replicates clustered together, suggesting a high reliability of our sequencing data (Figure 3B). Moreover, two main groups were displayed, including one group (G1) for the green-colored leaf samples (HD-2) and one group (G2) for the purple-red-colored samples (HD-1, ZD-1 and ZD-2). Also, G2 could be split into two subgroups, including G2_1 gathering samples from the young stage of both cultivars and G2_2 gathering samples from the mature stage of ZD. Overall, the sample clustering pattern was clearly according to the leaf color phenotype.

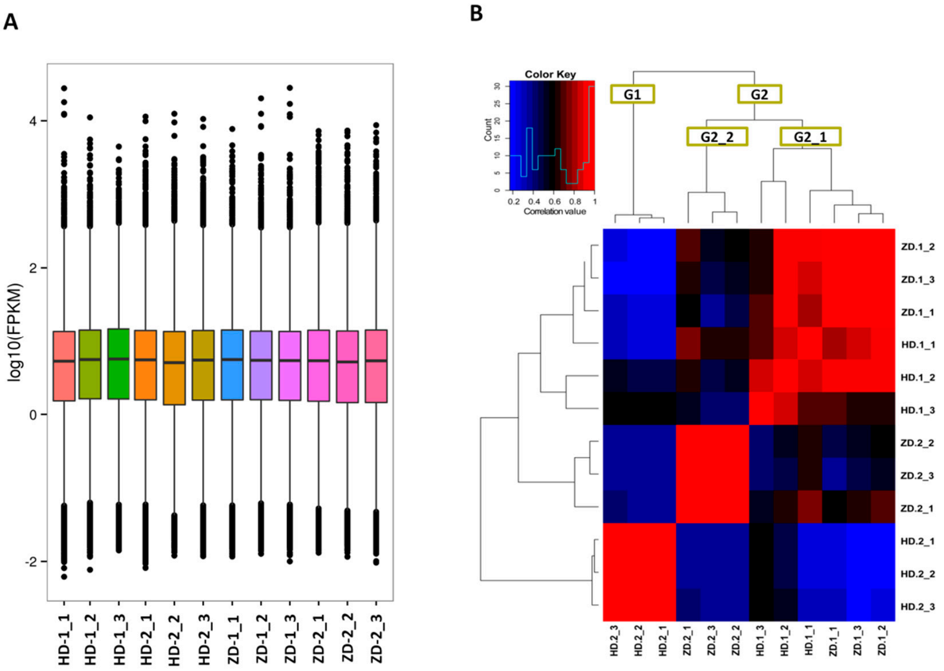


Figure 3. Overview of the transcriptome sequencing in *L. indica* leaves. (**A**) Gene expression profiles in the 12 libraries. HD represents the cultivar Lagerstroemia Dynamite, while ZD represents the cultivar Lagerstroemia Ebony Embers; (**B**) heatmap clustering showing correlation among samples based on global expression profiles. G1 and G2 represent the two major groups of samples, while G2_1 and G2_2 represent the subgroups of the group G2.

2.3. Differentially Expressed Genes between the Two L. indica *Cutlivars*

To uncover the genes involved in the different levels of anthocyanins in leaves of HD and ZD, the gene expression values expressed as FPKM were compared between cultivars and developmental stages. The differentially expressed genes (DEG) were detected with the following parameters: fold change >2 and a false discovery rate correction set at $p < 0.01$. The results showed large numbers of DEGs between compared pair of samples HD-1_vs_HD-2 (9247), ZD-1_vs_ZD-2 (9852), ZD-1_vs_HD-1 (13990), and ZD-2_vs_HD-2 (14688) (Figure 4). Gene ontology (GO) enrichment analysis was performed for these four types of DEGs (Figure S1). The metabolic process and cellular process were the most enriched GO terms in the biological process, the cell and cell part were the most enriched cellular component GO terms, while catalytic activity and binding were clearly enriched as molecular functions. These results suggest that transcription factors (binding activity) and high enzymatic activity are involved in the modulation of leaf coloration in *L. indica*. The highest numbers of DEGs (approximately 1/3 of total expressed genes) were observed by comparing the two cultivars independently of the stages, which indicates a large variation in their genetic make up. We focused our analysis on the genes related to the flavonoid–anthocyanin biosynthesis and MYB transcription factors detected within the DEGs.

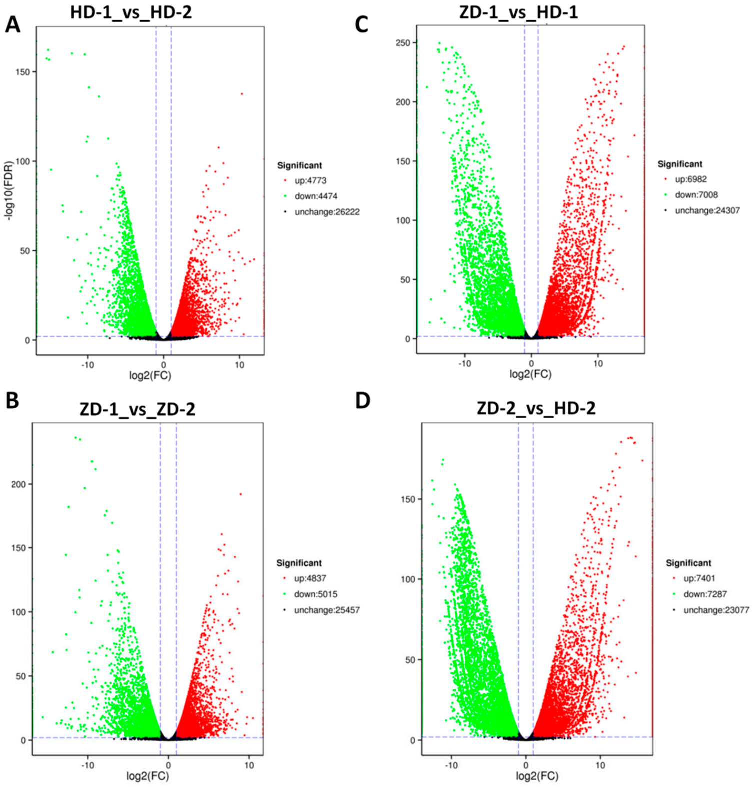


Figure 4. Volcano plot showing the up-regulated genes, down-regulated genes, and genes that were not regulated between pairs of compared samples. (**A**) HD-1_vs_HD-2, (**B**) ZD-1_vs_ZD-2; (**C**) ZD-1_vs_HD-1; and (**D**) ZD-2_vs_HD-2. HD represents the cultivar Lagerstroemia Dynamite, while ZD represents the cultivar Lagerstroemia Ebony Embers.

2.4. DEGs Related to the Flavonoid–Anthocyanin Biosynthesis and Mechanisms Underlying the Differential Leaf Color Phenotypes

Since we observed a higher content of total anthocyanins in the leaves of ZD than HD at the two developmental stages (Figure 1E), we further compared the expressed genes related to the flavonoid–anthocyanin biosynthesis between the two cultivars at each stage. We obtained 74 and 71 DEGs at Stage 1 and Stage 2, respectively, resulting in a total of 96 DEGs with the majority of these DEGs being higher expressed in ZD than HD, particularly at Stage 2 (Table S3). This result is not intriguing, and shows that a stronger activity of genes related to the flavonoid–anthocyanin biosynthesis in the leaves of ZD promotes a higher synthesis and accumulation of anthocyanins. To elucidate the molecular mechanism underlying the change in leaf color observed in HD while the leaf color of ZD remained stable, we investigated the DEGs between the two developmental stages in each cultivar. In total, 74 (10 up-regulated and 64 down-regulated) and 52 (nine up-regulated and 43 down-regulated) genes involved in the flavonoid–anthocyanin biosynthesis were differentially expressed from Stage 1 to Stage 2 in HD and ZD, respectively. The higher number of altered genes, particularly the down-regulated genes, in HD as compared to ZD highlights a mechanism to strongly limit anthocyanin biosynthesis. Thirty DEGs displayed the same pattern between the two cultivars, suggesting that they are not involved in the differential leaf color phenotype (Table S4). In contrast, 44 DEGs showed either opposite patterns between the two cultivars or a huge difference in the gene expression fold change from Stage 1 to Stage 2. We infer that these genes are crucial for the differential leaf color phenotype observed in the two cultivars. To better understand how these genes affect the leaf color, we mapped them on the flavonoid–anthocyanin biosynthesis pathways (Figure 5), which have been well characterized in plants [11,12]. The main precursors for flavonoids are 4-coumaroyl CoA and three molecules of malonyl CoA that produce chalcones by chalcone synthase (CHS) [13].

We identified 17 chalcone synthase [EC:2.3.1.74] (CHS) genes, including 16 strongly down-regulated in HD from Stage 1 to Stage 2, but these genes were unaltered or just slightly down-regulated in ZD. Flavanones are produced from chalcones via chalcone isomerase [EC:5.5.1.6] (CHI). We detected four CHI down-regulated in HD, but they were all unaffected in ZD at Stage 2. The pathway is further catalyzed by flavanone 3-hydroxylase [EC:1.14.11.9] (F3H) to yield dihydrokaempferol and subsequently by flavonoid 3′-monooxygenase [EC:1.14.14.82] (F3′H) to yield dihydroquercetin. We found 10 F3′H DEGs, nine of which were strongly silenced in HD from Stage 1 to Stage 2, but were unaffected or just slightly down-regulated in ZD. Dihydroflavonol 4-reductase (DFR) catalyzes the synthesis of leucoanthocyanidins, which could be converted into anthocyanidins (by anthocyanidin synthase [EC:1.14.20.4] (ANS)) or proanthocyanidins (by leucoanthocyanidin reductase [EC:1.17.1.3] (LAR)). In contrast to the previous flavonoid–anthocyanin biosynthetic DEGs, we found only one LAR (F01_transcript/54491) up-regulated in ZD but not affected in HD from Stage 1 to Stage 2, denoting a mechanism toward a high accumulation of proanthocyanidins in ZD leaves. Finally, anthocyanidins are converted into anthocyanins via UDP-flavonoid glucosyl transferase (UFGT) [13]. UFGT genes are active in the last step of anthocyanin modifications and without their actions, anthocyanins are unstable and can not accumulate in the cells to give the purple-red pigmentation [23]. In this study, we observed 12 various DEGs involved in this last step of anthocyanin modification, including two flavonol 3-O-glucosyltransferase [EC:2.4.1.91] (UFGT), seven anthocyanidin 3-O-glucosyltransferase [EC:2.4.1.115] (UA3GT), two anthocyanidin 5,3-O-glucosyltransferase [EC:2.4.1.-] (GT1), and one anthocyanidin 3-O-glucoside 2′′′-O-xylosyltransferase [EC:2.4.2.51] (UGT). Interestingly, the expression levels of 11 out of these 12 genes were highly repressed from Stage 1 to Stage 2 in HD, while the majority was stably expressed in ZD. Distinctively, the gene F01_transcript/29830 (anthocyanidin 3-O-glucosyltransferase (UA3GT) was strongly up-regulated in ZD but was found to be repressed in HD. Collectively, our results demonstrate that in contrast to HD, ZD tends to keep a high activity level of key genes involved in the flavonoid–anthocyanin biosynthesis pathways throughout the leaf developmental stages in order to maintain the synthesis, accumulation, and modification of anthocyanins (probably proanthocyanidins, too).

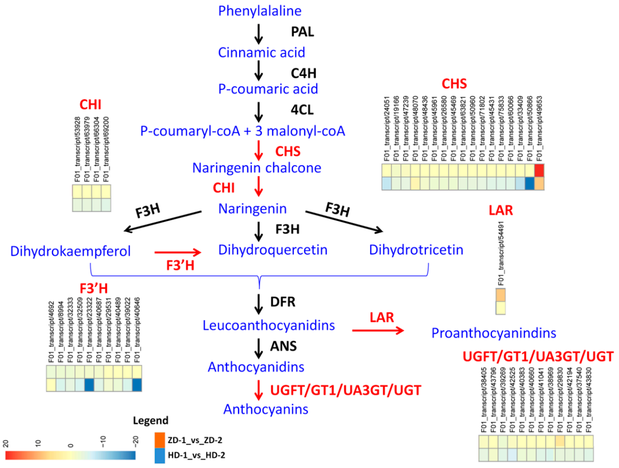


Figure 5. Flavonoid–anthocyanin biosynthetic genes in *L. indica*. The differentially expressed genes between HD and ZD from Stage 1 (young leaves) to Stage 2 (mature leaves) are highlighted in red color. Phenylalanine ammonia-lyase (PAL), cinnamic acid 4-hydroxylase (C4H), 4 coumarate CoA ligase (4CL), chalcone synthase (CHS), chalcone isomerase (CHI), flavanone 3-hydroxylase (F3H), flavonoid 3′-monooxygenase (F3′H), dihydroflavonol 4-reductase (DFR), by anthocyanidin synthase (ANS), leucoanthocyanidin reductase (LAR), UDP-flavonoid glucosyl transferase (UFGT), anthocyanidin 3-O-glucosyltransferase (UA3GT), anthocyanidin 5,3-O-glucosyltransferase (GT1), and anthocyanidin 3-O-glucoside 2′′′-O-xylosyltransferase (UGT). The heatmap show the log2 fold change of the gene expression from Stage 1 to Stage 2. HD represents the cultivar Lagerstroemia Dynamite, while ZD represents the cultivar Lagerstroemia Ebony Embers.

2.5. Active MYB Transcripion Factors Regulating Gene Expression for the Differential Leaf Color Phenotypes

It has been documented in several plant species that the structural genes involved in the flavonoid–anthocyanin biosynthesis pathways are mainly regulated by MYB transcription factors [24]. In total, 663 and 620 TFs DEGs were involved in gene regulation activity from Stage 1 to Stage 2 in ZD and HD, respectively (Table S5). Among these TFs, we retrieved 61 and 60 MYB TFs in ZD and HD, respectively. Comparative analysis of the gene expression fold change of these MYB TFs showed that 36 MYBs were commonly differentially expressed in both cultivars with similar fold changes within each stage (Table S6). However, we uncovered 49 other MYBs genes, which exhibited contrasting expression patterns between the two cultivars and are likely to be the key regulators of the structural genes involved in the flavonoid–anthocyanin biosynthesis pathways in *L. indica* (Table S7).

The gene co-expression network approach constructs the network of genes (co-expressed modules) with co-activation across a group of samples. Genes with similar expression patterns under multiple, but resembling experimental conditions have a high probability of sharing similar functions or being involved in related biological pathways [25,26]. To better decipher the regulation pattern of the structural genes involved in the flavonoid–anthocyanin biosynthesis pathways by the candidate regulator MYBs, we performed a gene co-expression analysis [25]. To give more power to the gene co-expression analysis, we further sequenced the transcriptome from leaves of HD at two intermediate stages (IS-1 and IS-2) between Stage 1 and Stage 2, when the leaf color gradually changes form purple-red to green (Figure S2, Table S8). Gene co-expression analysis of a total of 18 RNA-seq data resulted into 22 co-expressed gene modules (Figure S3, Table S9). Interestingly, 19 key MYB regulators and 32 flavonoid–anthocyanin biosynthetic genes were co-expressed in three different modules: dark red, yellow, and blue. In each of these modules, the MYB transcription factors have a high probability of regulating the target co-expressed flavonoid–anthocyanin biosynthetic genes. In the dark red module, six MYB regulators are co-expressed with 18 structural genes, including CHS, CHI, F3′H, UFGT, and UA3G (Figure 6A,B). The MYBs were preferentially down-regulated in HD from Stage 1 to Stage 2, which correlated with the strong down-regulation of the target structural genes in HD. This suggests that MYBs from this module are positive modulators of color formation in leaves of *L. indica*. Similarly in the yellow module, six MYBs were strongly down-regulated in HD as compared to ZD, and this correlated with a more reduced expression level of the structural genes (CHS, F3′H, UFGT, GT1, and UA3G) in HD (Figure 6C,D). Finally, in the blue module, seven MYBs were preferentially up-regulated in ZD to induce the expression levels of one LAR and one UA3G gene (Figure 6E,F). Globally, the gene co-expression analysis revealed that MYBs are positive modulators of the structural genes and the strong down-regulation of most of these MYB regulators from Stage 1 to Stage 2 observed in HD may limit the activity of the enzymes that catalyze the flavonoid–anthocyanin biosynthesis pathways, resulting in a reduced anthocyanin accumulation in the leaves.

To confirm the differential expression levels of the candidate structural genes and the co-expressed MYB transcription factors detected by the RNA-seq analysis, we conducted a quantitative real-time PCR on 21 selected genes from all modules. As expected, the qRT-PCR results were well correlated with the RNA-seq report (R^2 = 0.84; Figure S4), demonstrating the reliability of the report from this study.

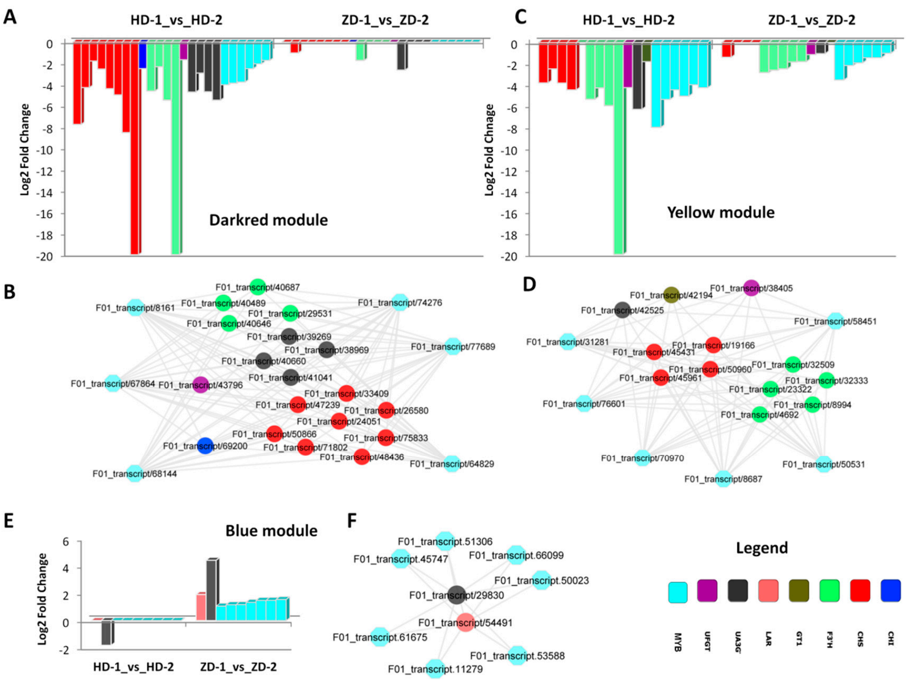


Figure 6. Gene co-expression analysis of the MYB regulators and their target genes related to the flavonoid–anthocyanin biosynthesis pathways in *L. indica*. (**A**,**B**) expression fold change and network of the genes co-expressed in the dark red module; (**B**,**C**) expression fold change and network of the genes co-expressed in the yellow module; (**D**,**E**) expression fold change and network of the genes co-expressed in the blue module. HD represents the cultivar Lagerstroemia Dynamite, while ZD represents the cultivar Lagerstroemia Ebony Embers.

3. Discussion

Although it is well known that anthocyanins are the key pigments coloring plant organs [10], their composition and concentration greatly vary among plant species [27]; therefore, it is impossible to predict the key molecules underlying specific colorations in plants without a detailed metabolic profiling. Anthocyanidins are the aglycone units of anthocyanins, and there are six major types found widely in plants, namely pelargonidin, cyanidin, peonidin, delphinidin, petunidin, and malvidin [28]. There was no previous report of the leaf anthocyanin profile in *L. indica*, but several studies were conducted on the flowers of different cultivars. Collectively, four anthocyanins were reported in *L. indica* flowers, including delphinidin 3-O-glucoside, petunidin 3-Oglucoside, cyanidin 3-O-glucoside, and malvidin 3-O-glucoside [29–32]. It is worth mentioning that these authors analyzed flowers with various colors, including purple-red, purple, purple-violet, violet, and white. In the present study, we obtained a less diverse set of anthocyanins, and only two of the flower anthocyanins (delphinidin 3-O-glucoside and cyanidin 3-O-glucoside) were detected in the leaves. This is understandable since only one leaf color was studied here. The intriguing findings in this study are those regarding the diversity of cyanidin glycoside-derived and methylated-derived compounds in both cultivars (cyanidin O-syringic acid, cyanidin 3-O-glucoside, cyanidin 3,5-O-diglucoside, and rosinidin O-hexoside), although cyanidin was only found in HD. The absence of cyanidin in ZD implies that it is systematically converted into the glycoside-derived and methylated-derived forms. Pelargonidin and cyanidin are the red series pigments in plants [29]. The absence of pelargonidin in the leaves of both cultivars indicate that cyanidin derivatives represent the main molecules conferring the purple-red coloration. Our results are in agreement with the previous report of Zhang et al. [32], who showed that cyanidin 3-O-glucoside was mainly concentrated in cultivars with purple-red flowers.

RNA sequencing offers the opportunity to simultaneously profile the expression levels of thousand of genes [33]. Zhang et al. [34] and Wang et al. [35] sequenced the leaf transcriptome in

L. indica to study the flowering regulatory genes and powdery mildew disease responsive genes, respectively. Globally, these authors assembled ~37000 genes, which is lower than the number of unigenes reported in the present study (45925). The difference in the numbers of detected genes may be attributed to the advanced sequencing platform and bioinformatic packages employed for unigene assembly in this work. Our goal was to explore the molecular mechanism underlying the differential leaf color phenotypes in the two cutivars, with a focus on the genes involved in the biosynthetic pathway of anthocyanins and their regulators [11,12]. In fact, the quantitative and qualitative variation of anthocyanins in plants are strongly correlated with the differential expression of key structural genes involved in the anthocyanin biosynthesis pathways [18,36]. In this study, several classes of structural genes related to the flavonoid–anthocyanin biosynthesis were differentially expressed between the two cultivars and have been mapped to the early steps (chalcone synthase (CHS), chalcone isomerase (CHI), and flavonoid 3′-monooxygenase (F3′H)) and late steps (leucoanthocyanidin reductase (LAR), UDP-flavonoid glucosyl transferase (UFGT), anthocyanidin 3-O-glucosyltransferase (UA3GT), anthocyanidin 5,3-O-glucosyltransferase (GT1), and anthocyanidin 3-O-glucoside 2′′′-O-xylosyltransferase (UGT)) (Figure 5) [12]. These genes were globally down-regulated from the young leaf stage to the mature stage in both cultivars; however, we noticed that ZD tends to maintain a stronger activity as compared to HD (Table S3), which presumably favors the observed high accumulation of anthocyanins in ZD (Figure 1E). More often, genes belonging to either the early steps or the late steps, but not both simultaneously have been reported to differentially modulate anthocyanin contents in contrasting colored samples. For example, Chen et al. [37] demonstrated that low expression levels of C4H, CHS, and F3H in white petals, contrarily to the red petals of peach, reduce the formation of dihydro-kaempferol (DHK), and thereby inhibit the anthocyanin accumulation. In addition, Jiao et al. [38] showed that PAL was weakly expressed in the white-flesh peach and limits anthocyanin production. In contrast, Zhuang et al. [21] showed that a strong anthocyanin accumulation in purple turnip was attributed to an up-regulation of DFR, ANS, and UFGT genes. LAR converts leucoanthocyanidins into proanthocyanidins. In this study, the up-regulation of the LAR gene (*F01_transcript/54491*) in ZD suggests an increment of the proanthocyanidins content from Stage 1 to Stage 2, but this mechanism may not be relevant to the stable leaf coloration observed in ZD, since proanthocyanidins are colorless in nature [39]. The class of genes involved in the modification of anthocyanidins (UDP-flavonoid glucosyl transferase (UFGT), anthocyanidin 3-O-glucosyltransferase (UA3GT), anthocyanidin 5,3-O-glucosyltransferase (GT1), and anthocyanidin 3-O-glucoside 2′′′-O-xylosyltransferase (UGT)) was particularly enriched (Figure 5). This was expected, since the anthocyanins detected in leaf samples were mainly glycoside-derived compounds (Table S1). Anthocyanidins are highly unstable and easily susceptible to degradation; therefore, glycosylation is essential to stabilize them [40]. Furthermore, glycosylation serve as a signal for transport of the anthocyanins to vacuoles, where they can function as pigments [41]. Since most of these genes were higher expressed in ZD than HD, correlating with the stronger content of glycoside-derived anthocyanins, we deduce that the glycosylation of anthocyanins (particularly cyanidin) is a key mechanism for the stable purple-red colored leaf phenotype observed in ZD, exactly as previously demonstrated in peach [42].

The expression levels of structural genes involved in the flavonoid–anthocyanin pathway are in part regulated by transcription factors (TF), particularly by the MYB family members [24]. We uncovered 49 candidate MYBs that are likely to be the key regulators of the structural genes involved in the flavonoid–anthocyanin pathway (Table S7). Many studies have reported several differentially expressed MYB genes as potential regulators, but the target genes of each specific MYB regulator and the regulatory network are often overlooked. The mechanisms of gene expression regulation by a TF could be simple (direct binding to the binding motif in the promoter region of the targets) or more complex, involving other cofactors. An example of the complex regulation mechanism is the feed-forward loop mechanism where a three-gene pattern is composed of two input transcription factors, one of which regulates the other, which both jointly regulate a target gene [43]. Hence, it is essential to clearly delineate the network of interaction between candidate MYBs and their targets in order to facilitate the directional manipulation of the expression levels of the structural genes involved in the flavonoid–anthocyanin pathway. In this study, we revealed three co-expressed modules containing candidate MYB regulators and their target structural genes (Figure 6). Overall, we found that MYBs are positive regulators of these structural genes; therefore, increasing the activity levels of some MYBs from these co-expressed modules, particularly those from the dark red and yellow modules, may have high potential to confer stable purple-red coloration in the leaves of HD and other *L. indica* cultivars.

4. Materials and Methods

4.1. Plant Materials

Two cultivars of *Lagerstroemia indica* L. were used as plant materials. The cultivar Lagerstroemia Dynamite was developed by the Carl Whitcomb breeding program, (Carl Whitcomb Lacebark Inc. Stillwater, OK, USA) and features purple-red young leaves, which gradually turn into a green color when they mature (Figure 1, Figure S2). The second cultivar, Lagerstroemia Ebony Embers released by the USDA, displays stable purple-red leaves throughout all the leaf developmental stages. In this study, Lagerstroemia Dynamite and Lagerstroemia Ebony Embers were named as HD and ZD, respectively. Both cultivars were grown under natural environmental conditions in the experimental station of the Hunan Academy of Forestry, China. Leaf blades were collected at different developmental stages (Figure 1, Figure S2) from three independent plants (2 years old) of each cultivar, quickly frozen in liquid nitrogen, and stored at −80 °C until further use.

4.2. Anthocyanin Analysis

The sample preparation, extract analysis, anthocyanin identification and quantification were performed at Wuhan MetWare Biotechnology Co., Ltd. (www.metware.cn) following their standard procedures and previously described by Cao et al. [22]. Before the data analysis, quality control (QC) analysis was conducted to confirm the reliability of the data. The QC sample was prepared by the mixture of sample extracts and inserted into every four samples to monitor the changes in repeated analyses. Data matrices with the intensity of the metabolite features from the samples were uploaded to the Analyst 1.6.1 software (AB SCIEX, Ontario, Canada) for statistical analyses. The supervised multivariate method, partial least squares-discriminant analysis (PLS-DA), was used to maximize the metabolome differences between the pair of samples. The relative importance of each metabolite to the PLS-DA model was checked using the parameter called variable importance in projection (VIP). Metabolites with VIP $\geq$1 and fold change $\geq$2 or fold change $\leq$0.5 were considered as differential metabolites for group discrimination [22].

4.3. Transcriptome Sequencing and Data Analysis

RNA extraction, transcriptome library preparation, sequencing, and bioinformatics analysis were conducted at the Biomarker Technologies (Beijing, China, www.biomarker.com.cn) following their standard procedures and previously described by Zhu et al. [44]. Briefly, total RNA was extracted from the leaf samples using a Spin Column Plant total RNA Purification Kit (Sangon Biotech, Shanghai, China) according to the manufacturer's instructions. Sequencing libraries were constructed following the protocol of the Gene Expression Sample Prep Kit (Illumina, San Diego, CA, USA). The first-strand cDNAs were synthesized from the total RNA with random hexamer primers, followed by second-strand cDNAs synthesis using DNA polymerase I (New England BioLabs, Ipswich, MA, USA) and RNase H (Invitrogen, Waltham, MA, USA). After end repair, adaptor ligation, and the addition of index codes for each sample, PCR amplification was conducted. The purity and quality of the libraries were measured by an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and Qubit 2.0 (Life Technologies, Carlsbad, CA, USA). Then, the libraries were pair-end sequenced by using the Illumina HiSeq 2500 platform (Illumina Inc., San Diego, USA).

The raw RNA-seq reads were quality-checked with the FastQC package (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and adaptor sequences and low-quality reads (containing >50% bases with a Phred quality score <15 and reads with more than 1% ambiguous residues N) were removed. The high-quality reads from all the libraries were de novo assembled into transcripts using the software Trinity (version r20140717, [45]) by employing the paired-end method. Next, the transcripts were realigned to construct unigenes. The assembled unigenes were annotated by searching against various databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [46], Gene Ontology (GO) [47], Clusters of Orthologous Groups (COG) [48], PfAM [49], Swiss-Prot [50], eggNOG [51], NR [52], and euKaryotic Orthologous Groups (KOG) [53] using BLAST [54] with a threshold of E-value $<1.0 \times 10^{-5}$. The software KOBAS2.0 [55] was employed to get the unigene KEGG orthology. The analogs of the unigene amino acid sequences were searched against the Pfam database [48] using the HMMER tool [56] with a threshold of E-value $<1.0 \times 10^{-10}$. The unigenes were counted and normalized into fragments per kilobase of transcript per million fragments mapped reads (FPKM) value using RSEM [57]. Differentially expressed genes (DEGs) between pairs of samples were determined using the EdgeR Bioconductor package [58]. False discovery rate values less than 0.01 and |fold change|≥2 were set as criteria to decide the significant differences in gene expression.

4.4. Gene Co-Expression Analysis

Weighted Gene Co-Expression Network Analysis (WGCNA) package version 1.61 [59] was used to construct the gene co-expression networks from the normalized log2-transformed FPKM matrix as described by Lv et al. [25] and Dossa et al. [60]. Network visualization for the co-expressed gene modules related to MYB and flavonoid–anthocyanin biosynthesis pathways was performed using the Cytoscape software version 3.6.1 [61].

4.5. Quantitative RT-PCR Analysis

Quantitative PCR was performed using the SYBR Premix Ex Taq™ Kit (Takara, Dalian, China) according to the manufacturer's instructions on the StepOne plus Real time PCR Platform (Applied Biosystems, CA, USA) with the following protocol: 95 °C for 10 min, followed by 40 cycles of 95 °C for 15 s, and at 60 °C for 60 s [62]. Each reaction was performed using a 20-μL mixture containing 10 μL of 2 × ChamQ SYBR qPCR Master Mix, 6 μL of nuclease-free water, 1 μL of each primer (10 mM), and 2 μL of four-fold diluted cDNA. All of the reactions were run in 96-well plates, and each cDNA was analyzed in triplicate. Specific primer pairs of 21 selected genes were designed using the Primer Premier 5.0 [63] (Table S10). The *Actin* gene was used as the internal control. Data are presented as relative transcript levels based on the $2^{-\Delta\Delta Ct}$ method [64].

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/20/22/5636/s1. Table S1. Quantification of the detected anthocyanins in the two *L. indica* cultivars at Stage 1 (young leaves) and Stage 2 (mature leaves); Table S2. Genes encoding transcription factors detected in *L. indica* transcriptome; Table S3. List of the differentially expressed genes between ZD and HD at Stage 1 and Stage 2 and their log2 fold change values; Table S4. List of the differentially expressed genes between Stage 1 and Stage 2 in HD and ZD that display the same fold-change patterns between the two cultivars; Table S5. Differentially expressed genes encoding transcription factors form Stage 1 to Stage 2; Table S6. List of the differentially expressed MYB genes between Stage 1 and Stage 2 in HD and ZD that display the same fold-change patterns between the two cultivars; Table S7. List of the differentially expressed MYB genes between Stage 1 and Stage 2 in HD and ZD that display different patterns of fold change of a huge difference between the two cultivars; Table S8. Overview of the transcriptome sequencing dataset and quality check of the leaves from HD at the intermediate stage (IS1 and IS2); Table S9. List of the genes belonging of each of the 22 modules detected through gene co-expression analysis; Table S10. The primer sequences of genes used for quantitative real time PCR; Figure S1. Gene ontology enrichment analysis of the differentially expressed genes between (A) HD-1_vs_HD-2, (B) ZD-1_vs_ZD-2; (C) ZD-1_vs_HD-1; (D) ZD-2_vs_HD-2. HD represents the cultivar Lagerstroemia Dynamite, while ZD represents the cultivar Lagerstroemia Ebony Embers. Figure S2. The phenotypes of HD leaves at the intermediate Stage 1 and Stage 2 when the leaf color is gradually turning from purple-red to green; Figure S3. Dendrogram clustering of the genes and identification of the co-expressed modules; Figure S4. qRT-PCR ($2^{-\Delta\Delta Ct}$) analysis of 21 selected genes within the differentially expressed genes detected in this study. Correlation analysis between qRT-PCR and RNA-seq (log2 fold change).

Author Contributions: Y.C. and N.C. designed and supervised the study. Z.Q., X.W., and S.L. conducted the experiments. Z.Q., H.Z., and Y.L. performed data analysis. X.W. prepared plant materials and carried out qRT-PCR analysis. Z.Q. drafted the manuscript. All authors read and approved the final manuscript.

Funding: This research was funded by Hunan Provincial Natural Science Foundation (2019JJ50305) and Changsha Science and Technology Plan Project (kq1801026).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pounders, C.; Rinehart, T.; Edwards, N.; Knight, P. An analysis of combining ability for height, leaf out, bloom date, and flower color for crapemyrtle. *HortScience* **2007**, *42*, 1496–1499. [CrossRef]
2. Liu, Y.S.; Zetter, R.; Ferguson, D.K.; Zou, C. Lagerstroemia (Lythraceae) pollen from the Miocene of eastern China. *Grana* **2008**, *47*, 262–271. [CrossRef]
3. Cabrera, R.I. Evaluating and promoting the cosmopolitan and multipurpose Lagerstroemia. In Proceedings of the XXVI International Horticultural Congress, Toronto, ON, Canada, 11–17 August 2002; pp. 177–184.
4. Pounders, C.; Scheffler, B.E.; Rinehart, T.A. 'Ebony Embers', 'Ebony Fire', 'Ebony Flame', 'EbonyGlow', and 'Ebony and Ivory' Dark-leaf Crapemyrtles. *HortScience* **2013**, *48*, 1568–1570. [CrossRef]
5. Zhengkang, P. Cultivation Managements of *Lagerstroemia indica* and its Application in the Landscape. *J. Anhui Agric. Sci.* **2006**, *34*, 5832–5833.
6. Knox, G. New Crapemyrtles with Burgundy Leaves from Spring through Fall. UF/IFAS Extension. 21 July 2014. Available online: http://nwdistrict.ifas.ufl.edu/hort/2014/07/21/new-crapemyrtles-with-burgundy-leaves-from-spring-through-fall/ (accessed on 15 February 2019).
7. Pounders, C.T.; Blythe, E.K.; Fare, D.C.; Knox, G.W.; Sibley, J.L. Crapemyrtle genotype × environment interactions, and trait stability for plant height, leaf-out, and flowering. *HortScience* **2010**, *45*, 198–207. [CrossRef]
8. Gilman, E.F.; Watson, D.G.; Klein, R.W.; Koeser, A.K.; Hilbert, D.R.; McLean, D.C. *Lagerstroemia indica*: Crapemyrtle. UF/IFAS Extension, 2018, ENH-501. Available online: https://edis.ifas.ufl.edu (accessed on 15 February 2019).
9. Mohan Jain, S.; Brar, D.S. *Molecular Techniques in Crop Improvement*, 2nd ed.; Springer: Dordrecht, The Netherlands, 2010. [CrossRef]
10. Chen, C. Overview of plant pigments. In *Pigments in Fruits and Vegetables*; Springer: New York, NY, USA, 2015; pp. 1–7.
11. Tanaka, Y.; Ohmiya, A. Seeing is believing: Engineering anthocyanin and carotenoid biosynthetic pathways. *Curr. Opin. Biotechnol.* **2008**, *19*, 190–197. [CrossRef] [PubMed]
12. Jaakola, L. New insights into the regulation of anthocyanin biosynthesis in fruits. *Trends Plant Sci.* **2013**, *18*, 477–483. [CrossRef] [PubMed]
13. Dixon, R.A.; Steele, C.L. Flavonoids and isoflavonoids—A gold mine for metabolic engineering. *Trends Plant Sci.* **1999**, *4*, 394–400. [CrossRef]
14. Moyano, E.; Martinez-Garcia, J.F.; Martin, C. Apparent redundancy in myb gene function provides gearing for the control of flavonoid biosynthesis in antirrhinum flowers. *Plant Cell* **1996**, *8*, 1519–1532. [PubMed]
15. To, K.Y.; Wang, C.K. Molecular breeding of flower color. In *Floriculture Ornamental and Plant Biotechnology: Advances and Topical Issues Volume I*; Silva, T.D., Ed.; Global Science Books: London, UK, 2006; pp. 300–310.
16. Li, Z.; Zhao, M.; Jin, J.; Zhao, L.; Xu, Z. Anthocyanins and their biosynthetic genes in three novel-colored Rosa rugosa cultivars and their parents. *Plant Physiol. Biochem.* **2018**, *129*, 421–428. [CrossRef] [PubMed]
17. Jiang, G.; Li, Z.; Song, Y.; Zhu, H.; Lin, S.; Huang, R.; Jiang, Y.; Duan, X. LcNAC13 Physically Interacts with LcR1MYB1 to Coregulate Anthocyanin Biosynthesis-Related Genes during Litchi Fruit Ripening. *Biomolecules* **2019**, *9*, 135. [CrossRef] [PubMed]
18. Lloyd, A.; Brockman, A.; Aguirre, L.; Campbell, A.; Bean, A.; Cantero, A.; Gonzalez, A. Advances in the MYB–bHLH–WD repeat (MBW) pigment regulatory model: Addition of a WRKY factor and co-option of an anthocyanin MYB for betalain regulation. *Plant Cell Physiol.* **2017**, *58*, 1431–1441. [CrossRef] [PubMed]
19. Zhou, H.; Lin-Wang, K.; Wang, H.; Gu, C.; Dare, A.P.; Espley, R.V.; He, H.; Allan, A.C.; Han, Y. Molecular genetics of blood-fleshed peach reveals activation of anthocyanin biosynthesis by NAC transcription factors. *Plant J.* **2015**, *82*, 105–121. [CrossRef] [PubMed]

20. He, L.; Tang, R.; Shi, X.; Wang, W.; Cao, Q.; Liu, X.; Wang, T.; Sun, Y.; Zhang, H.; Li, R.; et al. Uncovering anthocyanin biosynthesis related microRNAs and their target genes by small RNA and degradome sequencing in tuberous roots of sweet potato. *BMC Plant Biol.* **2019**, *19*, 232. [CrossRef] [PubMed]
21. Zhuang, H.; Lou, Q.; Liu, H.; Han, H.; Wang, Q.; Tang, Z.; Ma, Y.; Wang, H. Differential Regulation of Anthocyanins in Green and Purple Turnips Revealed by Combined De Novo Transcriptome and Metabolome Analysis. *Int. J. Mol. Sci.* **2019**, *20*, 4387. [CrossRef] [PubMed]
22. Cao, H.; Ji, Y.; Li, S.; Lu, L.; Tian, M.; Yang, W.; Li, H. Extensive Metabolic Profiles of Leaves and Stems from the Medicinal Plant Dendrobium officinale Kimura et Migo. *Metabolites* **2019**, *9*, 215. [CrossRef] [PubMed]
23. Fosket, D.E. The Genetic Basis of Plant Development. In *Fosket, Plant Growth and Development*; Donald, E., Ed.; Academic Press: Cambridge, MA, USA, 1994; pp. 41–78. [CrossRef]
24. Allan, A.C.; Hellens, R.P.; Laing, W.A. MYB transcription factors that colour our fruit. *Trends Plant Sci.* **2008**, *13*, 99–102. [CrossRef] [PubMed]
25. Lv, Y.; Xu, L.; Dossa, K.; Zhou, K.; Zhu, M.; Xie, H.; Tang, S.; Yu, Y.; Guo, X.; Zhou, B. Identification of putative drought-responsive genes in rice using gene co-expression analysis. *Bioinformation* **2019**, *15*, 480–489. [CrossRef] [PubMed]
26. Lee, H.K.; Hsu, A.K.; Sajdak, J.; Qin, J.; Pavlidis, P. Overexpression analysis of human genes across many microarray data sets. *Genome Res.* **2004**, *14*, 105–1094. [CrossRef] [PubMed]
27. Veberic, R.; Slatnar, A.; Bizjak, J.; Stampar, F.; Mikulic-Petkovsek, M. Anthocyanin composition of different wild and cultivated berry species. *LWT Food Sci. Technol.* **2015**, *60*, 509–517. [CrossRef]
28. Kong, J.M.; Chia, L.S.; Goh, N.K.; Chia, T.F.; Brouillard, R. Analysis and biological activities of anthocyanins. *Phytochemistry* **2003**, *64*, 923–933. [CrossRef]
29. Egolf, D.R.; Santamour, F.S. Anthocyanin pigments and breeding potential in crape myrtle (*Lagerstreomia indica* L.) and rose of Sharon (Hibiscus syriacus L.). *HortScience* **1975**, *10*, 223–224.
30. Toki, K. Anthocyanin pigments and breeding potential of blue flowers in *Lagerstreomia indica*. *BioHort* **1989**, 73–77.
31. Toki, K.; Katsuyama, N. Pigments and color variation in flowers of *Lagerstroemia indica*. *J. Jpn. Soc. Hortic. Sci.* **1995**, *63*, 853–861. [CrossRef]
32. Zhang, J.; Wang, L.-S.; Gao, J.-M.; Shu, Q.-Y.; Li, C.-H.; Yao, J.; Hao, Q.; Zhang, J.-J. Determination of Anthocyanins and Exploration of Relationship between Their Composition and Petal Coloration in Crape Myrtle (*Lagerstroemia* hybrid). *J. Integr. Plant Biol.* **2008**, *50*, 581–588. [CrossRef] [PubMed]
33. Van Dam, S.; Craig, T.; de Magalhães, J.P. GeneFriends: A human RNA-seq-based gene and transcript co-expression database. *Nucleic Acids Res.* **2015**, *43*, D1124–D1132.
34. Zhang, Z.Y.; Wang, P.; Li, Y.; Ma, L.L.; Li, L.F.; Yang, R.T.; Ma, Y.Z.; Wang, S.; Wang, Q. Global transcriptome analysis and identification of the flowering regulatory genes expressed in leaves of *Lagerstroemia indica*. *DNA Cell Biol.* **2014**, *33*, 680–688. [CrossRef] [PubMed]
35. Wang, X.; Shi, W.; Rinehart, T. Transcriptomes That Confer to Plant Defense against Powdery Mildew Disease in *Lagerstroemia indica*. *Int. J. Genom.* **2015**. [CrossRef] [PubMed]
36. Naing, A.H.; Kim, C.K. Roles of R2R3-MYB transcription factors in transcriptional regulation of anthocyanin biosynthesis in horticultural plants. *Plant Mol. Biol.* **2018**, *98*, 1–18. [CrossRef] [PubMed]
37. Chen, Y.; Mao, Y.; Liu, H.; Yu, F.; Li, S.; Yin, T. Transcriptome analysis of differentially expressed genes relevant to variegation in peach flowers. *PLoS ONE* **2014**, *9*, e90842. [CrossRef] [PubMed]
38. Jiao, Y.; Ma, R.J.; Shen, Z.J.; Yan, J.; Yu, M.L. Gene regulation of anthocyanin biosynthesis in two blood-flesh peach (*Prunus persica* (L) Batsch) cultivars during fruit development. *J. Zhejiang Univ. Sci. B* **2014**, *15*, 809–819. [CrossRef] [PubMed]
39. He, F.; Pan, Q.H.; She, Y.; Duan, C.Q. Biosynthesis and genetic regulation of proanthocyanidins in plants. *Molecules* **2008**, *13*, 2674–2703. [CrossRef] [PubMed]
40. Hellström, J.; Mattila, P.; Karjalainen, R.O. Stability of anthocyanins in berry juices stored at different temperatures. *J. Food Compos. Anal.* **2013**, *31*, 12–19. [CrossRef]
41. Ono, E.; Homma, Y.; Horikawa, M.; Kunikane-Doi, S.; Imai, H.; Takahashi, S.; Kawai, Y.; Ishiguro, M.; Fukui, Y.; Nakayama, T. Functional differentiation of the glycosyltransferases that contribute to the chemical diversity of bioactive flavonol glycosides in grapevines (Vitis vinifera). *Plant Cell* **2010**, *22*, 2856–2871. [CrossRef] [PubMed]

42. Cheng, J.; Wei, G.; Zhou, H.; Gu, C.; Vimolmangkang, S.; Liao, L.; Han, Y.P. Unraveling the mechanism underlying the glycosylation and methylation of anthocyanins in peach. *Plant Physiol.* **2014**, *166*, 1044–1058. [CrossRef] [PubMed]
43. Mangan, S.; Alon, U. Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 11980–11985. [CrossRef] [PubMed]
44. Zhu, C.; Li, X.; Zheng, J. Transcriptome profiling using Illumina- and SMRT-based RNAseq of hot pepper for in-depth understanding of genes involved in CMV infection. *Gene* **2018**, *666*, 123–133. [CrossRef] [PubMed]
45. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full length transcriptome assembly from RNA Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [CrossRef] [PubMed]
46. Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **2004**, *32*, D277–D280. [CrossRef] [PubMed]
47. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef] [PubMed]
48. Tatusov, R.L.; Galperin, M.Y.; Natale, D.A. The COG database: A tool for genome scale analysis of protein functions and evolution. *Nucleic Acids Res.* **2000**, *28*, 33–36. [CrossRef] [PubMed]
49. Finn, R.D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; et al. Pfam: The protein families database. *Nucleic Acids Res.* **2013**, *42*, 222–230. [CrossRef] [PubMed]
50. Apweiler, R.; Bairoch, A.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; et al. UniProt: The Universal Protein knowledgebase. *Nucleic Acids Res.* **2004**, *32*, D115–D119. [CrossRef] [PubMed]
51. Huerta-Cepas, J.; Szklarczyk, D.; Forslund, K.; Cook, H.; Heller, D.; Walter, M.C.; Rattei, T.; Mende, D.R.; Sunagawa, S.; Kuhn, M.; et al. eggNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **2015**, *44*, 286–293. [CrossRef] [PubMed]
52. Deng, Y.Y.; Li, J.Q.; Wu, S.F.; Zhu, Y.P.; Chen, Y.W.; He, F.C. Integrated nr Database in Protein Annotation System and Its Localization. *Comput. Eng.* **2006**, *32*, 71–74.
53. Koonin, E.V.; Fedorova, N.D.; Jackson, J.D.; Jacobs, A.R.; Krylov, D.M.; Makarova, K.S.; Mazumder, R.; Mekhedov, S.L.; Nikolskaya, A.N.; Rao, B.S.; et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **2004**, *5*, R7. [CrossRef] [PubMed]
54. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]
55. Xie, C.; Mao, X.; Huang, J.; Ding, Y.; Wu, J.; Dong, S.; Kong, L.; Gao, G.; Li, C.; Wei, L. KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* **2011**, *39*, W316–W322. [CrossRef] [PubMed]
56. Eddy, S.R. Profile hidden Markov models. *Bioinformatics* **1998**, *14*, 755–763. [CrossRef] [PubMed]
57. Li, B.; Colin, N.D. RSEM: Accurate transcript quantification from RNA Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [CrossRef] [PubMed]
58. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef] [PubMed]
59. Su, G.; Morris, J.H.; Demchak, B.; Bader, G.D. Biological network exploration with Cytoscape 3. *Curr. Protoc. Bioinform.* **2014**, *47*, 1–24.
60. Dossa, K.; Mmadi, M.A.; Zhou, R.; Zhang, T.; Su, R.; Zhang, Y.; Wang, L.; You, J.; Zhang, X. Depicting the Core Transcriptome Modulating Multiple Abiotic Stresses Responses in Sesame (*Sesamum indicum* L.). *Int. J. Mol. Sci.* **2019**, *20*, 3930. [CrossRef] [PubMed]
61. Langfelder, P.; Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **2008**, *9*, 559. [CrossRef] [PubMed]
62. Dossa, K.; Mmadi, M.A.; Zhou, R.; Zhou, Q.; Yang, M.; Cisse, N.; Diouf, D.; Wang, L.; Zhang, X. The contrasting response to drought and waterlogging is underpinned by divergent DNA methylation programs associated with transcript accumulation in sesame. *Plant Sci.* **2018**, *277*, 207–217.

63. Lalitha, S. Primer premier 5. *Biotechnol. Softw. Internet Rep.* **2000**, *1*, 270–272.
64. Livak, K.J.; Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta Ct}$ method. *Methods* **2001**, *25*, 402–408. [CrossRef] [PubMed]

International Journal of
Molecular Sciences

Article

Engineered Artificial MicroRNA Precursors Facilitate Cloning and Gene Silencing in Arabidopsis and Rice

Dandan Zhang [1,2,†], Nannan Zhang [1,3,†], Wenzhong Shen [1,*] and Jian-Feng Li [1,*]

[1] State Key Laboratory of Biocontrol, Guangdong Provincial Key Laboratory of Plant Resources, MOE Key Laboratory of Gene Function and Regulation, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China; ddzhang@iastate.edu (D.Z.); zhangnn23@mail2.sysu.edu.cn (N.Z.)

[2] Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50011, USA

[3] Guangdong Provincial Key Laboratory of Sugarcane Improvement and Biorefinery, Guangdong Bioengineering Institute, Guangzhou 510316, China

[*] Correspondence: shenwzh5@mail.sysu.edu.cn (W.S.); lijfeng3@mail.sysu.edu.cn (J.-F.L.); Tel.: +86-20-39943513 (W.S. & J.-F.L.)

[†] These authors contributed equally to this work.

Received: 24 September 2019; Accepted: 8 November 2019; Published: 10 November 2019

Abstract: Plant genome sequences are presently deciphered at a staggering speed, due to the rapid advancement of high-throughput sequencing technologies. However, functional genomics significantly lag behind due to technical obstacles related to functional redundancy and mutant lethality. Artificial microRNA (amiRNA) technology is a specific, reversible, and multiplex gene silencing tool that has been frequently used in generating constitutive or conditional mutants for gene functional interrogation. The routine approach to construct amiRNA precursors involves multiple polymerase chain reactions (PCRs) that can increase both time and labor expenses, as well as the chance to introduce sequence errors. Here, we report a simplified method to clone and express amiRNAs in Arabidopsis and rice based on the engineered Arabidopsis miR319a or rice miR528 precursor, which harbor restriction sites to facilitate one-step cloning of a single PCR product. Stem-loop reverse-transcriptase quantitative PCR (RT-qPCR) and functional assays validated that amiRNAs can be accurately processed from these modified precursors and work efficiently in plant protoplasts. In addition, Arabidopsis transgenic plants overexpressing the modified miR319a precursor or its derived amiRNA could exhibit strong gene silencing phenotypes, as expected. The simplified amiRNA cloning strategy will be broadly useful for functional genomic studies in Arabidopsis and rice, and maybe other dicotyledon and monocotyledon species as well.

Keywords: plant genome; artificial microRNA; gene silencing; Arabidopsis; rice

1. Introduction

With the advent of whole-genome sequencing technologies, plant genomic data are expanding at an explosive rate. In the post-genomic era, analyzing these genomic data and studying the functions of newly discovered genes is critical for understanding the nature of plant genomes and accelerating the process of crop improvement. One of the most frequently used strategies to study gene function is to create loss-of-function mutants. In past decades, a large number of mutant libraries in model plant species, such as Arabidopsis and rice, have been constructed through physical, chemical, or biological (T-DNA and transposon insertion) mutagenesis [1–3]. However, tedious large-scale screening is required to identify the genes responsible for desired mutant phenotypes [4]. Additionally, random mutagenesis could not broadly cover the whole plant genome. Recently, the powerful CRISPR/Cas9 technology, which enables targeted genome modifications, has already revolutionized plant genome research [5]. Although the CRISPR/Cas9 system is simple, efficient, and highly specific, there are still some limitations related to its application in plant research. First, CRISPR/Cas9-mediated gene disruption is less efficient in targeting heterochromatic regions [6], limiting the range of targetable genes. Second, permanent deletion of essential genes by CRISPR/Cas9 can result in lethality [7,8]. Third, transcripts of many plant genes undergo alternative splicing (AS) in the same or different cell types, producing multiple proteins with different structural domains [9]. However, the CRISPR/Cas9 system is unable to specifically inactivate a certain AS isoform in a cell type-specific manner.

MicroRNAs (miRNAs), a class of endogenous small noncoding RNAs with the size of 21–24 nucleotides, can mediate post-transcriptional and translational gene regulation. miRNAs play important roles in diverse aspects of plant development and plant responses to biotic and abiotic stresses [10,11]. The biogenesis of miRNA is a multistep process that begins with the transcription of a miRNA gene into a primary transcript (pri-miRNA) [12]. Pri-miRNA is sequentially processed into a stem-loop structured precursor (pre-miRNA) by DICER-LIKE1 (DCL1), and pre-miRNA is then processed into miRNA/miRNA* duplex and stabilized by methyltransferase HUA ENHANCER1 (HEN1) [13]. The methylated miRNA duplex is eventually loaded into the ARGONAUTE (AGO) protein to form the so-called RNA-induced silencing complexes (RISCs), followed by the release and degradation of miRNA* [14]. By targeting complementary sequences, RISCs negatively regulate gene expression through mRNA degradation and/or translation inhibition [14,15].

Artificial microRNA (amiRNA) technology has already been successfully developed to silence target gene expression by producing artificially designed miRNAs using the naturally existing miRNA precursor as a backbone [16,17]. Compared to genome editing tools, the amiRNA technology offers more flexibility and reversibility in generating loss-of-function mutants without altering DNA sequences. Since the expression of amiRNAs can be tightly controlled by chemical-inducible or cell/tissue-specific promoters [17], amiRNAs are widely utilized for investigating gene functions associated with mutant lethality [18,19]. Moreover, amiRNA has a high silencing specificity and only recognizes target sequences with less than 5 mismatches [17,20], making it an ideal tool to silence individual AS isoforms or multiple genes sharing short conserved sequences [17,21].

In general, amiRNA-expressing plasmids are constructed according to the method described by Schwab et al. [17], as follows: The miRNA and miRNA* of pre-miR319a are replaced by amiRNA/amiRNA* sequences through site-directed mutagenesis using overlapping polymerase chain reactions (PCRs). However, this method is time-consuming and cost ineffective because it involves four PCRs using three pairs of primers. Here, we report a simplified method for amiRNA cloning. We modified the most commonly used miRNA precursor backbones, pre-miR319a for Arabidopsis or related dicot species, and pre-miR528 for rice or related monocot species, by introducing restriction sites using PCR. With the modified amiRNA backbones, only one PCR is needed to amplify the stem-loop fragment containing a newly designed amiRNA/amiRNA* duplex with restriction enzyme sites, which can then be easily inserted into the engineered pre-miR319a or pre-miR528 in the expression vectors. We also provided evidence that amiRNAs produced in this way can be equally effective in protoplasts or transgenic plants as those produced using the traditional approach.

2. Results

2.1. Strategy for Simplified amiRNA Construction Using a Modified Arabidopsis miRNA319a Backbone

The previous overlapping PCR strategy to assemble a new amiRNA precursor involves four PCRs in two rounds (Figure 1A) [17]. To simplify the procedure and accelerate the amiRNA construction process, we tried to engineer the pre-miR319a backbone by introducing minor changes in its DNA sequences to create restriction sites for amiRNA sequence insertion (Figure 1A–C). For many plant miRNA precursors, the lower stem located ~15 nt below miRNA/miRNA* is critical for miRNA processing. A single change in the lower stem of pre-miRNA can completely abolish miRNA processing [22–24]. The ssRNA (single strand RNA) region, an unpaired region downstream of the lower stem seems to be less important for miRNA production [23,25]. Thus, we modified the pre-miR319a backbone by mutating GAATTG and TCTTGA sequences within the ssRNA region to *Eco*RI (GAATTC) and *Xba*I (TCTAGA) restriction sites, respectively (Figure 1B,C). After the modifications, a single PCR product of the stem-loop fragment containing the amiRNA/amiRNA* sequences can be inserted into the amiRNA backbone (Table S1) using *Eco*RI/*Xba*I, which greatly simplifies the construction procedure and enables possible high-throughput amiRNA construction.

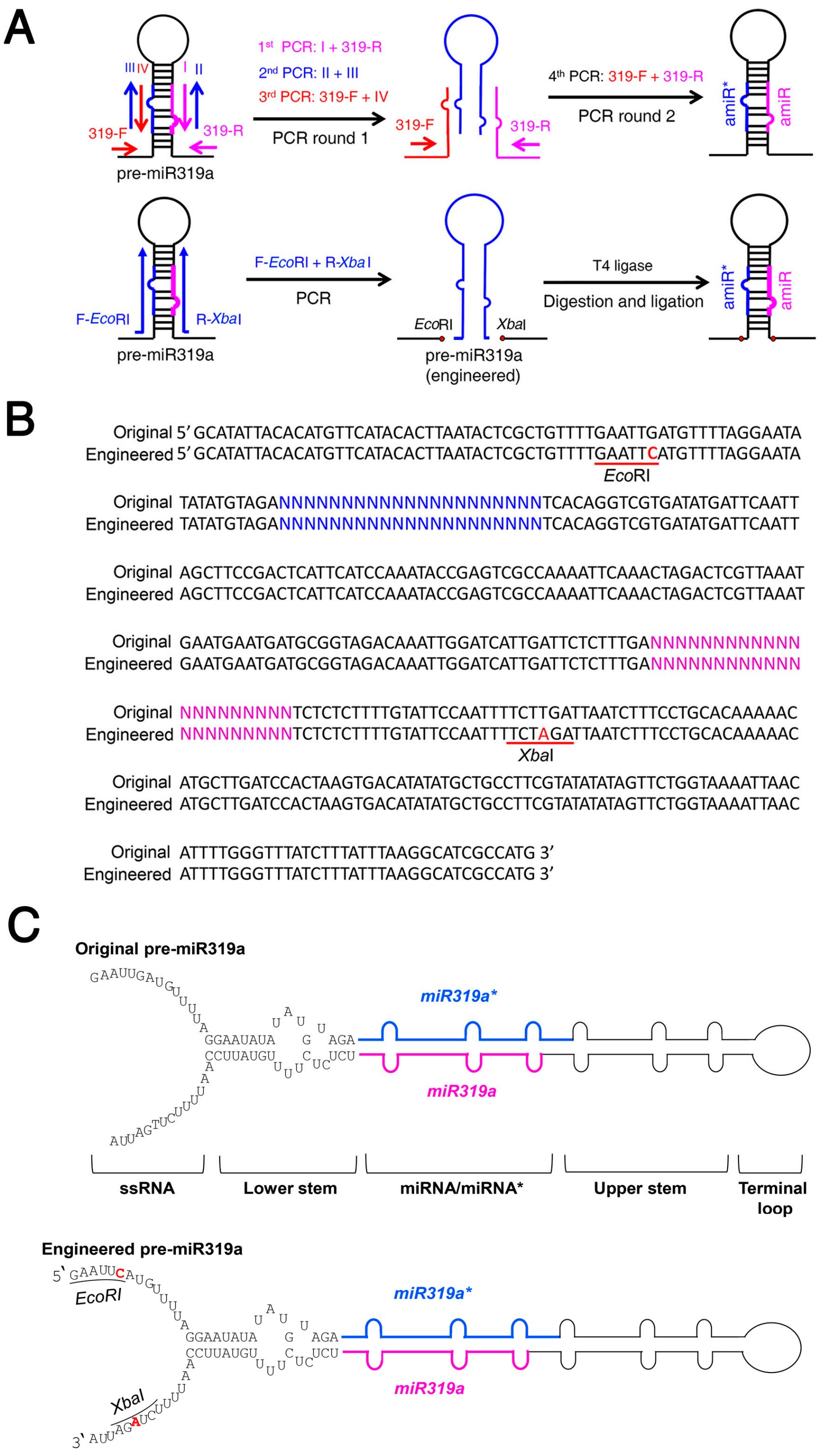

A
III IV
I II
319-F
319-R
pre-miR319a
1st PCR: I + 319-R
2nd PCR: II + III
3rd PCR: 319-F + IV
PCR round 1
4th PCR: 319-F + 319-R
PCR round 2
amiR*
amiR
F-EcoRI
R-XbaI
pre-miR319a
F-EcoRI + R-XbaI
PCR
EcoRI
XbaI
pre-miR319a
(engineered)
T4 ligase
Digestion and ligation
amiR*
amiR
B
Original 5′ GCATATTACACATGTTCATACACTTAATACTCGCTGTTTTGAATTGATGTTTTAGGAATA
Engineered 5′ GCATATTACACATGTTCATACACTTAATACTCGCTGTTTTGAATTCATGTTTTAGGAATA
EcoRI
Original TATATGTAGANNNNNNNNNNNNNNNNNNNNNTCACAGGTCGTGATATGATTCAATT
Engineered TATATGTAGANNNNNNNNNNNNNNNNNNNNNTCACAGGTCGTGATATGATTCAATT
Original AGCTTCCGACTCATTCATCCAAATACCGAGTCGCCAAAATTCAAACTAGACTCGTTAAAT
Engineered AGCTTCCGACTCATTCATCCAAATACCGAGTCGCCAAAATTCAAACTAGACTCGTTAAAT
Original GAATGAATGATGCGGTAGACAAATTGGATCATTGATTCTCTTTGANNNNNNNNNNNN
Engineered GAATGAATGATGCGGTAGACAAATTGGATCATTGATTCTCTTTGANNNNNNNNNNNN
Original NNNNNNNNNTCTCTCTTTTGTATTCCAATTTTCTTGATTAATCTTTCCTGCACAAAAAC
Engineered NNNNNNNNNTCTCTCTTTTGTATTCCAATTTTCTAGATTAATCTTTCCTGCACAAAAAC
XbaI
Original ATGCTTGATCCACTAAGTGACATATATGCTGCCTTCGTATATATAGTTCTGGTAAAATTAAC
Engineered ATGCTTGATCCACTAAGTGACATATATGCTGCCTTCGTATATATAGTTCTGGTAAAATTAAC
Original ATTTTGGGTTTATCTTTATTTAAGGCATCGCCATG 3′
Engineered ATTTTGGGTTTATCTTTATTTAAGGCATCGCCATG 3′
C
Original pre-miR319a
GAAUUGAUGUUUA
GGAAUAUA
CCUUAUGU
AAUUUCUGAUUA
miR319a*
miR319a
ssRNA
Lower stem
miRNA/miRNA*
Upper stem
Terminal loop
Engineered pre-miR319a
5′ GAAUUCAUGUUUA
EcoRI
XbaI
3′ AUUAGAUCUUUAA
miR319a*
miR319a

Figure 1. Engineered Arabidopsis miR319a precursor allows a one-step construction of a new amiRNA

precursor. (**A**) Diagram of amiRNA construction using the engineered miR319a precursor (pre-miR319a). The upper diagram describes an overlapping PCR strategy for generating a new amiRNA precursor by 4 PCRs in two rounds. In the first round, 3 independent PCRs are performed using the indicated primers. A mixture of 3 PCR products is used as a template to conduct the second round of PCR (4th PCR) using the indicated primers. The lower diagram describes a restriction enzyme-based strategy to assemble a new amiRNA precursor. The *Eco*RI/*Xba*I sites are created at the base of miR319a stem-loop in the engineered pre-miR319a. A single PCR product amplified using a pair of mega-primers containing customized amiRNA/amiRNA* sequences, and *Eco*RI/*Xba*I sites are digested by *Eco*RI/*Xba*I and inserted into the same digested engineered pre-miR319a. In the resulting amiRNA precursor, the amiRNA and amiRNA* are colored in magenta and blue, respectively. (**B**) The engineered pre-miR319a contains a G-to-C mutation and a T-to-A mutation that create *Eco*RI and *Xba*I sites (underlined), respectively. The nucleotides in magenta and blue correspond to amiRNA and amiRNA*, respectively. (**C**) Diagram of the original and engineered pre-miR319a. Mutated nucleotides are highlighted in red.

2.2. Engineered Pre-miR319a Generated Functional miR319a as Demonstrated by the Silencing Phenotype

To test whether the engineered pre-miR319a remains functional, we generated Arabidopsis transgenic plants overexpressing the original and engineered pre-miR319a, respectively. It has been previously reported that miR319a controls Arabidopsis leaf development and morphogenesis through targeting and down-regulating the expression of several TCP (Teosinte branched1/Cycloidea/Proliferating cell factor) family members [26–29]. Arabidopsis gain-of-function mutant *jaw-D* with overexpression of miR319a exhibits a jagged and wavy leaf phenotype [26]. As expected, the transgenic plants overexpressing both the original and engineered pre-miR319a showed a curly and serrated leaf phenotype (Figure 2A). The relative abundances of mature miR319a in these transgenic plants were further determined using the stem-loop RT-qPCR technique that is specialized for accurate quantification of mature miRNA [30]. We detected comparable production of mature miR319a from engineered pre-miR319a as original pre-miR319a (Figure 2B). These results suggest that mature miR319a can be generated from the modified pre-miRNA319, as well as from the native pre-miR319a.

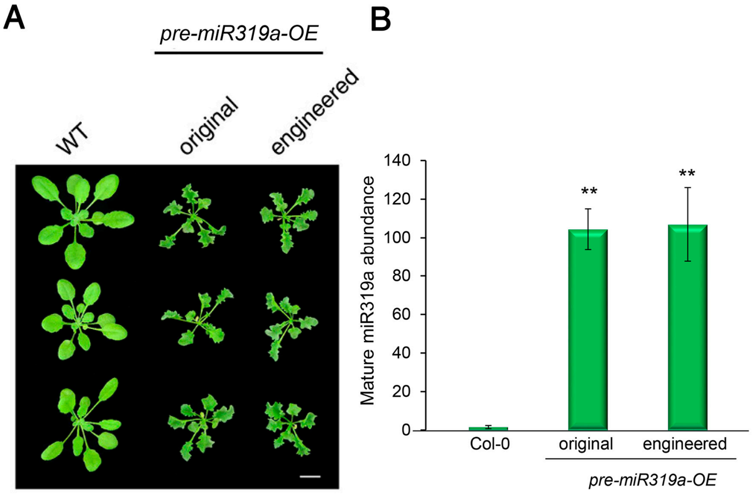


Figure 2. Engineered pre-miR319a retains its function in transgenic plants. (**A**) Phenotypic comparison of transgenic Arabidopsis plants expressing the original or engineered miR319a precursor. Four-week-old plants are shown. Scale bar, 0.5 cm. (**B**) Stem-loop RT-qPCR validates comparable production of mature miR319a from the original or engineered pre-miR319a in transgenic plants. The quantitative PCR data are presented as means ± SD of at least three independent repeats with endogenous *snoR101* expression level set as 1. ** $p < 0.01$ (student's *t* test).

2.3. amiRNAs Produced from Engineered Pre-miR319a Have Comparable Efficiencies in Gene Silencing

To provide more evidence that the modifications of pre-miR319a would not affect amiRNA processing and maturation, we compared side-by-side the silencing efficiencies of amiRNAs produced from the original or engineered pre-miR319a using an ETPamir assay [31]. In ETPamir assay, a target gene encoding epitope-tagged target protein is co-expressed with individual amiRNAs in protoplasts, and the silencing efficiency of each amiRNA is inversely reflected by the accumulation of target proteins, which can be monitored by immunoblotting using anti-tag antibodies [31,32]. By targeting Arabidopsis *PHYTOENE DESATURASE 3* (*PDS3*), *MAP/ERK KINASE KINASE 1* (*MEKK1*) or *MAP KINASE KINASE KINASE 3* (*MAPKKK3*), we found that the amiRNAs produced from the engineered pre-miR319a appeared to be as efficient as or even slightly more effective than those from the original pre-miRNA319a (Figure 3A). We also measured the abundances of mature amiRNAs produced in ETPamir assays by stem-loop RT-qPCR and found that the engineered pre-miR319a could produce comparable or even higher amounts of mature amiRNAs than the original pre-miR319a (Figure 3B). These results imply that the engineered pre-miR319a is fully functional in generating mature amiRNAs.

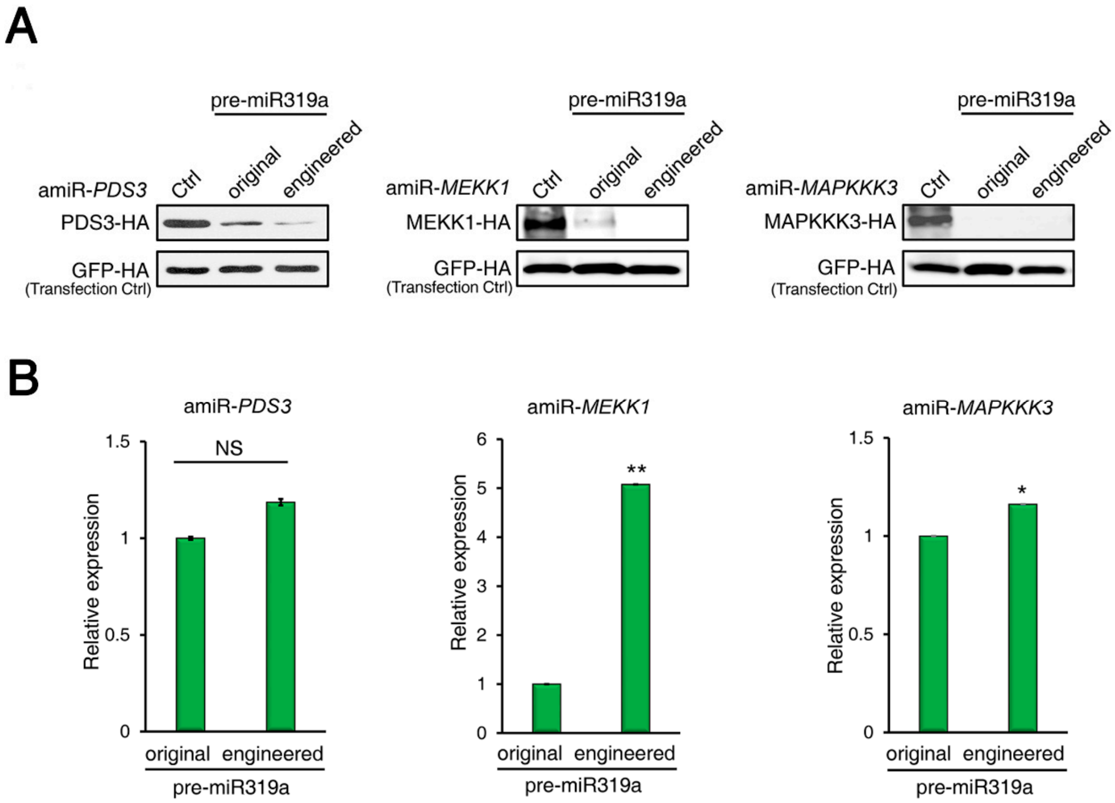


Figure 3. amiRNAs produced from engineered pre-mir319a exhibit equal efficacy in silencing target gene expression. (**A**) Comparison of the performance of amiRNAs produced from the original or engineered pre-miR319a using the ETPamir assay. amiRNAs expressed from engineered pre-miR319a are slightly more effective in silencing Arabidopsis *PDS3*, *MEKK1*, and *MAPKKK3* expression than those expressed from original pre-miR319a in protoplasts. Three independent repeats with GFP-HA as an untargeted internal control produced similar results. (**B**) Detection of mature amiRNAs produced from the original or engineered pre-miR319a in the ETPamir assay. Mature amiRNAs were detected using stem-loop RT-qPCR. The quantitative PCR data represent means ± SD of at least three independent repeats using *mCherry-HA* as a transfection control. * $p < 0.05$, ** $p < 0.01$ (student's t test).

2.4. amiRNAs Produced from Engineered Pre-miR319a Could Effectively Silence Target Gene Expression in Transgenic Plants

Next, we evaluated the efficiencies of amiRNAs produced from engineered pre-miR319a in planta. *CNGC4* (CYCLIC NUCLEOTIDE-GATED CATION CHANNEL 4) was selected as the target gene as the null phenotype of *CNGC4* has been reported [33,34] and is easy to observe. Three amiRNAs targeting *CNGC4* were constructed using the engineered pre-miR319a as backbone and their activities were assessed first by the ETPamir assay. The results showed that three amiR-*CNGC4s* could all suppress *CNGC4* expression, but they displayed different silencing efficiencies. amiR-*CNGC4*-1 could almost completely silence *CNGC4* expression (Figure 4A), whereas amiR-*CNGC4*-2 and amiR-*CNGC4*-3 were less effective. So, we chose amiR-*CNGC4*-1 to silence endogenous *CNGC4* in our transgenic plants. The engineered or original pre-amiR-*CNGC4*-1 construct was subsequently introduced into Arabidopsis Col-0 plants. Transgenic plants overexpressing engineered or original pre-amiR-*CNGC4*-1 both exhibited smaller leaves and shorter petioles relative to the wild-type plants, resembling the dwarf phenotype of *cngc4* T-DNA null mutant (Figure 4B). These results validate that the engineered pre-miR319a can be utilized to produce effective amiRNAs for target gene silencing.

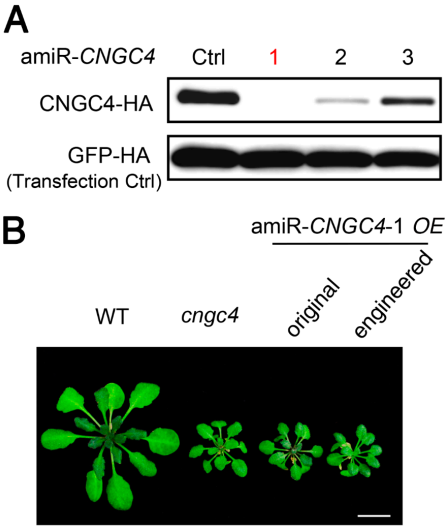


Figure 4. amiRNAs produced from the engineered pre-miR319a are fully functional. (**A**) Comparison of the silencing efficiency of three *CNGC4*-targeting amiRNAs expressed from the engineered pre-miR319a using the ETPamir assay. Note that amiR-*CNGC4*-1 (red) is the most potent amiRNA for silencing *CNGC4*. Three independent repeats with GFP-HA as an untargeted internal control produced similar results. (**B**) Comparison of the performance of amiR-*CNGC4*-1 produced from the original or engineered amiRNA precursor in transgenic plants. *cngc4* is a T-DNA insertion null mutant of *CNGC4*. Scale bar, 0.5 cm.

2.5. Strategy for Simplified amiRNA Construction Using a Modified Rice miRNA528 Backbone

Rice miR528 precursor (pre-miR528) is frequently used for generating amiRNAs and gene silencing in many monocot species [35–37]. To test whether the same strategy can be applied for amiRNA production using pre-miR528, we engineered pre-miR528 by mutating AGGTCT and GAAGTT sequences in the ssRNA region to *Stu*I (AGGCCT) and *Eco*RI (GAATTC) restriction sites, respectively (Figure 5A–C). Therefore, PCR products of the stem-loop fragment containing the amiRNA/amiRNA* duplex can be generated using a pair of mega primers (Table S2) and be readily inserted into the engineered pre-miR528 after *Stu*I/*Eco*RI digestion and ligation.

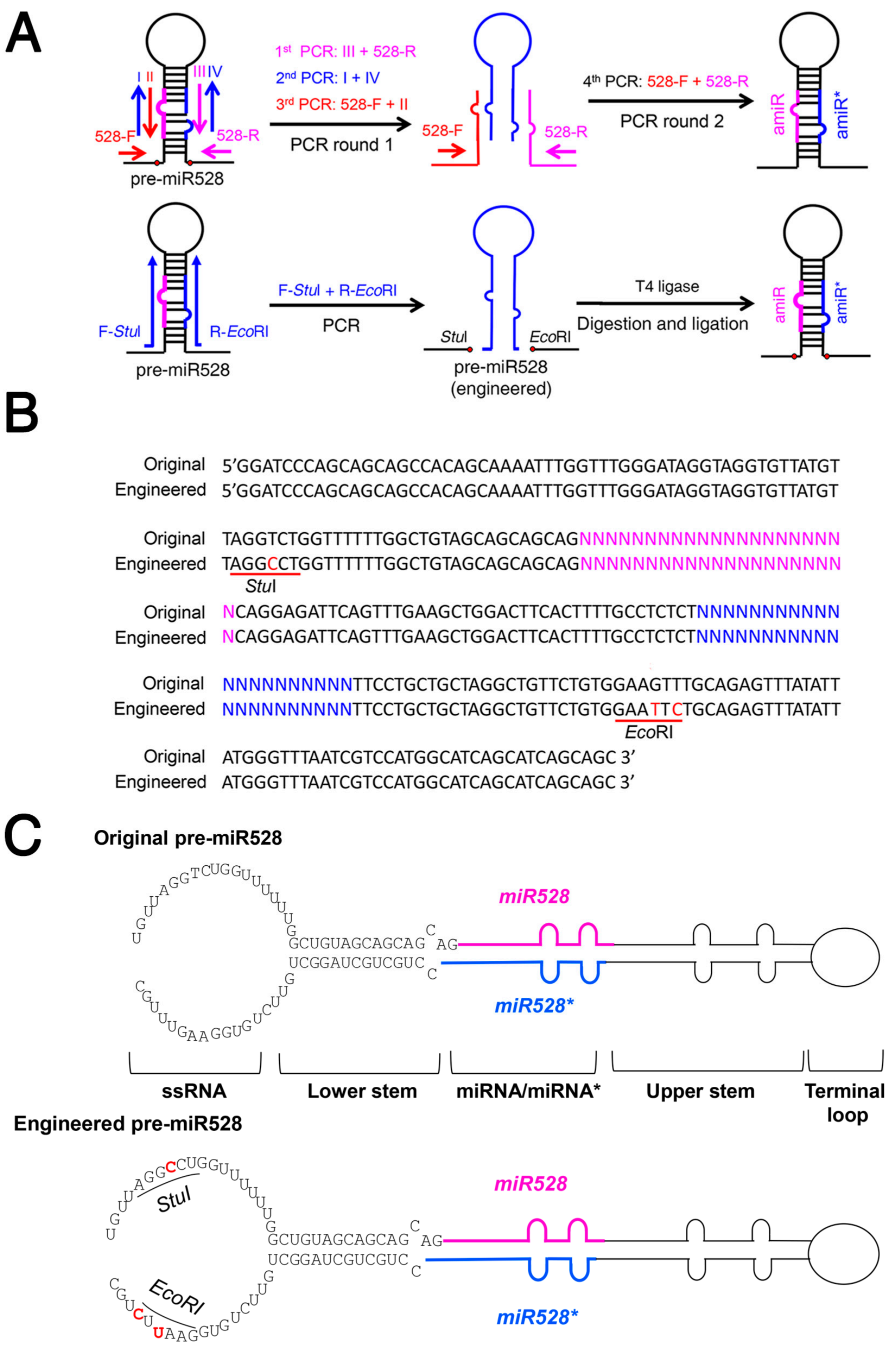

A
I II
III IV
528-F
528-R
pre-miR528
1st PCR: III + 528-R
2nd PCR: I + IV
3rd PCR: 528-F + II
PCR round 1
528-F
528-R
4th PCR: 528-F + 528-R
PCR round 2
amiR
amiR*
F-StuI
R-EcoRI
pre-miR528
F-StuI + R-EcoRI
PCR
StuI
EcoRI
pre-miR528
(engineered)
T4 ligase
Digestion and ligation
amiR
amiR*
B
Original 5’GGATCCCAGCAGCAGCCACAGCAAAATTTGGTTTGGGATAGGTAGGTGTTATGT
Engineered 5’GGATCCCAGCAGCAGCCACAGCAAAATTTGGTTTGGGATAGGTAGGTGTTATGT
Original TAGGTCTGGTTTTTTGGCTGTAGCAGCAGCAGNNNNNNNNNNNNNNNNNNNN
Engineered TAGGCCTGGTTTTTTGGCTGTAGCAGCAGCAGNNNNNNNNNNNNNNNNNNNN
StuI
Original NCAGGAGATTCAGTTTGAAGCTGGACTTCACTTTTGCCTCTCTNNNNNNNNNNN
Engineered NCAGGAGATTCAGTTTGAAGCTGGACTTCACTTTTGCCTCTCTNNNNNNNNNNN
Original NNNNNNNNNNTTCCTGCTGCTAGGCTGTTCTGTGGAAGTTTGCAGAGTTTATATT
Engineered NNNNNNNNNNTTCCTGCTGCTAGGCTGTTCTGTGGAATTCTGCAGAGTTTATATT
EcoRI
Original ATGGGTTTAATCGTCCATGGCATCAGCATCAGCAGC 3’
Engineered ATGGGTTTAATCGTCCATGGCATCAGCATCAGCAGC 3’
C
Original pre-miR528
miR528
miR528*
GCUGUAGCAGCAG
UCGGAUCGUCGUC
ssRNA
Lower stem
miRNA/miRNA*
Upper stem
Terminal loop
Engineered pre-miR528
StuI
EcoRI
miR528
miR528*

Figure 5. The engineered rice miR528 precursor allows a one-step construction of a new amiRNA precursor. (**A**) Diagram of amiRNA construction using the engineered miR528 precursor (pre-miR528). In the engineered pre-miR528, *Stu*I/*Eco*RI sites are created at the base of miR528 stem-loop. The amiRNA and amiRNA* are colored in magenta and blue, respectively. (**B**) The engineered pre-miR528 contains mutations that can create *Stu*I and *Eco*RI sites (underlined), respectively. The nucleotides in magenta and blue correspond to amiRNA and amiRNA*, respectively. (**C**) Diagram of the original and engineered pre-miR528. Mutated nucleotides are highlighted in red.

We next evaluated the silencing efficiencies of amiRNAs produced from the engineered pre-miR528 using the ETPamir assay. The amiRNAs produced from both original and engineered pre-miR528 could trigger efficient silencing of the foreign gene *GFP* and the rice endogenous gene *OsCEBiP* (*CHITIN ELICITOR BINDING PROTEIN*) in rice cells (Figure 6). There is no detectable difference in silencing efficiencies between amiRNAs produced from the original or engineered pre-miR528

(Figure 6). These data indicate that the same strategy could be applied to pre-miR528 engineering, which leads to production of functional amiRNAs in rice, but with a simple cloning procedure.

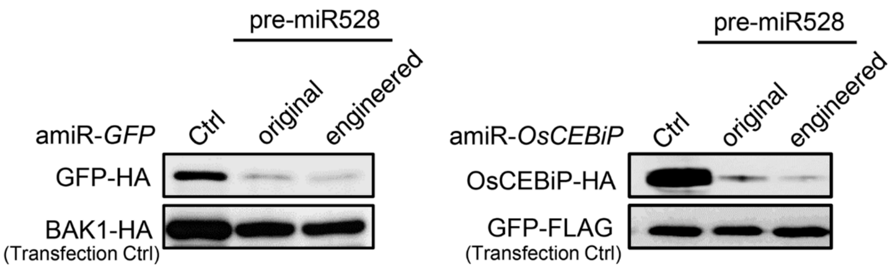


Figure 6. amiRNAs produced from the engineered pre-miR528 are functional in rice cells. Comparison of the performance of amiRNAs produced from the original or engineered pre-miR528 was conducted by the ETPamir assay. Three independent repeats with GFP-FLAG or BAK1-HA as an untargeted internal control produced similar results.

3. Discussion

The amiRNA technology is not only a powerful genetic tool for generating loss-of-function mutants in basic plant research, but is also an effective strategy to engineer crops for beneficial agronomic traits [38,39] and enhanced disease resistance against pathogens [40–43] or pests [44,45]. As amiRNA is produced from an endogenous plant miRNA precursor, and the promoter and terminator of an amiRNA expression cassette can be derived from plants, this technology may raise minimal concerns about introducing foreign genetic elements into engineered crops.

Selecting a suitable miRNA precursor backbone to express amiRNAs is vital for successfully silencing target gene expression. Many plant miRNA precursors such as Arabidopsis miR319a, miR172a, miR395, miR390 and rice miR390, and miR528 have been used as backbones for expressing amiRNAs to confer specific gene silencing [17,35,46–48]. The Arabidopsis miR319a and rice miR528 precursors are the most commonly used amiRNA expression backbones that have been widely used to generate loss-of-function mutants in dicot and monocot plant species, since they are highly conserved across the plant kingdom [17,49,50]. However, the traditional overlapping PCR approach to construct amiRNA plasmid is tedious, time-consuming, and inefficient, especially for high-throughput application [17]. In this study, we provide a simplified method by mutating the sequence of pre-miR319a and pre-miR528 to create restriction sites for subsequent insertion of PCR products (Figures 1 and 5). Therefore, the customized amiRNA/amiRNA* sequences can be easily inserted into the backbone of pre-miR319a and pre-miR528 by one single PCR, followed by restriction digestion and ligation. Our new strategy could dramatically improve the efficiency of amiRNA construction.

Many plant miRNAs such as Arabidopsis miR172a and miR169a are processed in a canonical “base to loop” manner [22–24]. For these miRNA precursors, the secondary structure of the lower stem is essential for miRNA processing. Disruption of the closing bulge structure in the lower stem by point mutation affects miRNA accumulation [23,24]. Meanwhile, other miRNAs such as Arabidopsis miR319a and miR159a have been reported to be processed in a “loop to base” direction [25]. Although complete removal of the lower stem sequences seemed to have little impact on the miR319 production, overexpressing the pre-miR319a lacking the lower stem caused a less severe leaf crinkled phenotype compared with overexpressing the full-length pre-miR319a [25]. We speculated that deletion of the lower stem bases may impair the accuracy of miR319a processing. Thus, full-length precursors are maintained as backbones for constructing engineered amiRNA vectors. In our case, the mutated sequences of the engineered pre-miR319a and pre-miR528 are located in the unpaired ssRNA region (Figures 1 and 5), which may be less important for miRNA processing [25]. Indeed, the modifications on the pre-miR319a and pre-miR528 have little influence on their processing and functionality (Figures 2–4 and 6). It has been reported that miRNAs repress target gene expression through two modes of action, mRNA cleavage and translation inhibition [51]. However, the evaluation of efficacy of amiRNA in many studies is largely based on the measurement of target mRNA [45–48], without checking the abundance of target proteins. Using the ETPamir assay, we proved that amiRNAs produced by engineered pre-miR319a and pre-miR528 could effectively block the target protein accumulation (Figures 3 and 6). The high efficacy of amiRNA expressed from the engineered pre-miR319a was further confirmed by the dwarf phenotype of amiR-*CNGC4*-1 overexpression lines (Figure 4). Taken together, we reason that the engineered amiRNA backbones should be fully functional as their original precursors.

In comparison to other amiRNA construction methods, our approach offers two advantages, as follows: First, full-length precursors are used as backbones and no changes are made in the lower stem, allowing amiRNAs to be processed accurately. Although some simplified amiRNA construction methods have been previously reported, those studies used the precursors either without the lower stem [47,52,53] or with a mutated lower stem [46,54,55] to express amiRNAs. Although the precursors in those methods could successfully produce amiRNAs and suppress target gene expression, at least in some cases there is no convincing evidence to prove that these truncated or mutated precursors are functioning equally like the full-length natural precursors. Second, we used the conventional restriction digestion-ligation strategy to construct amiRNA vectors, which balances the cloning efficiency and cost. Compared with the Gateway cloning system [56] and TA-based cloning system [52,54], whose cloning efficiency largely relies on commercial cloning kits or relatively expensive enzymes, our method is apparently more cost-saving.

In conclusion, we explored a simple and efficient method to construct amiRNA expression cassettes by creating restriction sites within the basal region of Arabidopsis and rice amiRNA precursors. We demonstrated that these modified amiRNA precursors are fully functional in plant protoplasts and transgenic plants. Hopefully, this new amiRNA cloning strategy will be useful for genome research in dicot and monocot plant species.

4. Materials and Methods

4.1. Plant Growth

Wild-type Col-0 or transgenic *Arabidopsis thaliana* plants were grown in a plant growth room on moistened Jiffy soil (Jiffy Substrates ®, Jiffy Group, Pärnumaa, Estonia), which are high-quality sphagnum peat-based growing substrates with a high organic content and water capacity to encourage rapid rooting and uniform growth. The Arabidopsis growth conditions are fixed at 65% humidity and 75 μmol·m^{-2}·s^{-1} light intensity under photoperiods of 12 h light at 23 °C and 12 h dark at 20 °C. Zhonghua 11 rice (*Oryza sativa*) plants were grown on Jiffy soil in a plant growth chamber under photoperiods of 12 h light (200 μmol·m^{-2}·s^{-1}) at 30 °C and 12 h dark at 27 °C, with a constant humidity of 70%.

4.2. Plasmid Construction

Routine molecular cloning procedures were followed for plasmid construction. The original sequences of Arabidopsis miR319a or rice miR528 backbones were mutagenized by PCR-based mutagenesis to generate engineered miRNA precursors (pre-amiRNA). The amiRNA vectors *HBT-amiR-MEKK1*, *HBT-amiR-PDS3*, and *HBT-amiR-MAPKKK3* were constructed using pre-miR319a as the backbone, while *HBT-amiR-GFP* and *HBT-amiR-OsCEBiP* were constructed using pre-miR528 as the backbone. The original amiRNA vectors were cloned by traditional overlapping PCR, described by Schwab et al. [17]. The engineered amiRNA vectors were constructed as follows. Briefly, mega-primers containing customized amiRNA/amiRNA* sequences were used for PCR amplification of primary miRNA fragment containing a new amiRNA/amiRNA* duplex using the original pre-miR319a or pre-miR528 as PCR template. PCR amplicons were digested by *Eco*RI/*Xba*I or *Stu*I/*Eco*RI and inserted into the same digested HBT vector harboring the engineered pre-miR319a or pre-miR528. For plant transformation, the pre-amiRNA was digested by *Bam*HI/*Pst*I and inserted into the same digested pCB302 binary vector.

To express a target gene encoding double HA- or FLAG-tagged target protein in protoplasts, the full-length coding sequences of target genes were amplified by RT-PCR, digested by *Bam*HI/*Stu*I and inserted into the same digested *HBT-2HA* or *HBT-2FLAG* vector, where the target gene expression is driven by the *35S* promoter.

4.3. Protoplast Isolation

Four-week-old Arabidopsis or 10-day-old *Oryza sativa* (Zhonghua 11 rice) seedlings were used for protoplast isolation according to the procedure described previously [32,57]. Briefly, leaves of Arabidopsis or sheaths of rice were cut into 0.5-mm strips with a sterile razor blade. The strips were digested in 10 mL enzyme solution (1.5% cellulase R10, 0.2% macerozyme R10, 0.4 M mannitol, 20 mM KCl, 20 mM MES, pH 5.7, 10 mM $CaCl_2$, and 0.1% BSA) at room temperature for 3 h under a dark condition. After mixing with 10 mL W5 solution (154 mM NaCl, 125 mM $CaCl_2$, 5 mM KCl, and 2 mM MES, pH 5.7), the digestion mixture was filtered through a 75 µm FALCON cell strainer. Protoplasts were collected by centrifugation in a CL2 clinical centrifuge (Thermo Scientific, Weaverville, North Carolina, USA) for 2 min at 100× *g* for Arabidopsis or 5 min at 200× *g* for rice. Cells were resuspended with 10 mL W5 solution and rested on ice for 30 min. Before transfection, protoplasts were pelleted by centrifugation for 1 min at 100× *g* for Arabidopsis or 3 min at 200× *g* for rice, and were then resuspended with MMg solution (0.4 M mannitol, 15 mM $MgCl_2$, and 4 mM MES, pH 5.7) to a final concentration of 2×10^5 cells per ml.

4.4. Protoplast Transfection and ETPamir Assay

DNA transfection was performed in a 2-mL round-bottom microcentrifuge tube, where 200 µL protoplasts were mixed with 21 µL (2 µg/µL) DNA cocktail and 220 µL PEG solution (40% PEG4000, *v*/*v*, 0.2 M mannitol and 0.1 M $CaCl_2$), gently. After incubated at room temperature for 5 min (light) for Arabidopsis protoplasts or 15 min (dark) for rice protoplasts, transfection was quenched by adding 880 µL W5 solution. Transfected protoplasts were collected by centrifugation for 2 min at 100× *g* for Arabidopsis or 5 min at 200× *g* for rice, and were resuspended with 100 µL W5 solution. The cells were then transferred into 1 mL WI solution (0.5 M mannitol, 4 mM MES, pH 5.7, and 20 mM KCl) in a 6-well plate and were incubated in the dark.

The ETPamir assay was conducted according to the method described previously [31,32]. Briefly, 200 µL protoplasts were transfected with a DNA cocktail (2 µg/µL) containing 16 µL amiRNA expression construct, 4 µL target gene-HA/FLAG expression construct, and 1 µL transfection control plasmid expressing GFP-HA/FLAG. In parallel, a negative control was set up by replacing the amiRNA expression construct with an equal amount of empty vector. After co-transfection, protoplasts were incubated for 18–36 h in dark and then were collected for western blot analysis. The amiRNA performance was inversely correlated with the target protein accumulation.

4.5. Western Blot

After centrifugation, protoplasts were directly lysed with the lysis buffer (10 mM HEPES, pH 7.5, 100 mM NaCl, 1 mM EDTA, 10% Glycerol). The lysates were mixed with 6 × SDS-PAGE loading buffer and heated at 95 °C for 5 min or 55 °C for 10 min. Total proteins were subjected to SDS-PAGE (10%) and immunoblotting with anti-HA (Roche) or anti-FLAG (Sigma-Aldrich, Saint Louis, Missouri, USA) antibodies.

4.6. Generation and Screen of Transgenic Plants

The recombinant pCB302 binary plasmids were introduced into *Agrobacterium tumefaciens GV3101* cells by electroporation, which were in turn used for floral dip-mediated Arabidopsis transformation [58]. Transgenic Arabidopsis plants were selected on 1/2 MS medium containing 12.5 mg/L glufosinate ammonium.

4.7. RNA Extraction and Mature amiRNA Detection

For mature amiRNA detection in protoplasts, a total of 400 µL Arabidopsis protoplasts co-transfected with *HBT-pre-amiRNA* (original or engineered) plasmid and *pAN-mCherry-HA* plasmid were used for RNA extraction. For mature miR319a detection in transgenic plants, 30 mg rosette leaves of Col-0 and pre-miR319a overexpression lines were used for RNA extraction. Total RNA was extracted using the RNAiso Plus reagent (TaKaRa) according to the manufacturer's instructions. The protocol described earlier [30] with minor modifications was used for mature amiRNA detection. Briefly, 1 µg total RNA was converted into the first-strand cDNA with stem-loop RT primers for amiRNA and gene specific primer of *mCherry* using a PrimeScript™ RT reagent Kit with genomic DNA Eraser (TaKaRa) according to the manufacturer's instructions. RT-qPCR was performed in a LightCycler 96 Instrument (Roche) using the SYBR® *Premix Ex Taq*™ Kit (TaKaRa). Accumulation of mature amiRNAs produced from the original or engineered precursor in Arabidopsis protoplasts or pre-miR319a overexpression transgenic plants were normalized to the transcript levels of the transfection control *mCherry-HA* or *snoR101* (Small Nucleolar RNA 101), respectively.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/20/22/5620/s1.

Author Contributions: D.Z. designed and constructed modified amiRNA vectors and performed preliminary tests. N.Z. and W.S. conducted the majority of experiments to evaluate the engineered amiRNA vectors in protoplasts and transgenic plants. W.S., D.Z., and J.-F.L. wrote the manuscript. J.-F.L. supervised the study.

Funding: This work was supported by the Foundation of Guangzhou Science and Technology Key Project (No. 201904020041).

Acknowledgments: We thank Jen Sheen for the support of the initiation of this research.

Conflicts of Interest: No conflict of interest declared.

Abbreviations

AS	Alternative splicing
amiRNA	Artificial microRNA
CRISPR	Clustered regularly interspaced short palindromic repeats
Cas9	CRISPR-associated protein 9
miRNA	MicroRNA
pre-miRNA	MiRNA precursor
pri-miRNA	Primary miRNA
T-DNA	Transfer DNA
PCR	Polymerase chain reaction

References

1. Sundaesan, V.; Springer, P.; Volpe, T.; Haward, S.; Jones, J.D.G.; Dean, C.; Ma, H.; Martienssen, R. Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes Dev.* **1995**, *9*, 1797–1810. [CrossRef] [PubMed]
2. Alonso, J.M.; Stepanova, A.N.; Leisse, T.J.; Kim, C.J.; Chen, H.; Shinn, P. Genome wide insertional mutagenesis of *Arabdiopsis thaliana*. *Science* **2003**, *301*, 653–657. [CrossRef] [PubMed]
3. Kim, Y.; Schumarker, K.S.; Zhu, J.K. EMS mutagenesis of Arabidopsis. *Methods Mol. Biol.* **2006**, *323*, 101–103. [PubMed]
4. Young, J.C.; Krysan, P.J.; Sussman, M.R. Efficient screening of Arabidopsis T-DNA insertion lines using degenerate primers. *Plant Physiol.* **2001**, *125*, 513–518. [CrossRef]
5. Bortesi, L.; Fischer, R. The CRISPR/Cas9 system for plant genome editing and beyond. *Biotechnol. Adv.* **2015**, *33*, 41–52. [CrossRef]

6. Wang, H.F.; Russa, M.L.; Qi, L.S. CRISPR/Cas9 in genome editing and beyond. *Annu. Rev. Biochem.* **2016**, *85*, 227–264. [CrossRef]
7. Dai, X.; Zhang, Y.; Zhang, D.; Chen, J.; Gao, X.; Estelle, M.; Zhao, Y. Embryonic lethality of *Arabidopsis abp1-1* is caused by deletion of the adjacent *BSM* gene. *Nat. Plants* **2015**, *1*, 15183. [CrossRef]
8. Budzisewski, G.J.; Lewis, S.P.; Glvoer, L.W.; Reineke, J.; Jones, G.; Ziemnik, L.S. Arabidopsis genes essential for seedling viability: Isolation of insertional mutants and molecular cloning. *Genetics* **2001**, *159*, 1765–1778.
9. Syed, N.H.; Kalyna, M.; Marquez, Y.; Barta, A.; Brown, J.W.S. Alternative splicing in plants-coming of age. *Trends Plant Sci.* **2012**, *17*, 616–623. [CrossRef]
10. Reinhart, B.J.; Weinstein, E.G.; Rhoades, M.W.; Bartel, B.; Bartel, D.P. MicroRNAs in plants. *Genes Dev.* **2002**, *16*, 1616–1626. [CrossRef]
11. Song, X.; Li, Y.; Gao, X.; Qi, Y. MicroRNAs and their regulatory roles in plant-environment interactions. *Annu. Rev. Plant Biol.* **2019**, *70*, 489–525. [CrossRef] [PubMed]
12. Kim, Y.J.; Zheng, B.; Yu, Y.; Won, S.Y.; Mo, B.; Chen, X. The role of mediator in small and long noncoding RNA production in *Arabidopsis thaliana*. *EMBO J.* **2011**, *30*, 814–822. [CrossRef] [PubMed]
13. Fang, X.; Cui, Y.; Li, Y.; Qi, Y. Transcription and processing of primary microRNAs are coupled by elongator complex in *Arabidopsis*. *Nat. Plants* **2015**, *1*, 15075. [CrossRef] [PubMed]
14. Fang, Y.; Spector, D.L. Identification of nuclear dicing bodies containing proteins for microRNA biogenesis in living *Arabidopsis* plants. *Curr. Biol.* **2007**, *17*, 818–823. [CrossRef]
15. Manavella, P.A.; Koeing, D.; Weigel, D. Plant secondary siRNA production determined by microRNA-duplex structure. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 2461–2466. [CrossRef]
16. Ossowski, S.; Schwab, R.; Weigel, D. Gene silencing in plants using artificial microRNAs and other small RNAs. *Plant J.* **2008**, *53*, 674–690. [CrossRef]
17. Schwab, R.; Ossowski, S.; Riester, M.; Warthmann, N.; Weigel, D. Highly specific gene silencing by artificial microRNAs in *Arabidopsis*. *Plant Cell* **2006**, *18*, 1121–1133. [CrossRef]
18. Zhang, Y.; Nasser, V.; Pisanty, O.; Omary, M.; Wulff, N.; Donato, M.D. A transprotome-scale amiRNA-based screen identifies redundant roles of *Arabidopsis* ABCB6 and ABCB20 in auxin transport. *Nat. Commun.* **2018**, *9*, 4204. [CrossRef]
19. Zhang, Z.; Guo, X.; Ge, C.; Ma, Z.; Jiang, M.; Li, T. KETCH1 imports HYL1 to nucleus for miRNA biogenesis in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 4011–4016. [CrossRef]
20. Brennecke, J.; Stark, A.; Russell, R.B.; Cohen, S.M. Principles of microRNA-target recognition. *PLoS Biol.* **2005**, *3*, e85. [CrossRef]
21. Hu, T.; Fu, Q.; Chen, P.; Ma, L.; Sin, O.; Guo, D. Construction of an artificial microRNA expression vector for simultaneous inhibition of multiple genes in mammalian cells. *Int. J. Mol. Sci.* **2009**, *10*, 2158–2168. [CrossRef] [PubMed]
22. Mateos, J.L.; Bologna, N.G.; Chorostecki, U.; Palatnik, J.F. Identification of microRNA processing determinants by random mutagenesis of *Arabidopsis* MIR172a precursor. *Curr. Biol.* **2010**, *20*, 49–54. [CrossRef] [PubMed]
23. Song, L.; Axtell, M.J.; Fedoroff, N.V. RNA secondary structural determinants of miRNA precursor processing in *Arabidopsis*. *Curr. Biol.* **2010**, *20*, 37–41. [CrossRef]
24. Werner, S.; Wollmann, H.; Schneeberger, K.; Weigel, D. Structure determinants for accurate processing of miR172a in *Arabidopsis thaliana*. *Curr. Biol.* **2010**, *20*, 42–48. [CrossRef] [PubMed]
25. Bologna, N.G.; Mateos, J.L.; Bresso, E.G.; Palatnik, J.F. A loop-to-base processing mechanism underlies the biogenesis of plant microRNAs miR319 and miR159. *EMBO J.* **2009**, *28*, 3646–3656. [CrossRef] [PubMed]
26. Palatnik, J.F.; Allen, E.; Wu, X.; Schommer, C.; Schwab, R.; Carrington, J.C.; Weigel, D. Control of leaf morphogenesis by microRNAs. *Nature* **2003**, *425*, 257–263. [CrossRef] [PubMed]
27. Palatnik, J.F.; Wollmann, H.; Schommer, C.; Schwab, R.; Boisbouvier, J.; Rodriguez, R. Sequence and expression differences underlie functional specialization of *Arabidopsis* microRNAs miR159 and miR319. *Dev. Cell* **2007**, 115–125. [CrossRef]
28. Nag, A.; King, S.; Jack, T. miR319a targeting of TCP4 is critical for petal growth and development in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 22534–22539. [CrossRef]
29. Koyama, T.; Sato, F.; Ohme-Takagi, M. Roles of miR319 and TCP transcription factors in leaf development. *Plant Physiol.* **2017**, *175*, 874–885. [CrossRef]
30. Varkonyi-Gasic, E.; Wu, R.; Wood, M.; Walton, E.F.; Hellens, R.P. Protocol: A highly sensitive RT-PCR method for detection and quantification of microRNAs. *Plant Methods* **2007**, *3*, 12. [CrossRef]

31. Li, J.F.; Chung, H.S.; Niu, Y.; Bush, J.; McCormack, M.; Sheen, J. Comprehensive protein-based artificial microRNA screens for effective gene silencing in plants. *Plant Cell* **2013**, *25*, 1507–1522. [CrossRef] [PubMed]
32. Li, J.F.; Zhang, D.; Sheen, J. Epitope-tagged protein-based artificial miRNA screens for optimized gene silencing in plants. *Nat. Protoc.* **2014**, *9*, 939–949. [CrossRef] [PubMed]
33. Balagué, C.; Lin, B.; Alcon, C.; Flottes, G.; Malmström, S.; Köhler, C. HLM1, an essential signaling component in the hypersensitive response, is a member of the cyclic nucleotide-gated channel ion channel family. *Plant Cell* **2003**, *15*, 365–379. [CrossRef] [PubMed]
34. Tian, W.; Hou, C.; Ren, Z.; Wang, C.; Zhao, F.; Dahlbeck, D. A calmodulin-gated calcium channel links pathogen patterns to plant immunity. *Nature* **2019**, *572*, 131–135. [CrossRef] [PubMed]
35. Warthmann, N.; Chen, H.; Ossowski, S.; Weigel, D.; Hervé, P. Highly specific gene silencing by artificial miRNAs in rice. *PLoS ONE* **2008**, *3*, e1829. [CrossRef]
36. Petchthai, U.; Yee, C.S.L.; Wong, S. Resistance to CymMV and ORSV in artificial microRNA transgenic *Nicotiana benthamiana* plants. *Sci. Rep.* **2018**, *8*, 9958. [CrossRef]
37. Adkar-Purushothama, C.R.; Perreault, J.P. Alterations of the viroid regions that interact with the host defense genes attenuate viroid infection in host plant. *RNA Biol.* **2018**, *15*, 955–966. [CrossRef]
38. Butardo, V.M.; Fitzgerald, M.A.; Bird, A.R.; Gidley, M.J.; Flanagan, B.M.; Larroque, O. Impact of down-regulation of starch branching enzyme IIb in rice by artificial microRNA- and hairpin RNA-mediated RNA silencing. *J. Exp. Bot.* **2011**, *62*, 4927–4941. [CrossRef]
39. Chi, M.; Bhagwat, B.; Lane, W.D.; Tang, G.; Su, Y.; Sun, R. Reduced polyphenol oxidase gene expression and enzymatic browning in potato (*Solanum tuberosum L.*) with artificial microRNAs. *BMC Plant Biol.* **2014**, *14*, 62. [CrossRef]
40. Niu, Q.W.; Lin, S.S.; Reyes, J.L.; Chen, K.C.; Wu, H.W.; Yeh, S.D.; Chua, N.H. Expression of artificial microRNAs in transgenic *Arabidopsis thaliana* confers virus resistance. *Nat. Biotechnol.* **2006**, *24*, 1420–1428. [CrossRef]
41. Kis, A.; Tholt, G.; Ivanics, M.; Varallyay, E.; Jenes, B.; Havelda, Z. Polycistronic artificial miRNA-mediated resistance to wheat dwarf virus in barley is highly efficient at low temperature. *Mol. Plant Pathol.* **2016**, *17*, 427–437. [CrossRef] [PubMed]
42. Zhang, T.; Zhao, Y.L.; Zhao, J.H.; Wang, S.; Jin, Y.; Chen, Z.Q.; Fang, Y.Y.; Hua, C.L.; Ding, S.W.; Guo, H.S. Cotton plants export microRNAs to inhibit virulence gene expression in a fungal pathogen. *Nat. Plants.* **2016**, *2*, 16153. [CrossRef] [PubMed]
43. Wang, M.; Weiberg, A.; Lin, F.M.; Thomma, B.P.; Huang, H.D.; Jin, H. Bidirectional cross-kingdom RNAi and fungal uptake of external RNAs confer plant protection. *Nat. Plants* **2016**, *2*, 16151. [CrossRef] [PubMed]
44. Tian, B.; Li, J.; Oakley, T.R.; Todd, T.C.; Trick, H.N. Host-derived artificial microRNA as an alternative method to improve soybean resistance to Soybean Cyst Nematode. *Genes* **2016**, *7*, e122. [CrossRef]
45. Guo, H.; Song, X.; Wang, G.; Yang, K.; Wang, K.; Niu, L. Plant-generated artificial small RNAs mediated aphid resistance. *PLoS ONE* **2014**, *9*, e97410. [CrossRef]
46. Liang, G.; He, H.; Li, Y.; Yu, D. A new strategy for construction of artificial miRNA vectors in *Arabidopsis*. *Planta* **2012**, *235*, 1421–1429. [CrossRef]
47. Carbonell, A.; Takeda, A.; Fahlgren, N.; Johnson, S.C.; Cuperus, J.T.; Carrington, J.C. New generation of artificial microRNA and synthetic trans-acting small interfering RNA vectors for efficient gene silencing in *Arabidopsi*s. *Plant Physiol.* **2014**, *165*, 15–29. [CrossRef]
48. Carbonell, A.; Fahlgren, N.; Mitchell, S.; Cox Jr, K.L.; Reilly, K.C. Highly specific gene silencing in a monocot species by artificial microRNAs derived from chimeric miRNA precursors. *Plant J.* **2015**, *82*, 1061–1075. [CrossRef]
49. Yuan, S.; Li, Z.; Li, D.; Yuan, N.; Hu, Q.; Luo, H. Constitutive expression of rice MicroRNA528 alters plant development and enhances tolerance to salinity stress and nitrogen starvation in *Creeping Bentgrass*. *Plant Physiol.* **2015**, *169*, 576–593. [CrossRef]
50. Axtell, M.J.; Snyder, J.A.; Bartel, D.P. Common functions for diverse small RNAs of land plants. *Plant Cell* **2007**, *19*, 1750–1769. [CrossRef]
51. Li, S.; Liu, L.; Zhuang, X.; Yu, Y.; Liu, X.; Cui, X. microRNAs inhibit the translation of target mRNAs on the endoplasmic reticulum in *Arabidopsis*. *Cell* **2013**, *153*, 562–574. [CrossRef] [PubMed]
52. Wang, C.; Yin, X.; Kong, X.; Li, W.; Ma, L.; Sun, X. A series of TA-based and zero-background vectors for plant functional genomic. *PLoS ONE* **2013**, *8*, e59576. [CrossRef] [PubMed]

53. Ju, Z.; Cao, D.; Cao, C.; Zuo, J.; Zhai, B. A viral satellite DNA vector (TYLCCNV) for functional analysis of miRNAs and siRNAs in plants. *Plant Physiol.* **2017**, *173*, 1940–1952. [CrossRef] [PubMed]
54. Chen, S.; Songkumarn, P.; Liu, J.; Wang, G. A versatile zero background T-vector system for gene cloning and functional genomics. *Plant Physiol.* **2009**, *150*, 1111–1121. [CrossRef] [PubMed]
55. Liu, C.; Zhang, L.; Sun, J.; Luo, Y.; Wang, M.; Fan, Y.; Wang, L. A simple artificial microRNA vector based on ath-miR169d precursor from *Arabidopsis*. *Mol. Biol. Rep.* **2010**, *37*, 903–909. [CrossRef] [PubMed]
56. Karimi, M.; Depicker, A.; Hilson, P. Recombinational cloning with plant Gateway vectors. *Plant Physiol.* **2007**, *145*, 1144–1154. [CrossRef] [PubMed]
57. Zhang, Y.; Su, J.; Duan, S.; Ao, Y.; Dai, J.; Liu, J.; Wang, P.; Li, Y.; Liu, B.; Feng, D. A highly efficient rice green tissue protoplast system for transient gene expression and studying light/chloroplast-related processes. *Plant Methods* **2011**, *7*, 30. [CrossRef]
58. Clough, S.J.; Bent, A.F. Floral dip: A simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **1998**, 735–743. [CrossRef]

International Journal of
Molecular Sciences

Article

Comparatively Barcoded Chromosomes of *Brachypodium* Perennials Tell the Story of Their Karyotype Structure and Evolution

Joanna Lusinska [1], Alexander Betekhtin [1], Diana Lopez-Alvarez [2,3], Pilar Catalan [3,4,5], Glyn Jenkins [6], Elzbieta Wolny [1,*] and Robert Hasterok [1,*]

[1] Institute of Biology, Biotechnology and Environmental Protection, Faculty of Natural Sciences, University of Silesia in Katowice, 40-032 Katowice, Poland; jlusinska@us.edu.pl (J.L.); alexander.betekhtin@us.edu.pl (A.B.)
[2] Faculty of Agricultural Sciences, National University of Columbia, Palmira 763533, Colombia; dianalopez430@gmail.com
[3] Department of Agriculture (Botany), High Polytechnic School of Huesca, University of Zaragoza, 22071 Huesca, Spain; pilar.catalan09@gmail.com
[4] Grupo de Bioquímica, Biofísica y Biología Computacional (BIFI, UNIZAR), Unidad Asociada al CSIC, 50018 Zaragoza, Spain
[5] Institute of Biology, Tomsk State University, Tomsk 634050, Russia
[6] Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth SY23 3DA, UK; gmj@aber.ac.uk
[*] Correspondence: elzbieta.wolny@us.edu.pl (E.W.); robert.hasterok@us.edu.pl (R.H.)

Received: 28 September 2019; Accepted: 2 November 2019; Published: 7 November 2019

Abstract: The *Brachypodium* genus is an informative model system for studying grass karyotype organization. Previous studies of a limited number of species and reference chromosomes have not provided a comprehensive picture of the enigmatic phylogenetic relationships in the genus. Comparative chromosome barcoding, which enables the reconstruction of the evolutionary history of individual chromosomes and their segments, allowed us to infer the relationships between putative ancestral karyotypes of extinct species and extant karyotypes of current species. We used over 80 chromosome-specific BAC (bacterial artificial chromosome) clones derived from five reference chromosomes of *B. distachyon* as probes against the karyotypes of twelve accessions representing five diploid and polyploid *Brachypodium* perennials. The results showed that descending dysploidy is common in *Brachypodium* and occurs primarily via nested chromosome fusions. *Brachypodium distachyon* was rejected as a putative ancestor for allotetraploid perennials and *B. stacei* for *B. mexicanum.* We propose two alternative models of perennial polyploid evolution involving either the incorporation of a putative $x = 5$ ancestral karyotype with different descending dysploidy patterns compared to *B. distachyon* chromosomes or hybridization of two $x = 9$ ancestors followed by genome doubling and descending dysploidy. Details of the karyotype structure and evolution in several *Brachypodium* perennials are revealed for the first time.

Keywords: *Brachypodium*; comparative chromosome barcoding; dysploidy; karyotype structure and evolution; model grass genus; molecular cytogenetics; polyploidy

1. Introduction

In recent years, the genus *Brachypodium* has become one of the most comprehensively studied genera among monocotyledonous plants primarily due to the exploitation of one of its flagship species, *B. distachyon*, as a functional model organism for temperate cereals and other economically important grasses [1–5]. It comprises three annuals, the diploids *B. distachyon* ($2n = 10$) and *B. stacei* ($2n = 20$)

and their derived allotetraploid *B. hybridum* (2*n* = 30), which have recently been proposed as a model system to study polyploidy and grass speciation [6,7]. Approximately 15 other representatives are perennials distributed worldwide [5,8]. All *Brachypodium* species have small and compact nuclear genomes, diverse (*x* = 5, 8, 9, 10) basic chromosome numbers and various ploidy levels [9–12], which are accompanied by complicated and still not fully resolved phylogenetic relationships [8,13,14]. Recent phylogenetic studies that were based mainly on combined analyses of some nuclear and plastid genes dated the origin and split of the crown *Brachypodium* ancestor in the Mid-Miocene (12.6 Ma) [6,8,13,15]. *Brachypodium* perennials are very diverse in terms of their phenotype, origin, and geographical distribution. They range from the American short-rhizomatous *B. mexicanum* (2*n* = 40), which resembles the annual more than the perennial taxa [8,16], to the more recently evolved Eurasian and African long-rhizomatous diploid and allopolyploid species of the core-perennial clade, i.e., *B. arbuscula* (2*n* = 18), *B. boissieri* (2*n* = 42, 46), *B. glaucovirens* (2*n* = 16), *B. phoenicoides* (2*n* = 28), *B. pinnatum* (2*n* = 16, 18, 28), *B. retusum* (2n = 36, 38), *B. rupestre* (2*n* = 18, 28), and *B. sylvaticum* (2*n* = 18) [3,6,8,10,17]. One of the most widespread and best studied perennial species is *B. sylvaticum* which has considerable readily available genetic resources such as inbred lines, efficient transformation protocols, and genomic and transcriptomic tools. Because of these, it has been recently proposed as a new model plant to study perenniality [3,7,18,19]. In some earlier reports, *B. sylvaticum* was used to assist the molecular characterization of the *Ph1* locus in wheat [20] and to compare gene conservation and collinearity with orthologous regions from rice and wheat [21]. Given the economic importance of perennial grasses, comparative studies of more closely related *Brachypodium* annuals and perennials can also be of particular importance in identifying and testing candidate "perenniality" genes and creating a tractable model system for both fundamental research and crop improvement [7,18].

Synteny-based paleogenomics implies that the present-day karyotypes originated from ancestral genomes with the lowest number of historical polyploidization events [22,23]. This offers insight into the putative numbers of protochromosomes of the respective progenitors of the current species and provides an opportunity to link the karyotypes of extant species, including that of *B. distachyon*, with those of their hypothetical and extinct ancestors [4,24,25]. Comparative genomics identifies polyploidization and dysploidy events, which are complemented by minor genome rearrangements, as crucial factors in the evolution, divergence, and adaptive speciation of all flowering plants [26,27]. Despite its great importance, our ability to understand polyploid genome evolution, including that of economically important crops, is still constrained by a limited knowledge of the actual parents and incomplete lineage reconstruction during polyploid speciation [7,28]. Within the complexity of many plant species, polyploid series are often described either as intraspecific cytotypes showing different ploidy levels and very similar morphological features [29,30] or are considered different species [31]. For example, *B. distachyon* was initially described as a species with three cytotypes comprising an *x* = 5 basic chromosome number [11]. However, the seminal cytogenetic analyses of Hasterok et al. [32,33], coupled with later molecular studies [34] and a comprehensive taxonomic description and phylogenetic analysis [6], drove their reclassification into three separate species (*B. distachyon*, *B. stacei*, and *B. hybridum*). Moreover, fluorescence in situ hybridization (FISH)-based studies, including comparative chromosome painting (CCP) and comparative chromosome barcoding (CCB), concluded that various *Brachypodium* allopolyploids were derived from interspecific crosses of distinct diploid, perennial, and/or annual progenitors [9,12,35]. Although the exact taxonomic identity of the diploid (2*n* = 16, 18) and allotetraploid (2*n* = 28) cytotypes of *B. pinnatum* and *B. rupestre* remains unclear, there is a growing body of evidence [9,12,13] that they should be classified as separate species, thereby paralleling the case of the diploid–allotetraploid *B. distachyon* complex [6,32].

The availability of the *B. distachyon* whole-genome sequence [4] (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Bdistachyon) combined with FISH using low-repeat BAC (bacterial artificial chromosomes) clones as probes [36–38] enabled the precise dissection of the chromosome structure at the microscopic level via selective visualization of either their smaller (via CCB) or larger (including entire chromosomes: via CCP) regions. Apart from *Brachypodium*, similar approaches are limited in plants to a handful of small-genome taxa within Brassicaceae [39,40], *Cucumis* [41], rice [42], and, recently, also in maize [43]. To date, detailed analyses of the karyotype structure and evolution using CCB in *Brachypodium* essentially targeted the annuals [44]. The relatively few studies on perennials have been constrained in the past by a paucity of chromosome markers and unavailability of germplasm [14,35,45]. We address these technical obstacles here and present a comprehensive model of the karyotype structure and evolution of perennial *Brachypodium* species.

2. Results

The karyotypes of both diploid and allopolyploid *Brachypodium* perennials (Table 1) were compared using the CCB mapping approach and with reference to *B. distachyon*. The use of 86 low-repeat BAC clones as the FISH probes for *B. mexicanum* and 59 clones for other *Brachypodium* perennial species enabled us to precisely track and analyze the evolutionary rearrangements of individual chromosomes and, consequently, entire karyotypes. We used differentially labelled, overlapping triplets of single-locus BACs at contiguous positions on the physical map of a particular chromosome of *B. distachyon* (Table S1). To clarify the relationships among the mapped chromosomal regions of the *Brachypodium* karyotypes, additional FISH experiments were performed with specific non-adjacent pairs of single-locus BAC-based probes and a centromeric BAC BD_CBa0033J12 (CEN). The results of the CCB were analyzed with reference to the so-called Bd-genome of *B. distachyon* and published genomic data from the whole-genome comparison of *B. distachyon* and rice [4]. Cytogenetic maps of the chromosomes were constructed based upon the results of the cross-species chromosome mapping (Figures 1–4). We adopted the nomenclature for the chromosomes of the *Brachypodium* perennials according to their alignment with CoGe (https://genomevolution.org/coge/SynMap.pl), which is based on sequencing data for *B. sylvaticum* (https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=Bsylvaticum) with reference to their assignment to *B. distachyon* (https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=BdistachyonBd21_3). For consistency with the Bd and Bs genome designations that were assigned to the annuals *B. distachyon* and *B. stacei* [44], we used the Bp and Bm designations for the *Brachypodium* perennial genomes and *B. mexicanum* genomes, respectively. The chromosomes of the diploid *Brachypodium* perennials and *B. mexicanum* with the same or similar composition of mapped Bd genome-derived BAC clones were given the same chromosome numbers.

Table 1. General characteristics of the *Brachypodium* species that were used in this study.

Species	Accession Number	2*n*	*x*	Ploidy Level	Origin	Source *
B. distachyon	Bd21	10	5	2×	Iraq	USA
B. sylvaticum	PI 297868	18	9	2×	Australia	USA
	PI 269842	18	9	2×	Tunisia	USA
B. glaucovirens	PI 4202	16	8	2×	Greece, Crete	Germany
B. pinnatum	PI 185135	16	8	2×	Iraq	USA
	PI 230113	18	9	2×	Iran	USA
	PI 345982	18	9	2×	Norway	USA
	PI 249722	28	5 + 9	4×	Greece	USA
	PI 251445	28	5 + 9	4×	Turkey	USA
	PI 430277	28	5 + 9	4×	Ireland	USA
B. phoenicoides	PI 253503	28	5 + 9	4×	Spain	USA
	PI 89817	28	5 + 9	4×	Spain	USA
B. mexicanum	Bmex347	40	10 + 10	4×	Mexico	UK

* USA: United States Department of Agriculture—National Plant Germplasm System, Beltsville, MD; Germany: Botanical Garden Berlin-Dahlem; UK: University of Leicester, Leicester (from Clive A. Stace); PI: Plant introduction.

2.1. Comparative Mapping of the Chromosomes in the Perennial Diploids (2n = 18, 16 Chromosomes)

The arrangement of all the BACs mapped to the chromosomes of the diploids *B. sylvaticum* and *B. pinnatum* (both 2*n* = 18, *x* = 9) is shown in detail on a cytogenetic map (Figure 1). We observed no intraspecific differences in the pattern of clones between the *B. sylvaticum* genotypes PI 297868 and PI

269842 and *B. pinnatum* PI 230113 and PI 345982. Each of the clones hybridized to a single chromosome pair in both species, and their order and arm specificity were the same in the perennial diploid chromosomes and were consistent with their counterparts in the respective reference chromosomes of *B. distachyon*. The clones derived from chromosome Bd1 of *B. distachyon* consistently hybridized to three different chromosomes of *B. sylvaticum* and *B. pinnatum* (2*n* = 18) which were identified as Bp2, Bp6, and Bp7 (Figure 1 and Supplementary Materials Figure S1). The BACs Bd1S/1–5 and Bd1L/24–28 from the distal parts of both chromosome arms of Bd1 mapped consecutively along the entire chromosome Bp2. Clones Bd1S/7–10 and Bd1L/21–23 from the interstitial parts of Bd1 localized along the short and long arm of chromosome Bp6, respectively. The set of BACs Bd1S/11-Bd1L/19 from the central part of Bd1 localized to chromosome Bp7. It is known that chromosome Bd1 of *B. distachyon* arose from two separate nested chromosome fusions (NCF) of three ancestral chromosomes, which are equivalent to the ancestral Os3, Os7, and Os6 "rice-like" chromosomes [4]. Thus, these three ancestral rice chromosome equivalents (ARCEs) correspond to the entire Bp2, Bp6, and Bp7 chromosomes of the diploid *Brachypodium* perennials, respectively (Figure 1).

In the same species, the Bd2-derived clones hybridized to two different chromosomes identified as Bp1 and Bp8 (Figure 1 and Figure S3). The BAC clones Bd2S/2–6 and Bd2L/14–19, from both of the chromosome arms of Bd2 that corresponded to the Os1 ARCE, had an undisrupted linear arrangement along Bp1. The BACs Bd2S/8 to Bd2L/13 from the central part of chromosome Bd2 were localized on chromosome Bp8, which is the Os5 ARCE (Figure 1). These results demonstrate that in the karyotypes of diploid *Brachypodium sylvaticum* and *B. pinnatum*, the homoeologues of Bd2 are represented by two distinct chromosomes, which are equivalent to Os1 and Os5. Comparative mapping with the Bd3-derived BACs revealed two homoeologues, Bp4 and Bp3 (Figure 1 and Figure S4). The BACs Bd3S/1–3 and Bd3L/14–18, from the distal parts of a chromosome of Bd3, were mapped on chromosome Bp4, while the sets Bd3S/4–7 and Bd3L/9–12 from the proximal part of Bd3 localized along chromosome Bp3. Chromosome Bd3 of *B. distachyon* resulted from two separate NCFs of three ARCE—Os2, Os8, and Os10. Comparative mapping indicated that Os2 corresponded to the entire Bp4 and that both Os8 and Os10 corresponded to the Bp3 chromosome of the *Brachypodium* perennials. In the genomes of both diploid 18 chromosome perennials, the full set of Bd4-specific BACs mapped along only one homoeologous counterpart, Bp5 (Figure 1). Probes Bd4S/1–6 and Bd4L/7–13 mapped its entire short and long arm, respectively (Figure S5). All of the applied Bd4-derived probes corresponded to the Os12, Os9, and Os11 ARCEs that were localized together on chromosomes Bp5. Finally, CCB mapping with the Bd5-derived clones revealed their conservative arrangement along one chromosome, which was identified as Bp9. According to these results, the composition of Os12, Os9, and Os11 in the Bp5 and Os4 in the Bp9 chromosome resembled that in Bd4 and Bd5 of *B. distachyon*, respectively (Figure 1).

Interestingly, the karyotypes of some *Brachypodium* diploid perennials consist of only 16 chromosomes. We observed such an atypical, $x = 8$ basic chromosome number in *B. glaucovirens* and in one of the diploid *B. pinnatum* cytotypes (PI 185135). However, barcoding with Bd1–Bd5 chromosome-specific probes showed exactly the same number and position of breakpoint regions as the one in the 18-chromosome diploids. Simultaneous hybridization of Bd1- and Bd3-derived BACs showed that the probes identifying the homoeologues of the Bp3 and Bp6 chromosomes in the 2*n* = 18 chromosome species mapped to the same chromosome pair in the 16 chromosome species (Figure 2 and Figure S8). Such a result clearly indicates the presence of a unique descending dysploidy event via so-called end-to-end fusion (EEF), or a variant mimicking it, of two chromosomes similar to Bp6 and Bp3, resulting in a single chromosome designated Bp6+Bp3. Such convention was applied to all of the other chromosomes with "dual" origin. Among the *Brachypodium* perennial diploids studied to date, such a chromosome has only been found in *B. glaucovirens* and *B. pinnatum* PI 185135 and results in a decrease in their basic chromosome number from $x = 9$ to $x = 8$.

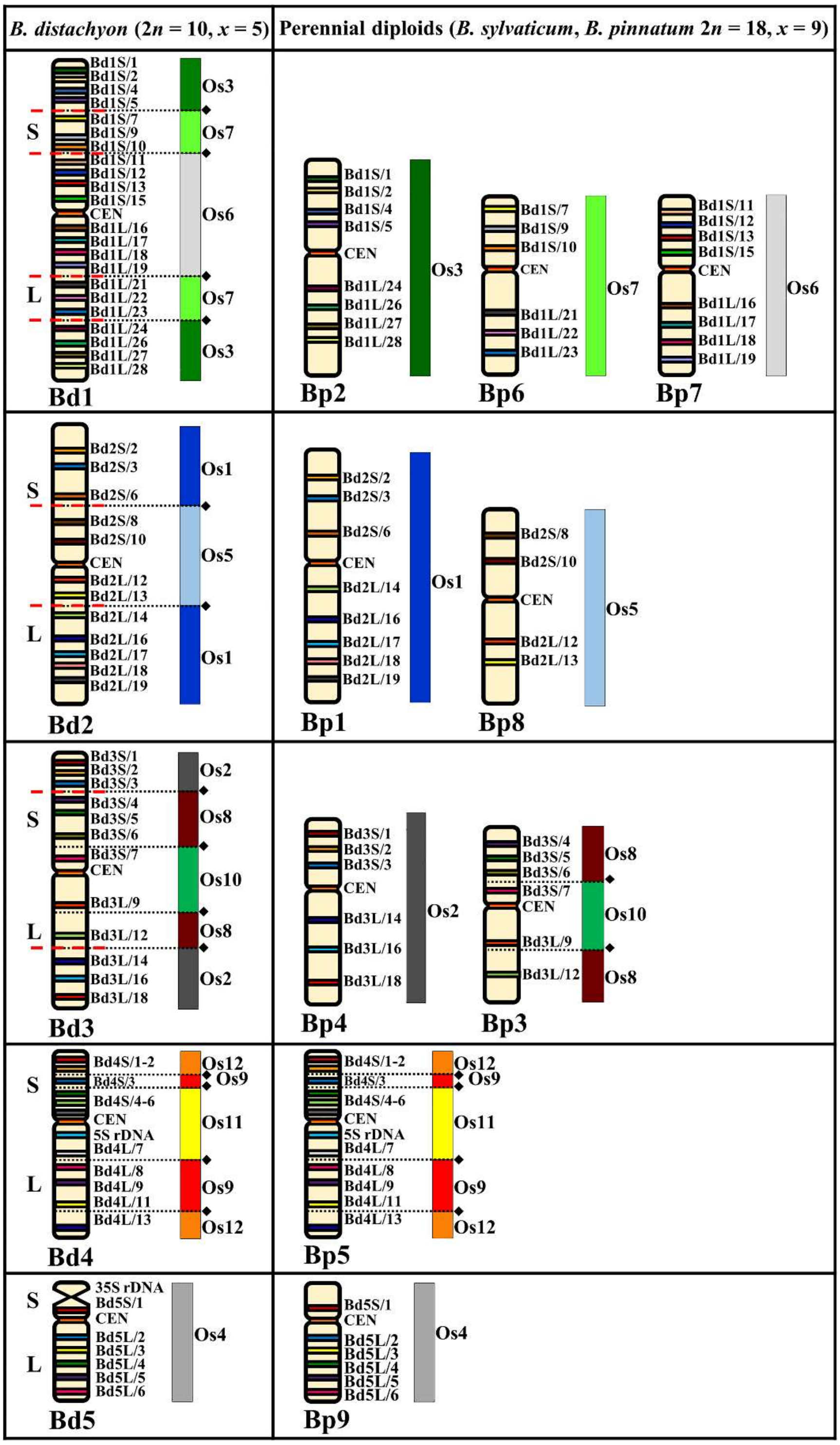


Figure 1. Distribution of the bacterial artificial chromosome (BAC) clones derived from chromosomes Bd1–Bd5 of *B. distachyon* (2*n* = 10, *x* = 5) that were comparatively mapped to the chromosomes of the *Brachypodium* perennial diploids (*B. sylvaticum* and *B. pinnatum*, 2*n* = 18, *x* = 9). Only one homologue from a pair is shown. The diagrams next to the *Brachypodium* (Bd, Bp) chromosomes align the BAC clones to the homoeologous regions (syntenic segments) in the relevant ancestral rice chromosome equivalents (ARCEs), Os1–Os12. Black diamonds and dotted lines indicate the hypothetical fusion points of the ARCE (adapted from IBI, [4]). Red, dashed lines indicate the chromosomal breakpoints in the Bp-genome chromosomes in *B. sylvaticum* and *B. pinnatum* 2*n* = 18 that were found by comparative chromosome barcoding.

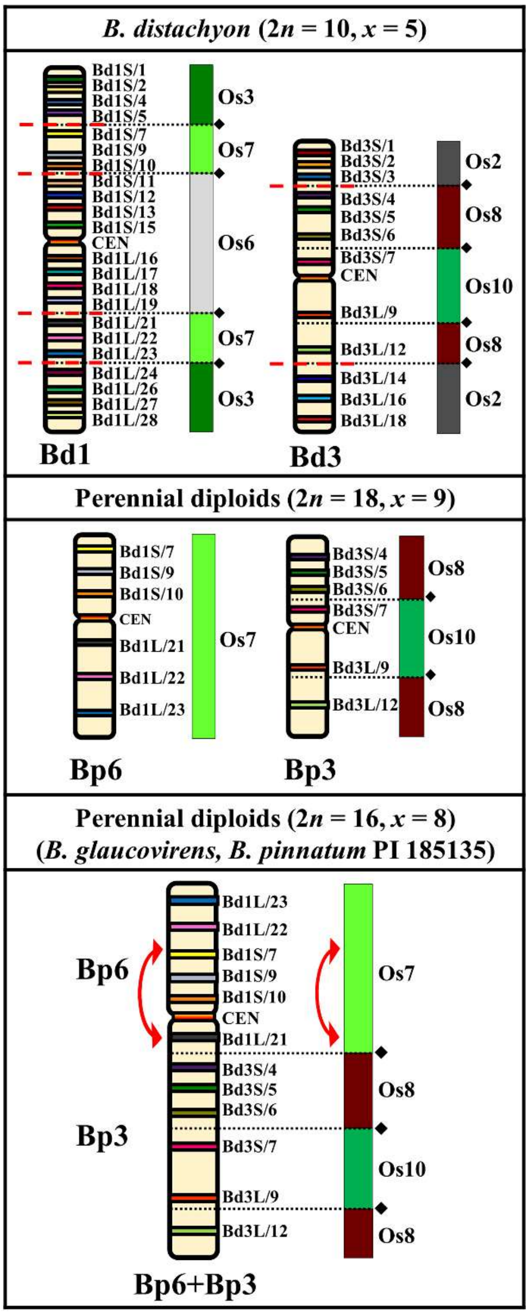


Figure 2. Distribution of the bacterial artificial chromosome (BAC) clones derived from chromosomes Bd1 and Bd3 of *B. distachyon* (2*n* = 10, *x* = 5) that were comparatively mapped to chromosomes Bp6 and Bp3 of the *Brachypodium* perennial diploids (2*n* = 18, *x* = 9) and to the chromosome Bp6+Bp3 of *B. glaucovirens* and *B. pinnatum* PI 185135 (both 2*n* = 16, *x* = 8). Only one homologue from a pair is shown. The diagrams next to the *Brachypodium* (Bd, Bp) chromosomes align the BAC clones to the homoeologous regions (syntenic segments) in the relevant ancestral rice chromosome equivalents (ARCEs). Black diamonds and dotted lines indicate the hypothetical fusion points of the ARCE (adapted from IBI, [4]). Red, dashed lines indicate the chromosomal breakpoints in the Bp-genome chromosomes of *B. sylvaticum* and *B. pinnatum* (2*n* = 18) that were found by comparative chromosome barcoding. The diagram for the *x* = 8 perennial diploids shows the specific end-to-end translocation of the putative Bp6 and Bp3 chromosomes which led to the formation of a specific Bp6+Bp3 chromosome. Red arrow points to the pericentric inversion that was found on this chromosome in *B. glaucovirens* and *B. pinnatum* PI 185135 2*n* = 16.

2.2. Comparative Mapping of the Chromosomes in the Perennial Allotetraploids (2n = 28 Chromosomes)

The CCB of the allotetraploids *B. pinnatum* and *B. phoenicoides*, both 2n = 28 chromosomes, revealed that each single-locus BAC had four hybridization sites that were located on two chromosome pairs. Several genotypes of these polyploids (Table 1) had no intraspecific variation in either the number or the arrangement of the FISH loci (Figures S6 and S7). We were able to distinguish two distinct groups of chromosomes in the karyotypes of these allotetraploids on the basis of their distinctive hybridization signals (Figure 3). One consisted of five pairs of chromosomes and the other nine which can be regarded as subgenomes of Bp with x = 5 and x = 9.

Although the basic chromosome number of subgenome x = 5 is the same as that of genome Bd in *B. distachyon*, CCB with different combinations of the probes from distinct Bd chromosomes shows a unique arrangement of the syntenic segments defined by the ARCE (Figure 3, Figures S6 and S7). Probes Bd1S/1–5 and Bd1L/24–28, which correspond to chromosome Bp2 in perennial diploids, hybridized to the distal parts of the long and short arms of one chromosome in the x = 5 subgenome (Figure S2). The central part of this chromosome had hybridization sites of Bd3S/4–7 and Bd3L/9–12 BAC clones, which corresponded to chromosome Bp3 (Figure S4). This indicates a fusion of two ancestral chromosomes that resemble the current Bp2 and Bp3 chromosomes. Based on its BAC clone composition, this chromosome was named Bp2+Bp3 (Figure 3, Figures S6 and S7). Additionally, some of the BAC loci in Bp2+Bp3 had an altered orientation, most likely indicating the presence of a pericentric inversion involving the region delimited by clones Bd3S/4–7 and CEN as well as a paracentric inversion (clones Bd3L/9–12) in the long arm (Figure 3; red arrows). Three other chromosomes in the x = 5 subgenome arose as a result of NCFs involving the ARCE, which were similar to those of the Bp x = 9 genome. Chromosome Bp4+Bp6 comprised the Bp6 equivalent marked by Bd1S/7–10 and Bd1L/21–23 BAC clones, and the Bp4 equivalent was marked by BACs Bd3S/1–3 and Bd3L/14–18 clones (Figure 3, Figures S2 and S4). Another chromosome was designated Bp5+Bp7 (Figure 3, Figures S6 and S7) as it contained all of the clones from Bd4 that corresponded to both chromosome arms of Bp5 (Figure S5) as well as BACs Bd1S/11–15 and Bd1L/16–19 which marked Bp7 (Figure S2). Heterologous mapping of Bd2-originated probes—BdS2/8–10 and Bd2L/12–13, which in the perennial diploids x = 9 corresponded to chromosome Bp8 (Figure S3), and Bd5-derived BACs Bd5S/1 and Bd5L/2–4, which mark chromosome Bp9 (Figure S5)—identified chromosome Bp9+Bp8 (Figure 3, Figures S6 and S7). The arrangement of all of the Bd2S/2–6 and Bd2L/14–19 BAC landmarks in the last chromosome of subgenome x = 5 was identical to their distribution along chromosome Bp1 in the perennial diploids. Taking into account the morphological similarity to its counterpart in the x = 9 subgenome, this chromosome was also designated Bp1 (Figure 3 and Figure S3).

The BAC–FISH signal distribution in all of the chromosomes belonging to the second group was identical to that found in the chromosomes of the x = 9 genome Bp of the *Brachypodium* perennial diploids (2n = 18). This observation provided strong evidence that this genome is conserved and constitutes one of the subgenomes of the perennial *Brachypodium* allotetraploids (Figures 1 and 3).

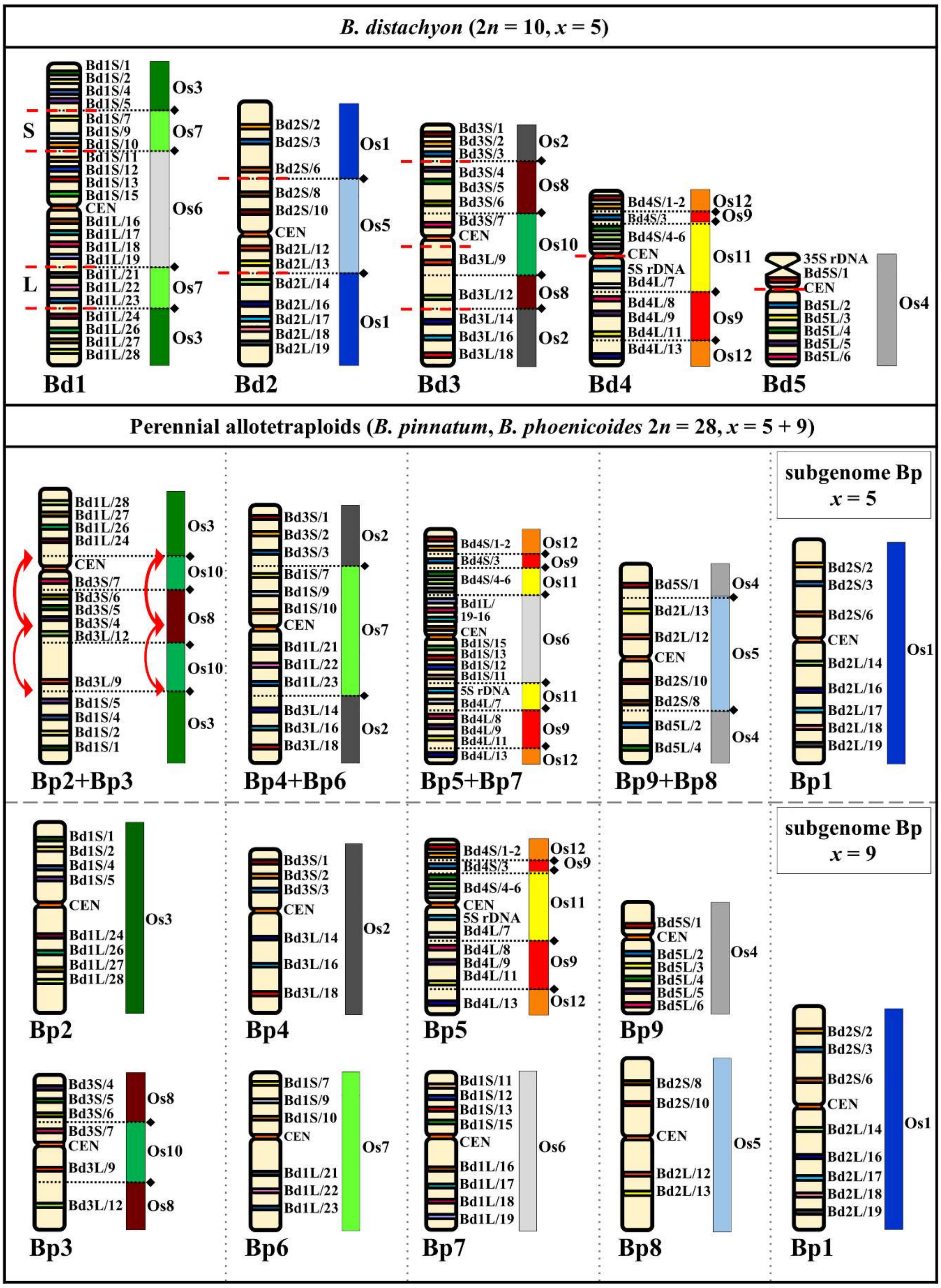


Figure 3. Distribution of the bacterial artificial chromosome (BAC) clones derived from chromosomes Bd1–Bd5 of *B. distachyon* (2*n* = 10, *x* = 5) that were comparatively mapped to the chromosomes of the *Brachypodium* perennial allotetraploids (2*n* = 28, *x* =5 + 9). Only one homologue from a pair is shown. The diagrams next to the *Brachypodium* (Bd, Bp) chromosomes align the BAC clones to the homoeologous regions (syntenic segments) in the relevant ancestral rice chromosome equivalents (ARCEs), Os1–Os12. Black diamonds and dotted lines indicate the hypothetical fusion points of the ARCE (adapted from IBI, [4]). Red, dashed lines indicate the chromosomal breakpoints in the chromosomes of two Bp subgenomes in *B. pinnatum* 2*n* = 28 and *B. phoenicoides* that were found by comparative chromosome barcoding. Red arrows point to the most likely one pericentric and one paracentric inversion that were found in chromosome Bp2+Bp3 in *B. pinnatum* 2*n* = 28 and *B. phoenicoides*.

2.3. Comparative Chromosome Barcoding of B. mexicanum (2n = 40 Chromosomes)

Each of the Bd-derived clones had four hybridization sites in the chromosome complement of *B. mexicanum* that were usually localized in two morphologically more or less diverse homoeologous chromosome pairs. This implies that *B. mexicanum* is a tetraploid consisting of two 10 chromosome subgenomes ($x = 10 + 10$) which were designated Bm and Bm′. Heterologous mapping of the BACs originating from chromosome Bd1 showed hybridization to seven different chromosomes (Figure 4). The BACs Bd1S/1–6 and Bd1L/24–29, which corresponded to Os3, hybridized to chromosomes Bm2 and Bm2′. However, their exact distribution in these chromosomes was different, suggesting the presence of a large pericentric inversion combined with a duplication of the region that hybridized with the clone Bd1S/1 from chromosome Bm2′ (Figure 4; red arrow). Other BAC clones derived from Bd1 that corresponded to the Os7 ARCE mapped to chromosomes Bm6 and Bm6′. A comparison of the arrangement of clones Bd1L/21–23 indicated the presence of a paracentric inversion in the long arm of Bm6′ (Figure 4; red arrow). The BAC clones that corresponded to the Os6 ARCE were also mapped on individual chromosomes of *B. mexicanum* (i.e., Bm7 and Bm7′) as well as to a short distal segment along Bm3′ (Figures S9 and S10). The CCB of the clones derived from Bd2 highlighted four homoeologues in *B. mexicanum* (Figure 4 and Figure S11). Two of them, Bm1 and Bm1′, carried Bd2S/1–6 and Bd2L/14–20 BAC clone loci corresponding to Os1 ARCE, whereas Os5 ARCE was represented by chromosomes Bm8 and Bm8′ which had the loci of the Bd2S/7–11 and Bd2L/12–13 clones. The Bd3-derived BACs mapped to five homoeologues (Figure 4, Figures S9, S10 and S12). The loci of clones Bd3S/1–3 and Bd3L/13–18 had a similar pattern on both the Bm4 and Bm4′ chromosomes, whereas the probes Bd3S/4–7 and Bd3L/8–12 hybridized to the three homoeologous chromosomes Bm3, Bm3′, and Bm7′. While all of the Bd3-derived BACs mapped to chromosome Bm3, and the clones derived from the short and long arms of Bd3 hybridized separately to Bm3′ and Bm7′, respectively (Figures S9, S10 and S12). Such specific arrangement of the Bd1- and Bd3-derived probes indicates the occurrence of a reciprocal translocation between chromosomes Bm3′ and Bm7′ (Figure 4). The set of clones from Bd4 mapped to four different chromosomes of *B. mexicanum*. The BACs Bd4S/1–3 and Bd4L/8–13, which correspond to the Os12+Os9 ARCE segments in Bd4, spanned the entire length of the short and long arms of chromosome Bm5, respectively. Another homoeologue, Bm5′, had a similar distribution of these clones except for the terminal fragment, which contained the clones Bd4L/12–13 that underwent an inter-arm translocation that was connected with the inversion (Figure 4; red arrows). Chromosome Bm5 was the only 35S rDNA-bearing chromosome with a distinct secondary constriction on its short arm, whereas Bm5′ contained a 5S rDNA site that was localized subterminally on the long arm. Clones Bd4S/4–6 and Bd4L/7, which in Bd4 are associated with the Os11 ARCE, hybridized with chromosomes Bm10 and Bm10′ (Figure S13). These BACs mapped in a conserved order along Bm10 as in Bd4, but the Bm10′ hybridization sites of Bd4L/7 and 5S rDNA were located together on the long arm with clones Bd4S/4–6. This suggests the presence of a small pericentric inversion that involved the proximal region of Bm10′ (Figure 4; red arrows). Bd5-derived BACs hybridized with two chromosomes, Bm9 and Bm9′, in the same manner as in the Bd genome (Figure 4 and Figure S14).

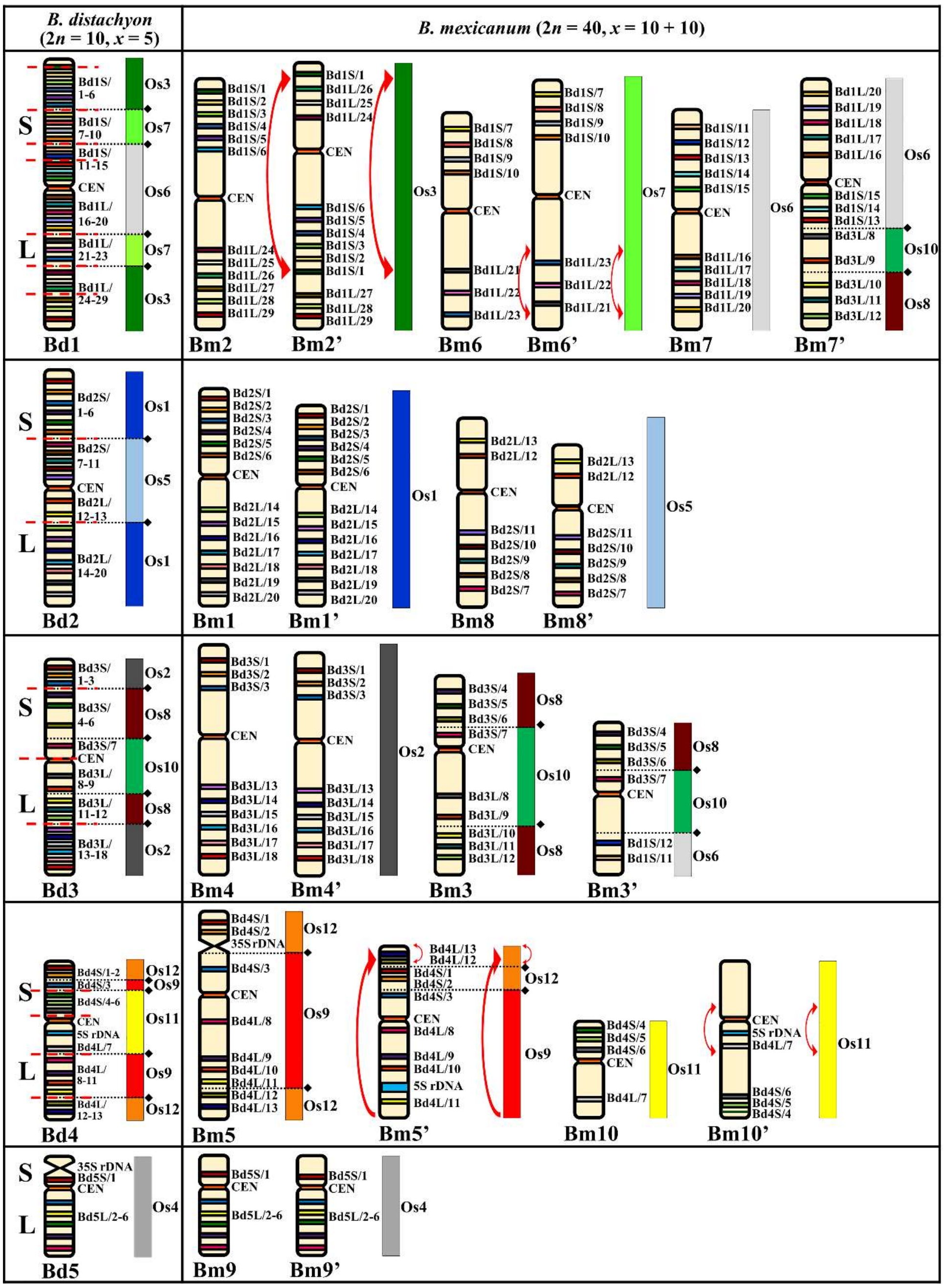


Figure 4. Distribution of the bacterial artificial chromosome (BAC) clones derived from chromosomes Bd1–Bd5 of *B. distachyon* (2*n* = 10, *x* = 5) that were comparatively mapped to the chromosomes of *B. mexicanum* (2*n* = 40, *x* = 10 + 10). Only one homologue from a pair is shown. The diagrams next to the *Brachypodium* (Bd, Bm) chromosomes align the BAC clones to the homoeologous regions (syntenic segments) in the relevant ancestral rice chromosome equivalents (ARCEs), Os1–Os12. Black diamonds and dotted lines indicate the hypothetical fusion points of the ARCEs (adapted from IBI, [4]). Red, dashed lines indicate the chromosomal breakpoints in the Bm-subgenome and Bm′-subgenome chromosomes of *B. mexicanum* that were found by comparative chromosome barcoding. Red arrows point to a pericentric inversion combined with a duplication of the region hybridizing with clone Bd1S/1 that was found on chromosome Bm2′; the paracentric inversion in the long arm of Bm6′; the translocation connected with the inversion of the terminal fragment containing clones Bd4/L12–13 in Bm5′; and a small pericentric inversion in the proximal region of chromosome Bm10′ that involves 5S rDNA and Bd4L/7 loci.

3. Discussion

3.1. Karyotype Evolution in the Perennial Diploids

Most of the *Brachypodium* species in this group have x = 9 chromosomes which suggests that some chromosome fusions must have occurred during the divergence of their karyotypes from the 12 chromosome Intermediate Ancestral Grass Karyotype (IAGK) [23]. We showed the same distribution pattern of BAC–FISH signals in *B. sylvaticum* and *B. pinnatum* (2n = 18) chromosomes (Figure 1). Their karyotypes had the same structure and pattern of dysploidy events. Chromosomes Bp3 and Bp5 were formed by NCFs, which involved Os8+Os10 and Os12+Os9+Os11, respectively. All of the seven remaining chromosomes (i.e., Bp2, Bp6, Bp7, Bp1, Bp8, Bp4, and Bp9) did not undergo NCF events and directly correspond to Os3, Os7, Os6, Os1, Os5, Os2, and Os4 ARCEs, respectively. The same Os12+Os9+Os11 fusions as in Bp5 were observed in Bd4 of the reference *B. distachyon* karyotype, whereas, in Bp3, only one NCF (Os8+Os10) was detected. This particular fusion was also present in chromosome Bs3 of the annual *B. stacei* and its allotetraploid derivative *B. hybridum* [44].

Moreover, it was also found in all homoeologues across the *Brachypodium* species, which suggests that it might be one of the most ancient NCF events involving two ancestral chromosomes that were fused in the putative x = 10 Ancestral *Brachypodium* Karyotype (ABK, Figure 5). In the perennial diploids, Os12, Os9, and Os11 comprised chromosome Bp5 (Figure 1) and in *B. distachyon* chromosome Bd4, while, in *B. stacei*, they were found in two chromosomes (i.e., Bs10 and Bs5) [44]. Thus, it can be inferred that the Os12+Os9+Os11 fusions occurred before the divergence of *B. stacei* (16.2 Ma), *B. distachyon* (10.6 Ma), and the core perennial clade (6.1 Ma) [6,8,13,15]. As was shown by Lusinska et al. [44], the Bs10 and Bs5 split was possibly the result of a Robertsonian rearrangement, which was responsible for an ascending dysploidy (Figure 5) in the Bs genome.

Most of the perennial diploids had 2n = 18 chromosomes but species with 2n = 16 have also been described. We revealed that, in *B. glaucovirens* and in *B. pinnatum* PI 185135, a combination of the Bd1- and Bd3-derived BAC-based probes hybridized to the same chromosome, indicating the presence of an EEF (or asymmetric reciprocal translocation between the ends of two metacentric chromosomes that mimics EEF) involving chromosomes similar to Bp6 and Bp3 which is responsible for the descending dysploidy from x = 9 to x = 8 via the formation of a unique chromosome, Bp6+Bp3, in this karyotype (Figure 2). Based on nuclear genome size estimates [12], it can be assumed that this dysploidy was not associated with genome downsizing. When two (sub)metacentric chromosomes are involved in EEF, at least one becomes telo- or acrocentric via a pericentric inversion [46]. Such an inversion was detected in the chromosome Bp6+Bp3, and it can be assumed that this rearrangement occurred in an ancestral chromosome that was similar to Bp6 before its EEF with Bp3. Pericentric inversions that accompany chromosome fusions were detected on the dense genetic and cytogenetic maps of some Brassicaceae [39,47] and *Cucumis* [41] representatives. In grasses, the occurrence of EEFs was reported in maize and Wang et al. [48] postulated that such a mechanism was responsible for the organization of several chromosomes. However, EEFs seem to be rare in *Brachypodium*.

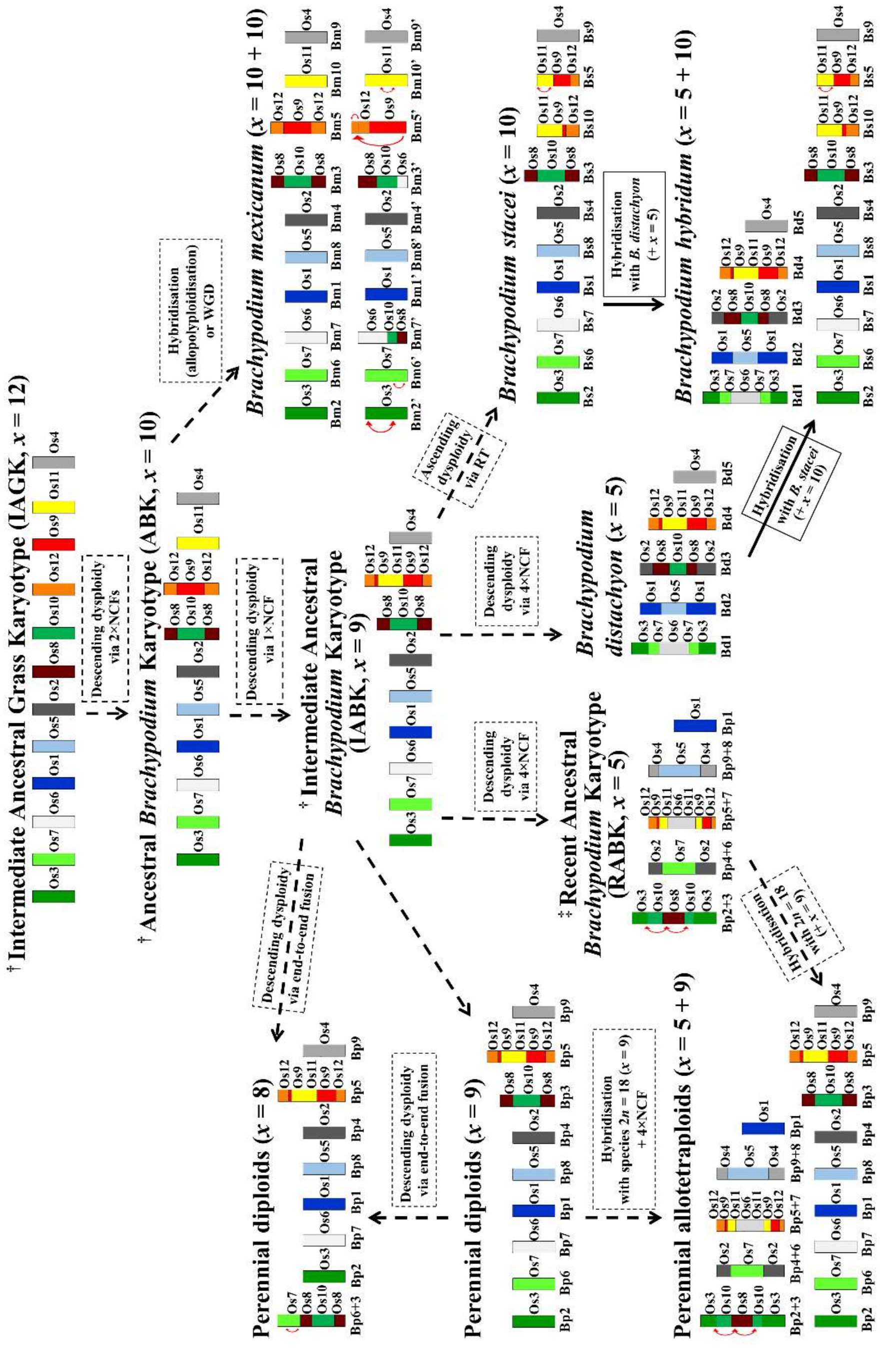

† Intermediate Ancestral Grass Karyotype (IAGK, x = 12)
Os3 Os7 Os6 Os1 Os5 Os2 Os8 Os10 Os12 Os9 Os11 Os4
Descending dysploidy via 2×NCFs
† Ancestral Brachypodium Karyotype (ABK, x = 10)
Descending dysploidy via 1×NCF
† Intermediate Ancestral Brachypodium Karyotype (IABK, x = 9)
Hybridisation (allopolyploidisation) or WGD
Brachypodium mexicanum (x = 10 + 10)
Bm2 Bm6 Bm7 Bm1 Bm8 Bm4 Bm3 Bm5 Bm10 Bm9
Bm2' Bm6' Bm7' Bm1' Bm8' Bm4' Bm3' Bm5' Bm10' Bm9'
Ascending dysploidy via RT
Brachypodium stacei (x = 10)
Bs2 Bs6 Bs7 Bs1 Bs8 Bs4 Bs3 Bs10 Bs5 Bs9
Hybridisation with B. distachyon (+ x = 5)
Brachypodium hybridum (x = 5 + 10)
Bd1 Bd2 Bd3 Bd4 Bd5
Descending dysploidy via 4×NCF
Brachypodium distachyon (x = 5)
Hybridisation with B. stacei (+ x = 10)
‡ Recent Ancestral Brachypodium Karyotype (RABK, x = 5)
Bp2+3 Bp4+6 Bp5+7 Bp9+8 Bp1
Hybridisation with 2n = 18 (+ x = 9)
Descending dysploidy via end-to-end fusion
Perennial diploids (x = 8)
Bp6+3 Bp2 Bp7 Bp1 Bp8 Bp4 Bp3 Bp5 Bp9
Perennial diploids (x = 9)
Bp2 Bp6 Bp7 Bp1 Bp8 Bp4 Bp3 Bp5 Bp9
Hybridisation with species 2n = 18 (x = 9) + 4×NCF
Perennial allotetraploids (x = 5 + 9)

Figure 5. A comprehensive model of karyotype evolution in the genus *Brachypodium* inferred from the Intermediate Ancestral Grass Karyotype based on IBI [4] and Murat et al. [22]. The model is based on the results of the comparative chromosome barcoding-based mapping in all of the *Brachypodium* perennial diploids, allotetraploids, and in the *B. mexicanum* that were analyzed in this study as well as in the *Brachypodium* annuals, *B. stacei* and *B. hybridum* [44]. Os—ancestral rice chromosome equivalents (ARCEs). Genome/subgenome designations: Bd—*B. distachyon*, Bp—*Brachypodium* perennials, Bm, Bm'—*B. mexicanum*, Bs—*B. stacei*. Dashed arrows indicate hypothetical evolutionary pathways of *Brachypodium* karyotypes. In some cases, two alternative pathways are shown. The solid arrow shows the origin of *B. hybridum* which is experimentally determined [49]. Red arrows point to the minor intrachromosomal rearrangements (inversions, translocations). † Species with such karyotypes are extinct. ‡ Diploid species with this karyotype are extinct or unknown.

3.2. Karyotype Evolution in the Perennial Polyploids

It is recognized that polyploidy followed by subsequent diploidization are major mechanisms that drive genomic diversity and evolution in angiosperms [22,50]. Current comparative karyotypic data suggest that the post-polyploid descending dysploidies are more common than the ascending ones. A return to a reduced diploid state usually occurs via reciprocal translocations, which are either NCFs that commonly occur in grasses [4,44,51,52] and some eudicots [40,41,53] or EEFs which seem to be common in eudicots [39,50,54].

The allotetraploid nature of many *Brachypodium* perennials has already been inferred from comparative genomic in situ hybridization [12] and CCP-based analyses [9]. Previous studies suggested that the 28 chromosome *Brachypodium* species can be allopolyploids, which may have been derived from diploid 2n = 18 (some genotypes of *B. sylvaticum* and *B. pinnatum*) and 2n = 10 (*B. distachyon*) progenitors [9,10,12]. However, the results of other cytomolecular analyses [14,55] and recent phylogenetic studies [13] suggested that the perennial allopolyploids had originated from the hybridizations of various 2n = 18 core-perennial diploids. In this study, we confirmed the allopolyploid nature of *B. phoenicoides* and *B. pinnatum* (2n = 28) and identified all of the chromosomes that had been derived from putative parental genomes (Figure 3). Those of the x = 9 subgenome corresponded to the nine Bp chromosomes that are found in the perennial diploids, whereas five chromosomes that belong to the second subgenome were characterized by a plethora of complex descending dysploidy events. We showed that three ancestral NCFs involving five ancestral chromosomes (Os8+Os10 and Os12+Os9+Os11) are present in the genomes of *Brachypodium* annuals [44] as well as in all *Brachypodium* perennials except *B. mexicanum*. Moreover, four additional NCFs were found in the subgenome x = 5 of *Brachypodium* perennial allotetraploids. These involved eight ancestral chromosomes and are probably more recent, since their patterns did not reflect any of the several rounds of descending dysploidy events known for *B. distachyon* chromosomes (Figure 3). Recently, an inferred homology among Triticeae, rice, and *B. distachyon* chromosomes revealed different chromosome evolution trajectories in the Triticeae and *B. distachyon* lineages. Seven Triticeae chromosomes resulted from four NCFs and one EEF of 12 ARCEs that constitute the IAGK, while five *B. distachyon* chromosomes arose through seven NCF events. Interestingly, neither a single fusion event that formed intermediate and/or extant chromosomes was shared by the Triticeae and *B. distachyon* lineages [56] nor by Triticeae and *Brachypodium* perennial allotetraploids in this study.

Initially, the basic chromosome number of *B. mexicanum* was suggested as five [16], but the results of more recent studies estimated it to be ten and indicated a possible allotetraploid nature of this species [8,9]. Our current study provides a strong indication that *B. mexicanum* is a tetraploid with a karyotype consisting of two subgenomes with x = 10 in each (Figure 4). Their individual Bm and Bm′ homoeologues display various degrees of similarity with several peri- and paracentric inversions and translocations identified in some of the homoeologues (Figure 4). Because of the similarity of its subgenomes, it is not clear if *B. mexicanum* is an allotetraploid or autotetraploid with structural changes in the Bm′ subgenome after a WGD. In contrast to other *Brachypodium* representatives, we revealed that *B. mexicanum* chromosomes carry only two ancient fusions, which is the lowest number within the genus to date. The first was an Os8+Os10 fusion, which was present in chromosomes Bm3, Bm3′, and Bm7′. The second fusion was the Os12+Os9 that was found in chromosomes Bm5 and Bm5′ (Figure 4). The Os8+Os10 and Os12+Os9 fusions were present in all of the *Brachypodium* species studied to date. Moreover, *B. mexicanum* is the only *Brachypodium* representative that does not have Os12+Os9 ARCEs fused with Os11.

3.3. Brachypodium Karyotype Evolution

The current study together with cytomolecular analyses of *Brachypodium* annuals [44] permitted the creation of a hypothetical model of *Brachypodium* karyotype evolution (Figure 5). It begins with the IAGK (x = 12) through separate descending dysploidy events which resulted in inferred putative ABK with x = 10 and an Intermediate Ancestral *Brachypodium* Karyotype (IABK) with x = 9 chromosomes.

Based on the results of cytomolecular mapping, we deduced that most of the perennial and annual species probably evolved from an ancestor that had IABK, because of the presence of the Os12, Os9, and Os11 segments that were already fused in their genomes. We inferred that *B. mexicanum* evolved directly from an x = 10 ancestor via autopolyploidization or allopolyploidization, which is evidenced by the lack of Os12+Os9 fused with Os11. The phylogenetic analysis of plastid and nuclear loci suggest the existence of an ancestral homoeologous subgenome not found in current diploid species and present only in *B. mexicanum* and the high polyploids (*B. boissieri*, *B. retusum*) [8,13]. This early split was followed by the split of diploid *B. stacei* and its close polyploid subgenomes, such as the one present in *B. mexicanum*, and the split of a more recent sister relation of diploid *B. distachyon* and the core perennial clade composed of diploid and polyploid species [8,13]. These phylogenetic data partially corroborate our findings in *B. mexicanum,* insofar as the first lineage that diverged from a common ancestor was characterized by x = 10. However, the arrangement of the fused Os12+Os9 and the separate Os11 in *B. mexicanum* seems to be in agreement with some of the phylogenetic data of Díaz-Pérez et al. [13] that indicated the involvement of an ancestral genome older than that of *B. stacei* in *B. mexicanum*. However, this conclusion is confounded by other data that identifies an additional subgenome homoeologous to that of *B. stacei.* Thus, the results of CCB clearly contradict the notion that one of the *B. mexicanum* subgenomes originated from the genome Bs, but this assumption is based only on the karyotypic data (Figure 5) [44]. The cytomolecular data also support the separate evolution of the diploid annuals from the ancestor with IABK (x = 9). We assume that the divergence of genome Bs could occur via ascending dysploidy and that the genome Bd could emerge via multiple descending dysploidy events (Figure 5) [44], though the evolutionary timing of these events could not be established from the current data.

The lineages of the extant *Brachypodium* perennial diploids (2n = 18) are likely to have originated from an ancestor that was characterized by x = 9 IABK. Unlike genome Bd, they did not undergo the series of NCFs that was responsible for the chromosome number reduction in *B. distachyon*. However, the x = 8 chromosome genomes Bp of *B. glaucovirens* and *B. pinnatum* PI 185135 might have arisen either directly from an x = 9 intermediate ancestor with IABK or from other perennial diploids via the occurrence of one EEF event resulting in the chromosome Bp6+Bp3 (Figure 5). The current data support two hypothetical models of the origin of the perennial allotetraploids. One infers the existence of a progenitor species with a Recent Ancestral *Brachypodium* Karyotype (RABK) consisting of five chromosomes (Figure 5) that contributed to the cross with the x = 9 diploids followed by genome doubling. This model explains well the striking conservation of the NCF patterns that are observed in the subgenome Bp x = 5 in various perennial allotetraploids. However, the existence of a RABK x = 5 progenitor is speculative. It cannot be ruled out that the NCFs that were observed in the chromosomes of the Bp x = 5 subgenome reflect a putative "ghost" genome which is now extinct from or unknown in the diploids.

The other model assumes that perennial allotetraploids resulted from the hybridization of two different x = 9 diploids followed by descending dysploidy via four NCFs (Figure 5) which was also postulated by Catalan et al. [8]. Considering this hypothetical pathway, it is likely that the descending dysploidy involves only one of the contributing ancestral genomes. This conclusion is supported by the fact that the inter-chromosomal fusions never involved the same ARCE. The NCF patterns that are specific for the subgenome Bp x = 5 chromosomes were highly conserved in several genotypes of both the *B. phoenicoides* and *B. pinnatum* allotetraploids (Figures S6 and S7). This implies that the genomic and possibly taxonomic variability between these taxa might be the result of their independent divergence occurring after polyploidization.

Recent phylogenetic data discriminate between homoeologous "ancestral" and "recently evolved" gene copies at the GIGANTEA locus and to a lesser extent also within the ITS and ETS of ribosomal DNA loci thus providing new insight into the origin of perennial allopolyploids [8,13]. Phylogenetic analyses indicated the presence of genome donors in *B. phoenicoides* that are homoeologous to *B. pinnatum* 2n = 18 and *B. sylvaticum*. However, the genomic composition of *B. pinnatum* 2n = 28 is still not fully resolved since its different cytotypes have alleles that are associated with the core genomes homoeologous to those of *B. glaucovirens*, *B. sylvaticum* and *B. arbuscula* [13]. Based on these findings, the assumption is that only the perennial genomes formed the allotetraploids *B. pinnatum* and *B. phoenicoides*, which contradicts our earlier hypothesis that *B. distachyon* is also one of the genome donors [9,12]. However, the CCB-based findings of the present study clearly indicate the contribution of an unknown Bp subgenome $x = 5$ that shares the same chromosome number but has a completely different syntenic segment composition of all of its chromosomes compared to Bd. This enables the complex karyotype organization in the perennial allotetraploids to be resolved.

4. Materials and Methods

4.1. Plant Material

Six diploid and six allopolyploid genotypes of five perennial *Brachypodium* species were used in this study with reference to the *B. distachyon* inbred line Bd21. Information about their origin and basic cytogenetic properties is provided in Table 1.

4.2. Chromosome Preparation

The multi-substrate chromosome preparations were made according to the protocols of Hasterok et al. [57] and Jenkins and Hasterok [37]. In brief, young seedlings were incubated for 24 h in a box of ice, then fixed for several hours in 3:1 (*v/v*) methanol/glacial acetic acid and stored at −20 °C until they were used. Excised root tips were digested in an enzyme mixture containing 8% (*v/v*) pectinase, 1% (*w/v*) cellulase (Sigma–Aldrich, St. Louis, MO, USA), and 1% (*w/v*) cellulase, "Onozuka R-10" (Serva, Heidelberg, Germany), for 2 h at 37 °C for all species except *B. mexicanum*, where these enzymes were used at concentrations of 8%, 0.5%, and 0.5%, respectively. For squashed chromosome preparations, the meristems of each species were dissected in a small volume of 45% acetic acid followed by a separate mounting of the digested material on a slide.

4.3. Probe Labelling and FISH

The BAC clones that were used in this study (Table S1) originated from the BD_ABa and BD_CBa genomic DNA libraries and were derived from the FingerPrinted Contigs that had been assigned to the respective reference chromosomes of *B. distachyon* [38]. The details regarding centromeric clone BD_CBa0033J12 and the selection of the low-repeat BAC clones are described in Lusinska et al. [44]. Each clone was mapped to chromosome preparations of several individuals of each species or accession in order to gauge intraspecific variation (Table 1).

The BAC DNA was isolated using the standard alkaline lysis method and then labelled by nick-translation with tetramethylrhodamine-5-dUTP by nick-translation (Roche, Basel, Switzerland) with digoxigenin-11-dUTP or biotin (both Roche). The nick-translated 25S and 5S ribosomal DNA probes were based on a clone that contained a 2.3 kb *Cla*I fragment of the 25S rRNA gene of *A. thaliana* [58] and on a pTa794 clone that contained the 5S rRNA gene from common wheat [59], respectively. The probe labelling and FISH followed the Jenkins and Hasterok [37] protocol with a minor modification by Lusinska et al. [44]. All of the images were acquired using an AxioCam Mrm high-sensitivity monochromatic camera attached to an AxioImager.Z.2 wide-field epifluorescence microscope (both Zeiss, Oberkochen, Germany) and processed uniformly using ZEN 2.3 Pro (Zeiss) and Photoshop CS3 (Adobe, San Jose, CA, USA).

5. Conclusions

Our current analyses in several *Brachypodium* species enabled the dissection of their karyotype organization, tracking of the evolutionary histories of individual chromosomes, and the identification of an additional $x = 5$ genome (RABK, $x = 5$). It contributed a subgenome to the perennial allotetraploids, such as *B. pinnatum* $2n = 28$ and *B. phoenicoides*, and is probably now extinct in diploids. It seems that NCFs are much more common players than EEFs in the descending dysploidy in *Brachypodium* genomes, although one such rare EEF event is responsible for the difference in chromosome number between the 2*n* = 18 and 2*n* = 16 perennial diploids. Interestingly, all perennials lack the split of a Bd4-like chromosome that causes the ascending dysploidy from *x* = 9 to *x* = 10 which is found in all annuals except *B. distachyon*. Thus, it seems that this structural event is exclusive to the genome Bs. Although our study offers significant insight into the organization of the *B. mexicanum* karyotype and places its subgenomes among the first to have diverged from ABK with *x* = 10, it does not provide a definite answer as to whether this species is of an allopolyploid origin or whether it represents a highly restructured autopolyploid. The findings of this study enabled us to propose a model of the karyotype evolution in the *Brachypodium* genus that is inferred from IAGK *x* = 12 and to provide the most comprehensive view on the organization of *Brachypodium* genomes at the chromosomal level to date.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/20/22/5557/s1.

Author Contributions: Conceptualization, J.L., A.B., E.W. and R.H.; Formal analysis, J.L., A.B., D.L.-A., P.C., G.J., E.W. and R.H.; Funding acquisition, R.H.; Investigation, J.L.; Methodology, J.L., A.B., E.W. and R.H.; Project administration, R.H.; Resources, R.H.; Supervision, E.W. and R.H.; Validation, J.L., A.B., D.L.-A., P.C., G.J., E.W. and R.H.; Visualization, J.L.; Writing—original draft, J.L., E.W. and R.H.; Writing—review and editing, J.L., A.B., P.C., G.J., E.W. and R.H.

Funding: This research was funded by the National Science Centre Poland (grant numbers DEC-2012/04/A/NZ3/00572 and DEC-2014/14/M/NZ2/00519).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ABK	Ancestral *Brachypodium* Karyotype (*x* = 10)
ARCE	Ancestral rice chromosome equivalent
Bd	Chromosome complement of *B. distachyon* (*x* = 5)
Bp	Chromosome complement(s) of *Brachypodium* perennials (*x* = 9, 8, and 5)
Bm (Bm′)	Chromosome complement(s) of *B. mexicanum* (*x* = 10 + 10)
Bs	Chromosome complement of *B. stacei* (*x* = 10)
BAC	Bacterial artificial chromosome
CCB	Comparative chromosome barcoding
CCP	Comparative chromosome painting
CEN	Centromeric BAC BD_CBa0033J12
EEF	End-to-end fusion
FISH	Fluorescence in situ hybridization
IABK	Intermediate Ancestral *Brachypodium* Karyotype (*x* = 9)
IAGK	Intermediate Ancestral Grass Karyotype (*x* = 12)
Ma	Million years
NCF	Nested chromosome fusion
RABK	Recent Ancestral *Brachypodium* Karyotype (*x* = 5)
WGD	Whole genome duplication

References

1. Mur, L.A.; Allainguillaume, J.; Catalan, P.; Hasterok, R.; Jenkins, G.; Lesniewska, K.; Thomas, I.; Vogel, J. Exploiting the Brachypodium tool box in cereal and grass research. *New Phytol.* **2011**, *191*, 334–347. [CrossRef] [PubMed]
2. Vogel, J.P. The rise of *Brachypodium* as a model system. In *Genetics and Genomics of Brachypodium*; Vogel, J.P., Ed.; Springer International Publishing: Cham, Switzerland, 2016; pp. 1–7. [CrossRef]
3. Scholthof, K.G.; Irigoyen, S.; Catalan, P.; Mandadi, K.K. *Brachypodium*: A monocot grass model genus for plant biology. *Plant Cell* **2018**, *30*, 1673–1694. [CrossRef] [PubMed]
4. IBI. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **2010**, *463*, 763–768. [CrossRef] [PubMed]

5. Catalan, P.; Chalhoub, B.; Chochois, V.; Garvin, D.F.; Hasterok, R.; Manzaneda, A.J.; Mur, L.A.J.; Pecchioni, N.; Rasmussen, S.K.; Vogel, J.P.; et al. Update on the genomics and basic biology of *Brachypodium*. *Trends Plant Sci.* **2014**, *19*, 414–418. [CrossRef]
6. Catalan, P.; Muller, J.; Hasterok, R.; Jenkins, G.; Mur, L.A.; Langdon, T.; Betekhtin, A.; Siwinska, D.; Pimentel, M.; Lopez-Alvarez, D. Evolution and taxonomic split of the model grass *Brachypodium distachyon*. *Ann Bot.* **2012**, *109*, 385–405. [CrossRef]
7. Gordon, S.P.; Liu, L.; Vogel, J.P. The genus *Brachypodium* as a model for perenniality and polyploidy. In *Genetics and Genomics of Brachypodium*; Vogel, J.P., Ed.; Springer International Publishing: Cham, Switzerland, 2016; pp. 313–325. [CrossRef]
8. Catalan, P.; López-Álvarez, D.; Díaz-Pérez, A.; Sancho, R.; López-Herránz, M.L. Phylogeny and evolution of the genus *Brachypodium*. In *Genetics and Genomics of Brachypodium*; Vogel, J.P., Ed.; Springer International Group: Cham, Switzerland, 2016; pp. 9–38. [CrossRef]
9. Betekhtin, A.; Jenkins, G.; Hasterok, R. Reconstructing the evolution of *Brachypodium* genomes using comparative chromosome painting. *PLoS ONE* **2014**, *9*, e115108. [CrossRef]
10. Khan, M.A.; Stace, C.A. Breeding relationships in the genus *Brachypodium* (Poaceae). *Nord J. Bot.* **1999**, *19*, 257–269. [CrossRef]
11. Robertson, I.H. Chromosome numbers in *Brachypodium* Beauv (Gramineae). *Genetica* **1981**, *56*, 55–60. [CrossRef]
12. Wolny, E.; Hasterok, R. Comparative cytogenetic analysis of the genomes of the model grass *Brachypodium distachyon* and its close relatives. *Ann. Bot.* **2009**, *104*, 873–881. [CrossRef]
13. Díaz-Pérez, A.; López-Álvarez, D.; Sancho, R.; Catalán, P. Reconstructing the origins and the biogeography of species' genomes in the highly reticulate allopolyploid-rich model grass genus *Brachypodium* using minimum evolution, coalescence and maximum likelihood approaches. *Mol. Phylogenet. Evol.* **2018**, *127*, 256–271. [CrossRef]
14. Wolny, E.; Lesniewska, K.; Hasterok, R.; Langdon, T. Compact genomes and complex evolution in the genus *Brachypodium*. *Chromosoma* **2011**, *120*, 199–212. [CrossRef] [PubMed]
15. Sancho, R.; Cantalapiedra, C.P.; Lopez-Alvarez, D.; Gordon, S.P.; Vogel, J.P.; Catalan, P.; Contreras-Moreira, B. Comparative plastome genomics and phylogenomics of *Brachypodium*: Flowering time signatures, introgression and recombination in recently diverged ecotypes. *New Phytol.* **2018**, *218*, 1631–1644. [CrossRef] [PubMed]
16. Catalan, P.; Shi, Y.; Armstrong, L.; Draper, J.; Stace, C.A. Molecular phylogeny of the grass genus *Brachypodium* P. Beauv. based on RFLP and RAPD analysis. *Bot. J. Linn Soc.* **1995**, *177*, 263–280. [CrossRef]
17. Gordon, S.P.; Contreras-Moreira, B.; Woods, D.P.; Des Marais, D.L.; Burgess, D.; Shu, S.; Stritt, C.; Roulin, A.C.; Schackwitz, W.; Tyler, L.; et al. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* **2017**, *8*, 2184. [CrossRef] [PubMed]
18. Steinwand, M.A.; Young, H.A.; Bragg, J.N.; Tobias, C.M.; Vogel, J.P. *Brachypodium sylvaticum*, a model for perennial grasses: Transformation and inbred line development. *PLoS ONE* **2013**, *8*, e75180. [CrossRef] [PubMed]
19. Fox, S.E.; Preece, J.; Kimbrel, J.A.; Marchini, G.L.; Sage, A.; Youens-Clark, K.; Cruzan, M.B.; Jaiswal, P. Sequencing and de novo transcriptome assembly of *Brachypodium sylvaticum* (Poaceae). *Appl. Plant Sci.* **2013**, *1*. [CrossRef]
20. Griffiths, S.; Sharp, R.; Foote, T.N.; Bertin, I.; Wanous, M.; Reader, S.; Colas, I.; Moore, G. Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. *Nature* **2006**, *439*, 749–752. [CrossRef]
21. Bossolini, E.; Wicker, T.; Knobel, P.A.; Keller, B. Comparison of orthologous loci from small grass genomes *Brachypodium* and rice: Implications for wheat genomics and grass genome annotation. *Plant J.* **2007**, *49*, 704–717. [CrossRef]
22. Murat, F.; Armero, A.; Pont, C.; Klopp, C.; Salse, J. Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat. Genet.* **2017**, *49*, 490–496. [CrossRef]
23. Pont, C.; Wagner, S.; Kremer, A.; Orlando, L.; Plomion, C.; Salse, J. Paleogenomics: Reconstruction of plant evolutionary trajectories from modern and ancient DNA. *Genome Biol.* **2019**, *20*, 29. [CrossRef]
24. Salse, J. Deciphering the evolutionary interplay between subgenomes following polyploidy: A paleogenomics approach in grasses. *Am. J. Bot.* **2016**, *103*, 1167–1174. [CrossRef] [PubMed]

25. Salse, J. Ancestors of modern plant crops. *Curr. Opin. Plant Biol.* **2016**, *30*, 134–142. [CrossRef] [PubMed]
26. Alix, K.; Gerard, P.R.; Schwarzacher, T.; Heslop-Harrison, J.S.P. Polyploidy and interspecific hybridization: Partners for adaptation, speciation and evolution in plants. *Ann. Bot.* **2017**, *120*, 183–194. [CrossRef] [PubMed]
27. Soltis, P.S.; Marchant, D.B.; Van de Peer, Y.; Soltis, D.E. Polyploidy and genome evolution in plants. *Curr. Opin. Genet. Dev.* **2015**, *35*, 119–125. [CrossRef] [PubMed]
28. Doyle, J.J.; Sherman-Broyles, S. Double trouble: Taxonomy and definitions of polyploidy. *New Phytol.* **2017**, *213*, 487–493. [CrossRef]
29. Čertner, M.; Fenclová, E.; Kúr, P.; Kolář, F.; Koutecký, P.; Krahulcová, A.; Suda, J. Evolutionary dynamics of mixed-ploidy populations in an annual herb: Dispersal, local persistence and recurrent origins of polyploids. *Ann. Bot.* **2017**, *120*, 303–315. [CrossRef]
30. Paule, J.; Wagner, N.D.; Weising, K.; Zizka, G. Ecological range shift in the polyploid members of the South American genus *Fosterella* (Bromeliaceae). *Ann. Bot.* **2017**, *120*, 233–243. [CrossRef]
31. Cuadrado, A.; de Bustos, A.; Jouve, N. On the allopolyploid origin and genome structure of the closely related species *Hordeum secalinum* and *Hordeum capense* inferred by molecular karyotyping. *Ann. Bot.* **2017**, *120*, 245–255. [CrossRef]
32. Hasterok, R.; Draper, J.; Jenkins, G. Laying the cytotaxonomic foundations of a new model grass, *Brachypodium distachyon* (L.) Beauv. *Chromosome Res.* **2004**, *12*, 397–403. [CrossRef]
33. Hasterok, R.; Marasek, A.; Donnison, I.S.; Armstead, I.; Thomas, A.; King, I.P.; Wolny, E.; Idziak, D.; Draper, J.; Jenkins, G. Alignment of the genomes of *Brachypodium distachyon* and temperate cereals and grasses using bacterial artificial chromosome landing with fluorescence in situ hybridization. *Genetics* **2006**, *173*, 349–362. [CrossRef]
34. Lopez-Alvarez, D.; Lopez-Herranz, M.L.; Betekhtin, A.; Catalan, P. A DNA barcoding method to discriminate between the model plant *Brachypodium distachyon* and its close relatives *B. stacei* and *B. hybridum* (Poaceae). *PLoS ONE* **2012**, *7*, e51058. [CrossRef] [PubMed]
35. Idziak, D.; Hazuka, I.; Poliwczak, B.; Wiszynska, A.; Wolny, E.; Hasterok, R. Insight into the karyotype evolution of *Brachypodium* species using comparative chromosome barcoding. *PLoS ONE* **2014**, *9*, e93503. [CrossRef] [PubMed]
36. Idziak-Helmcke, D.; Betekhtin, A. Methods for cytogenetic chromosome barcoding and chromosome painting in *Brachypodium distachyon* and its relative species. In *Brachypodium Genomics: Methods and Protocols*; Sablok, G., Budak, H., Ralph, P.J., Eds.; Springer New York: New York, NY, USA, 2018; pp. 1–19. [CrossRef]
37. Jenkins, G.; Hasterok, R. BAC 'landing' on chromosomes of *Brachypodium distachyon* for comparative genome alignment. *Nat. Protoc.* **2007**, *2*, 88–98. [CrossRef] [PubMed]
38. Febrer, M.; Goicoechea, J.L.; Wright, J.; McKenzie, N.; Song, X.; Lin, J.; Collura, K.; Wissotski, M.; Yu, Y.; Ammiraju, J.S.; et al. An integrated physical, genetic and cytogenetic map of *Brachypodium distachyon*, a model system for grass research. *PLoS ONE* **2010**, *5*. [CrossRef] [PubMed]
39. Lysak, M.A.; Berr, A.; Pecinka, A.; Schmidt, R.; McBreen, K.; Schubert, I. Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 5224–5229. [CrossRef] [PubMed]
40. Mandakova, T.; Guo, X.; Ozudogru, B.; Mummenhoff, K.; Lysak, M.A. Hybridization-facilitated genome merger and repeated chromosome fusion after 8 million years. *Plant J.* **2018**, *96*, 748–760. [CrossRef]
41. Yang, L.; Koo, D.-H.; Li, D.; Zhang, T.; Jiang, J.; Luan, F.; Renner, S.S.; Hénaff, E.; Sanseverino, W.; Garcia-Mas, J.; et al. Next-generation sequencing, FISH mapping and synteny-based modeling reveal mechanisms of decreasing dysploidy in *Cucumis*. *Plant J.* **2014**, *77*, 16–30. [CrossRef]
42. Hou, L.; Xu, M.; Zhang, T.; Xu, Z.; Wang, W.; Zhang, J.; Yu, M.; Ji, W.; Zhu, C.; Gong, Z.; et al. Chromosome painting and its applications in cultivated and wild rice. *BMC Plant Biol.* **2018**, *18*, 110. [CrossRef]
43. Albert, P.S.; Zhang, T.; Semrau, K.; Rouillard, J.M.; Kao, Y.H.; Wang, C.R.; Danilova, T.V.; Jiang, J.; Birchler, J.A. Whole-chromosome paints in maize reveal rearrangements, nuclear domains, and chromosomal relationships. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 1679–1685. [CrossRef]
44. Lusinska, J.; Majka, J.; Betekhtin, A.; Susek, K.; Wolny, E.; Hasterok, R. Chromosome identification and reconstruction of evolutionary rearrangements in *Brachypodium distachyon*, *B. stacei* and *B. hybridum*. *Ann. Bot.* **2018**, *122*, 445–459. [CrossRef]
45. Wolny, E.; Fidyk, W.; Hasterok, R. Karyotyping of *Brachypodium pinnatum* (2n = 18) chromosomes using cross-species BAC-FISH. *Genome* **2013**, *56*, 239–243. [CrossRef] [PubMed]

46. Schubert, I.; Lysak, M.A. Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends Genet.* **2011**, *27*, 207–216. [CrossRef] [PubMed]
47. Mandakova, T.; Lysak, M.A. Chromosomal phylogeny and karyotype evolution in x=7 crucifer species (Brassicaceae). *Plant Cell* **2008**, *20*, 2559–2570. [CrossRef] [PubMed]
48. Wang, X.; Jin, D.; Wang, Z.; Guo, H.; Zhang, L.; Wang, L.; Li, J.; Paterson, A.H. Telomere-centric genome repatterning determines recurring chromosome number reductions during the evolution of eukaryotes. *New Phytol.* **2015**, *205*, 378–389. [CrossRef] [PubMed]
49. Dinh Thi, V.H.; Coriton, O.; Le Clainche, I.; Arnaud, D.; Gordon, S.P.; Linc, G.; Catalan, P.; Hasterok, R.; Vogel, J.P.; Jahier, J.; et al. Recreating stable *Brachypodium hybridum* allotetraploids by uniting the divergent genomes of *B. distachyon* and *B. stacei*. *PLoS ONE* **2016**, *11*, e0167171. [CrossRef] [PubMed]
50. Mandakova, T.; Lysak, M.A. Post-polyploid diploidization and diversification through dysploid changes. *Curr Opin Plant Biol.* **2018**, *42*, 55–65. [CrossRef] [PubMed]
51. Murat, F.; Xu, J.H.; Tannier, E.; Abrouk, M.; Guilhot, N.; Pont, C.; Messing, J.; Salse, J. Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.* **2010**, *20*, 1545–1557. [CrossRef]
52. Daverdin, G.; Bahri, B.A.; Wu, X.; Serba, D.D.; Tobias, C.; Saha, M.C.; Devos, K.M. Comparative relationships and chromosome evolution in switchgrass (*Panicum virgatum*) and its genomic model, foxtail millet (*Setaria italica*). *BioEnergy Res.* **2015**, *8*, 137–151. [CrossRef]
53. Fonsêca, A.; Ferraz, M.E.; Pedrosa-Harand, A. Speeding up chromosome evolution in *Phaseolus*: Multiple rearrangements associated with a one-step descending dysploidy. *Chromosoma* **2015**, *125*, 413–421. [CrossRef]
54. Mandakova, T.; Hlouskova, P.; German, D.A.; Lysak, M.A. Monophyletic origin and evolution of the largest crucifer genomes. *Plant Physiol.* **2017**, *174*, 2062–2071. [CrossRef]
55. Idziak, D.; Betekhtin, A.; Wolny, E.; Lesniewska, K.; Wright, J.; Febrer, M.; Bevan, M.W.; Jenkins, G.; Hasterok, R. Painting the chromosomes of *Brachypodium*: Current status and future prospects. *Chromosoma* **2011**, *120*, 469–479. [CrossRef] [PubMed]
56. Wang, Z.; Wang, J.; Pan, Y.; Lei, T.; Ge, W.; Wang, L.; Zhang, L.; Li, Y.; Zhao, K.; Liu, T.; et al. Reconstruction of evolutionary trajectories of chromosomes unraveled independent genomic repatterning between Triticeae and *Brachypodium*. *BMC Genom.* **2019**, *20*, 180. [CrossRef] [PubMed]
57. Hasterok, R.; Dulawa, J.; Jenkins, G.; Leggett, M.; Langdon, T. Multi-substrate chromosome preparations for high throughput comparative FISH. *BMC Biotechnol.* **2006**, *6*, 20. [CrossRef] [PubMed]
58. Unfried, I.; Gruendler, P. Nucleotide sequence of the 5.8S and 25S rRNA genes and of the internal transcribed spacers from *Arabidopsis thaliana*. *Nucleic Acids Res.* **1990**, *18*, 4011. [CrossRef] [PubMed]
59. Gerlach, W.L.; Dyer, T.A. Sequence organization of the repeating units in the nucleus of wheat which contain 5S rRNA genes. *Nucleic Acids Res.* **1980**, *8*, 4851–4886. [CrossRef]

International Journal of
Molecular Sciences

MDPI

Article

Identification of Rice Large Grain Gene *GW2* by Whole-Genome Sequencing of a Large Grain-Isogenic Line Integrated with Japonica Native Gene and Its Linkage Relationship with the Co-integrated Semidwarf Gene *d60* on Chromosome 2

Motonori Tomita [1,]*, Shiho Yazawa [2] and Yoshimasa Uenishi [1]

[1] Research Institute of Green Science and Technology, Shizuoka University, 836 Ohya, Suruga-ku, Shizuoka City, Shizuoka 422-8529, Japan; u.yoshimasa0202@gmail.com

[2] Faculty of Agriculture, Tottori University, 4-101 Koyama Minami, Tottori 680-8550, Japan; 8as41.yazawa@gmail.com

* Correspondence: tomita.motonori@shizuoka.ac.jp

Received: 19 September 2019; Accepted: 18 October 2019; Published: 31 October 2019

Abstract: Genetic analysis of "InochinoIchi," an exceptionally large grain rice variety, was conducted through five continuous backcrosses with Koshihikari as a recurrent parent using the large grain F_3 plant in Koshihikari × Inochinoichi as a nonrecurrent parent. Thorough the F_2 and all $BCnF_2$ generations, large, medium, and small grain segregated in a 1:2:1 ratio, indicating that the large grain is controlled by a single allele. Mapping by using simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) markers with small grain homozygous segregants in the F_2 of Nipponbare × Inochinoichi, revealed linkage with around 7.7 Mb markers from the distal end of the short arm of chromosome 2. Whole-genome sequencing on a large grain isogenic Koshihikari (BC_4F_2) using next-generation sequencing (NGS) identified a single nucleotide deletion in *GW2* gene, which is located 8.1 Mb from the end of chromosome 2, encoding a RING protein with E3 ubiquitin ligase activity. The *GW2*-integrated isogenic Koshihikari showed a 34% increase in thousand kernel weight compared to Koshihikari, while retaining a taste score of 80. We further developed a large grain/semi-dwarf isogenic Koshihikari integrated with *GW2* and the semidwarfing gene *d60*, which was found to be localized on chromosome 2. The combined genotype secured high yielding while providing robustness to withstand climate change, which can contribute to the New Green Revolution.

Keywords: rice; large grain gene; large grain-isogenic Koshihikari; fine mapping; NGS; *GW2*; co-integration; gene recombination; semidwarf gene; *d60*; linkage; chromosome 2

1. Introduction

There is a demand for a dramatic increase in the production of rice, as it is a staple food for over half of the world's rapidly increasing population. In the period of the 20th century when most rice breeding took place, known as the "Green Revolution," improving rice stems to be shorter decreased the likelihood of plant lodging, which made heavy manuring and dense planting possible, and radically increased yields. The "semidwarfness" led to a twofold increase in rice yields worldwide between the 1960s and the 1990s [1]. However, the trend of rice productively has now begun to plateau [2]. On top of that, the suppression of stem length is dependent on a single semidwarfing gene called *sd1*. In preparation for future increases in population and risk of crop damage due to climate change, there is a renewed demand for a "New Green Revolution" for genetic improvements to increase yields and make rice plants more robust.

Grain size is a major factor in affecting rice yield. Six quantitative trait loci (QTLs) genes related to rice grain size have been isolated [3–8]. For four of these genes, a loss of function causes the grains to be bigger, which shows that there is a mechanism that suppresses grain size. In general, Japonica rice has short and round grains, while Indica rice has long and thin grains. Whole-genome analysis of cultivated rice and wild rice shows that the domestication of rice started near the middle of the Pearl River in China, when Japonica was derived from a population of *Oryza rufipogon* [9]. Subsequently, Indica arose from hybrids between strains of wild rice in Southeast Asia and South Asia and Japonica. The *qSW5* and *GS3* genes confer grain size traits in Japonica and Indica, respectively [5,9].

The japonica rice Koshihikari is the leading variety in Japan, accounting for 36.1% of rice acreage in the country. The patent on the plant variety protection of Koshihikari, which was registered in 1956, has expired; therefore, global competition with Koshihikari produced in foreign countries is now of concern. The 2016 Trans-Pacific Partnership (TPP) eliminated tariffs on 82% (2135) of the 2594 agriculture, forestry, and fishery products that are imported by Japan [10]. The 341 JPY/kg (140%) tariffs on rice were maintained, but the simultaneous buy-and-sell (SBS) tender system with the US and Australia, which already produce Koshihikari, provides a special import framework for 78,400 t of rice. Foreign-produced Koshihikari is roughly 35% less expensive than Japanese Koshihikari, and it is genetically the same, with no difference in taste. Consequently, the influx of inexpensive Koshihikari produced in foreign countries is a concern to Japan. Furthermore, after the Trump administration began, the US pulled out of the TPP and has been pushing for a Trade Agreement on Goods (TAG) between the US and Japan to take its place. In the future, rice trade will inevitably be liberalized. Moreover, if Japan’s self-sufficiency collapses, the country will lose its paddy fields, which would not be ideal for maintaining national land conservation. In order to resolve this critical situation, there is a need to develop a low-cost and high-yield “super Koshihikari” variety that can compete in the international market.

Intensified climate change due to global warming is causing damage to crops on a global scale. Global contributions due to the New Green Revolution could uphold the innovation policy. In 2018, Japan was hit with the Western Japan heavy rain and floods [11], and seven large typhoons with wind speeds over 54 m/s, which were the worst in Japan’s history [12]. Typhoons Jebi and Trami were equivalent to the Isewan Typhoon. These extreme weather phenomena have caused marked damage to agriculture, forestry, and fisheries (totaling 436.5 billion yen) [13]. There is a need to genetically improve the sturdiness and robustness of rice plants to withstand the intensified climate change [14,15] and improve rice for the global market by lowering costs and increasing yield.

The grain weight of a large grain variety, “Inochinoichi,” is approximately 1.5 times that of Koshihikari. However, the genetic mode of the large grain is unclear, so it is not used for plant breeding at all. The production of “Inochinoichi” is also limited to the area around Gifu Prefecture. If the causative gene for the large grain of inochinoichi were to be identified, its possibility of application to improve varieties, including Koshihiakri, or to develop new varieties would be expanded. In this study, we identified from this unused and buried genetic resource Inochinoichi, the gene responsible for the large grain, and then by applying the gene to develop a large-grain Koshihikari. First, we conducted genetic analysis of the large grain through five continuous backcrosses with Koshihikari as a recurrent parent using the large grain segregant in the F_2 generation of a Koshihikari × Inochinoichi as a nonrecurrent parent. Then, we conducted whole-genome analysis of the developed large grain isogenic Koshihikari and identified a gene responsible for the large grain. Furthermore, we developed a large grain/semi-dwarf isogenic line by integrating both the identified large grain gene and the semidwarfing gene *d60* [16].

2. Results

2.1. Inheritance and Phenotypic Expression of Large Grain Gene in an Isogenic Background

As shown in Figure 1A, in the F_2 generation of Koshihikari × Inochinoichi, grain diameters showed a bimodal distribution in 0.69–0.85 mm, which were comparable to both parenteral range. Namely large grain plants with grain diameters of 0.78–0.85 mm, the same as Inochinoichi, and small grain plants with grain diameters of 0.69–0.77 mm the same as Koshihikari were segregated in a ratio of 134 large grains:52 small grains, which fit to a 3:1 ratio, (χ^2 = 0.87, df = 1, 0.35 < P < 0.40). Next, we conducted a progeny test using 50 F_3 lines consisting of 50 randomly selected F_2 plants of Koshihikari × Inochinoichi. As a results, the mean values of grain size in each F_3 line were distributed as shown in Figure 2. Namely, the F_3 progeny of large grain F_2 plants (grain diameter: 0.78–0.85 mm) were classified into a line that had fixed in large grains with diameters of 0.8–0.83 mm and a line that segregated within the lines, whereas the F_3 progeny of small grain F_2 plants (grain diameter: 0.68–0.77 mm) fixed in small grain with diameters of 0.72–0.75 mm. In other words, F_3 lines segregated in a ratio of 10 large grain homozygous lines:32 heterozygous lines:8 small grain homozygous lines, consistent with the theoretical single gene ratio (χ^2 = 4.08, df = 1, 0.10 < P < 0.25). Using the fixed large grain homozygous plant in the F_3 generation (grain diameter: 0.75 mm) as a nonrecurrent parent, five times of continuous backcrosses with Koshihikari as a recurrent parent were conducted. The BC_1F_2 plants segregated in a ratio of 10 large grain plants (grain area: 24.6–25.9 mm^2):32 small and medium grain plants (grain area: 20.6–24.0 mm^2) (Figure 1A). Furthermore, a large grain segregant in the BC_1F_2 generation (grain area: 25.9 mm^2) was used in a second backcross with Koshihikari, which yielded a BC_2F_2 plants segregated in a ratio of 17 large grain plants (grain area: 23.6–25.1 mm^2):39 medium plants (grain area: 20.6–23.5 mm^2):14 small grain plants (grain area: 19.6–20.5 mm^2); in both generations, segregation ratios fit to the theoretical single gene ratio (χ^2 = 0.03, df = 1, 0.50 < P < 0.90; χ^2 = 1.17, df = 2, 0.55 < P < 0.60) (Figure 1A). Subsequently, the BC_3F_2 plants segregated in a ratio of 15 large grain (grain area: 19.6–20.5 mm^2):13 medium grain (grain area: 19.6–20.5 mm^2):7 small grain (grain area: 19.6–20.5 mm^2) (χ^2 = 5.97, df = 2, 0.05 < P < 0.10). Next, a large grain BC_3F_2 segregant (grain area: 23.05 mm^2) was used for the fourth backcross with Koshihikari, whose BC_4F_2 progenies segregated in a ratio of 8 large grain (grain area: 26.1–29.5 mm^2):18 medium grain (grain area: 23.6–26.0 mm^2):10 small grain (grain area: 21.1–23.5 mm^2) plants, that fit well to a 1:2:1 ratio (χ^2 = 0.20, df = 2, 0.85 < P < 0.90) (Figure 1A). As seen above, from genetic analyses of large grain through the four times of backcrosses with Koshihikari, each BC_2F_2 to BC_4F_2 progeny segregated in the theoretical ratio for single incomplete dominance gene, namely 1 large grain:2 medium grain:1 small grain (Figure 1A). This indicates that the large grain is definitely inherited as a single allele. Finally, the large grain isogenic Koshihikari (BC_5F_2), which produced by backcrossed with Koshihikari and a large grain BC_4F_2 segregant, showed a grain area 27.7% greater than that of Koshihikari (Koshihikari average grain area: 22.32 mm^2, large grain phenotype average: 28.50 mm^2), and the thousand kernel weight increased by 34%. Its taste score (80.0) was also equivalent to that of Niigata Koshihikari (81.0) (Table 1); thus, this isogenic Koshihikari holds promise as a Super Koshihikari, which is distinguishable from US-made Koshihikari.

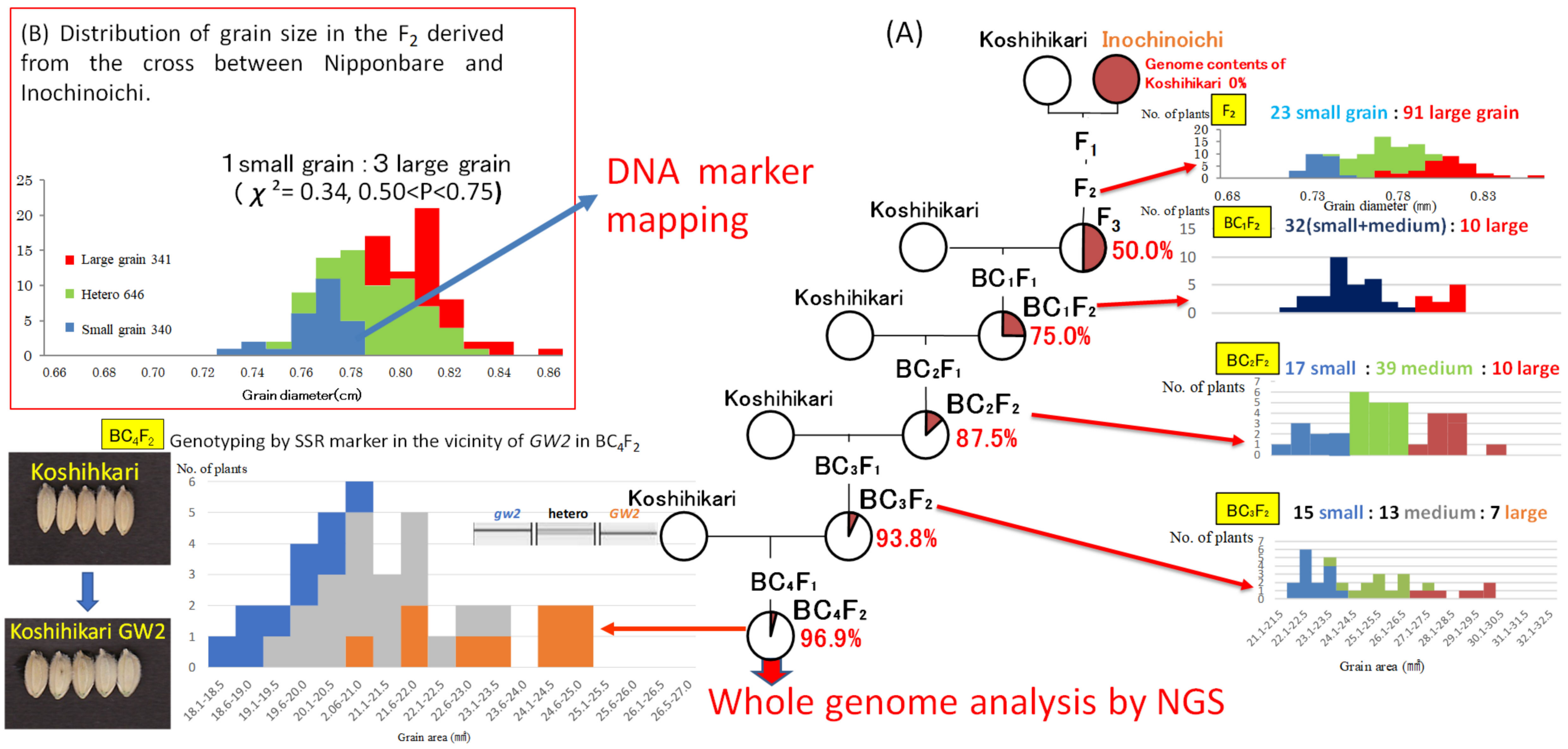


Figure 1. Procedure to identify large grain gene in Inochinoichi. (**A**) Whole-genome analysis of isogeneic line developed by continuous back cross integrating the large grain gene. Genetic analyses for large grain through the four times of backcrosses with Koshihikari, each BC_2F_2 to BC_4F_2 progeny segregated in the theoretical ratio for single incomplete dominance gene, namely 1 large grain:2 medium grain:1 small grain. This indicates that the large grain is definitely inherited as a single allele. (**B**) Molecular linkage analysis by using small grain homozygous F_2 derived from the cross of Nipponbare × Inochinoichi.

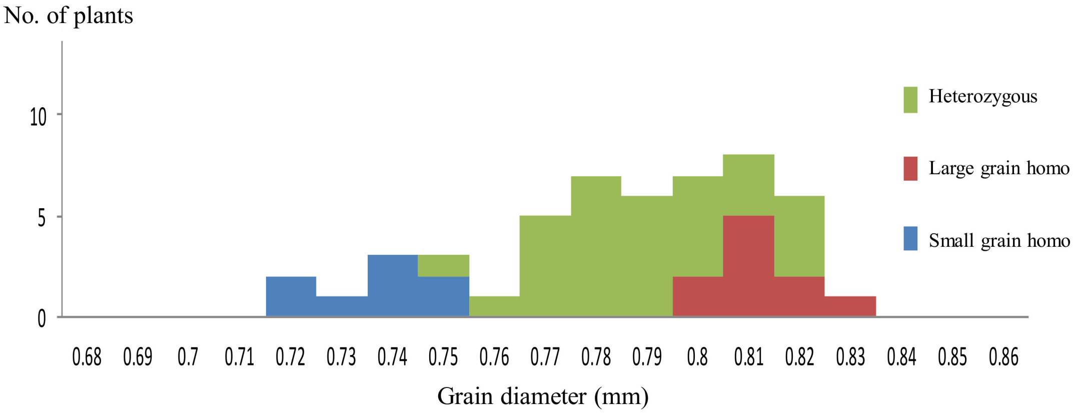


Figure 2. Distribution of mean value of grain diameter in 50 F_3 line derived from the cross, Koshihikari × Inochinoichi. F_3 lines segregated in a ratio of 10 large grain homozygous lines:32 heterozygous lines:8 small grain homozygous lines, which fit to the theoretical single gene ratio.

Table 1. Phenotypic expression of large-grain gene integrated isogenic Koshishikari.

		***GW2* Koshihikari**	***GW2 + d60* Koshihikari**	**Nigata Koshihikari**
Stem length (cm)		92	75	99
Weight of unpolished rice/1000 grains (g)		29.6 (×1.34)	28.8 (×1.31)	22.0
Polished rice	Taste value	80.0	80.0	81.0
	Protein	6.4	6.1	6.0
	Moisture	14.4	14.5	14.5
	Amylose	18.8	18.6	18.5
	Normal grain size	94.7	95.6	97.3
	Powdery grain	2.9	2.1	1.8
	Damaged grain	0.0	0.0	0.0
	Colored grain	0.0	0.0	0.0
	Split grain	0.2	0.3	0.3
	Crashed grain	2.2	2.0	0.6
	White degree	42.6	42.8	44.8

The large grain isogenic Koshihikari (BC_5F_2) showed the thousand kernel weight increased by 34%. Its taste score (80.0) was also equivalent to that of Niigata Koshihikari (81.0).

2.2. Candidate Region of Large Grain Gene

Using small grain homozygous F_2 segregants of a Nipponbare × Inochinoichi (Figure 1B), we genetically mapped the large grain gene by SSR and SNP markers across rice's 12 chromosomes. Our results showed that the recombinant values between DNA markers and the large grain gene were detected on chromosome 2. Namely, from the distal end of the short arm of chromosome 2, 21.7 at J521 (7.6 Mb), 17.5 at RM3390 (7.7 Mb), 15.2 at J527 (8.2 Mb), 19.6 at J529 (8.6 Mb), 28.3 at J536(9.1 Mb), 30.0 at RM6375 (9.6 Mb), and 34.5 at RM1358 (10.2 Mb), respectively (Figure 3). The RM3390-homozygous plant by the diagnosis, which is linked with *GW2*, showed that the mean grain area with Inochinoichi alleles was 23.3 mm^2, which is larger than 20.0 mm^2 in Koshihikari (Figure 1A).

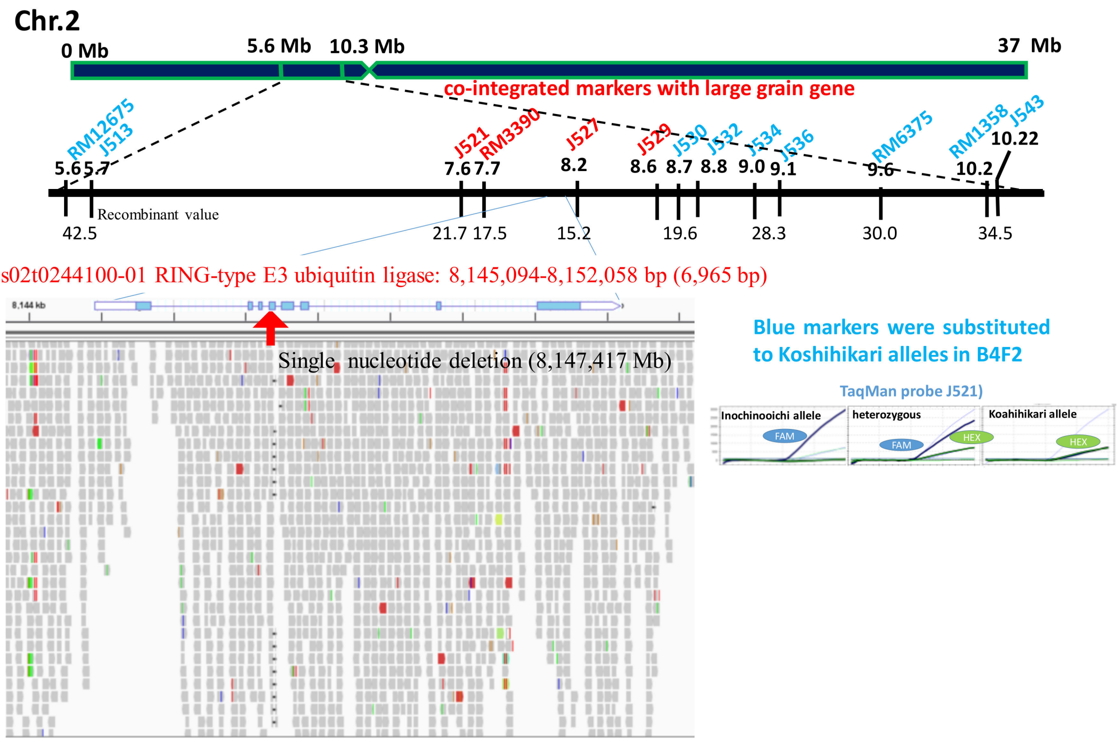


Figure 3. Identification of single nucleotide deletion in *GW2* responsible for large grain size of Inochinoichi. SNP allele-specific TaqMan probes were designed and labeled using the fluorescent dyes FAM or HEX. The real-time polymerase chain reaction (PCR) was used to amplify the allele-specific fluorescence. Blue DNA markers were substituted to Koshihikari alleles in B_4F_2. However, red DNA markers were inherited together with *GW2*.

2.3. Identification of DNA Variation Responsible for Large Grain Using Next-Generation Sequencing (NGS)

The read sequences of Koshihikari obtained by NGS were mapped using the Nipponbare genome as a reference sequence. The cover ratio was determined to be 99.05% and the mean depth was 32.43; Finally, a 372,912,445 bp long consensus sequence of the Koshihikari genome was constructed. Next, read sequences gained from the large grain isogenic Koshihikari (BC_4F_2) were mapped using the consensus sequence of Koshihikari as a reference sequence. In total, 187,159,213 reads were mapped, with a mapped read rate of 99.90%, a mean read length of 123.3 bp, and a 30.95× genome coverage.

Whole-genome sequencing detected a single nucleotide deletion (adenine) from the Koshihikari genome at the 8,147,417 bp position from the distal end of the short arm of chromosome 2 (Figure 3). This was the same as a single nucleotide deletion reported in the fourth exon of *GW2* (Os02g024410), the QTL gene responsible for grain width in the large grain Chinese rice WY3 [4]. *GW2* encodes a RING protein with E3 ubiquitin ligase activity, and a frame shift caused by a nucleotide deletion in this gene causes a loss-of-function [4]. *GW2* derived from Inochinoichi is 6,965 bp with 100% identical to that of WY3, i.e., a single deletion in the fourth exon. There are two SNPs that flank the coding region of the hydroquinone glucosyltransferase gene (Os02g0242900). Our results show that the gene responsible for the large grain size of Inochinoichi, a promising gene source for increasing yield, is *GW2*. DNA markers around *GW2*, namely from the distal end of the short arm of chromosome 2, RM12675(5.6 Mb), J513(5.7 Mb), J530(8.7 Mb), J532(8.8 Mb), J534(9.0 Mb), J536(9.1 Mb), RM6375 (9.6 Mb), and RM1358 (10.2 Mb) were substituted to Koshihikari alleles in the B_4F_2 (Figure 3). On the contrary, Inochinoichi alleles of J521 (7.6 Mb), RM3390 (7.7 Mb), J527 (8.2 Mb), J529 (8.6 Mb) were tightly inherited together with *GW2*.

2.4. Linkage Relationship Between Semidwarfing Gene d60 and Large Grain Gene GW2

The first backcross with Koshihikari was conducted with a large grain semi-dwarf plant (stalk length: 76 cm, grain diameter: 0.8 mm) as the nonrecurrent parent segregated in the F_2 between Koshihikari d60 (which was developed by integrating the Hokuriku 100-derived semidwarfing gene *d60* into the Koshihikari genome through seven times of backcrosses) and Inochinoichi (Figure 4A,B). Here, regarding the genetics of *d60*, in the F_1 hybrid (genotype *D60d60Galgal*) of Koshihikari (*D60D60galgal*) × Koshihikari d60(*d60d60GalGal*), male and female gametes having both *gal* and *d60* become gamete lethal and the pollen and seed fertility decrease to 75%. As a results, the F_2 progeny shows a unique mode of inheritance that is segregated into a ratio of 6 fertile long-culm (4*D60D60*:2*D60d60GalGal*: 2 partially sterile long-culm (*D60d60Galgal* = F_1 type):1 dwarf(*d60d60GalGal*) [16] (Figure S1). In this study in the BC_1F_2 of Koshihikari/(Koshihikari d60 × Inochinoichi F_2), the genotypic ratio for the *D60*/*d60* allele was 11 *d60* homozygous:26 partially sterile:75 long stem, which fit the theoretical ratio of 1:2:6 well ($\chi^2 = 0.22$, df = 2, $0.85 < P < 0.90$) (Figure 4C). However, in the relationship between grain area and stem length, this contrasts with the Koshihikari*1/Koshihikari/Inochinoichi BC_1F_2, where there was an extremely small number of large grain segregants, a large number of small grain segregants, and no large grain long-stem segregants (Figure 4C). In other words, for BC_1F_2 as a whole, the ratio of (*GW2* homozygous + heterozygous): *gw2* homozygous was 73:39. This ratio should be close to 5:4, which arises when *GW2* is completely linked with *D60*. Furthermore, while the segregation of the *GW2* allele in *d60* homozygous semi-dwarf plants was 10:1 for (large grain *GW2* homozygous + hetero): small grain *gw2* homozygous, in long-stem plants, the ratio of (*GW2* large grain homozygous + heterozygous): small grain *gw2* homozygous was 63:38. In other words, while large grain plants appeared at a higher rate in the semidwarf phenotype plants, they appeared at a lower rate in the long-stem phenotype plants (Figure 4C). Considerably deviated segregation in the *GW2* locus occurred while opposing to each genotype of *d60* allele. Furthermore, if *GW2* and *d60* are inherited independently, then the appearance rates of *GW2* homozygous long stem plants and that of *GW2* homozygous long stem partially sterile plants should be 6/36 (=(4*D60D60* + 2*D60d60*)/9 × 1*GW2GW2*/4)) and 2/36(=2*D60d60*/9 × 1*GW2GW2*/4), respectively. However, actually there were no *GW2* homozygotes among long stem or long stem partially sterile plants, respectively (Figure 4C). Thus, the fact that the segregation of the *GW2* allele was considerably deviated in each genotype of *d60* suggests linkage between *GW2* and *d60*.

A large grain semi-dwarf segregant in BC_1F_2 was backcrossed with Koshihikari d60 as a nonrecurrent parent to make Koshihikari d60//Koshihikari/((Koshihikarid60 × Inochinoichi)F_2) BC_2F_2, which segregated in a ratio of 3 large (grain area: 24.1–24.5 mm^2):14 medium:7 small (grain area: 19.6–22.0 mm^2) grain phenotypes = 1:2:1($\chi^2 = 2.00$, df = 2, $0.30 < P < 0.50$) (Figure 4D). Then, a *GW2d60* homozygous large grain semi-dwarf segregant in the BC_2F_2 was thirdly backcrossed with a Koshihikari d60 to make Koshihikari d60*2//Koshihikari/((Koshihikarid60 × Inochinoichi)F_2)BC_3F_2, which segregated in a ratio of 20 large (grain area: 27.1–32.5 mm^2):55 medium:19 small (grain area: 20.6–23.5 mm^2) grain phenotypes well fit to a 1:2:1 ratio ($\chi^2 = 2.74$, df = 2, $0.20 < P < 0.30$) (Figure 4E). Above all, the affect of the linkage between *GW2* and *d60* disappeared in the *d60* homozygous genetic background, so the large grain *GW2* locus segregated according to a Mendelian ratio of 1 *GW2* homozygous:2 heterozygous:1 *gw2* homozygous.

As above, a large grain semi-dwarf plant (BC_3F_2), namely Koshihikari d60 integrated with *GW2*, was fourthly backcrossed with Koshihikari to make a Koshihikari///Koshihikari d60*2//Koshihikari/((Koshihikari d60 × Inochinoichi)F_2 Gg genotype) BC_4F_2 generation (Figure 4A). The distribution of stem length in the BC_4F_2 was 17 (semi-dwarf 44.1–53.9 cm):36 (partially sterile):104 (long stem 53.9–74.1 cm) ≈ 1*d60d60GalGal*:2 *D60d60Galgal*:6(1*D60D60GglGal* + 2*D60D60Galgal* + 1*D60D60galgal* + 2*D60d60GalGal*)($\chi^2 = 0.01$, df = 2, $0.9 < P < 0.95$) (Figure 4F). Additionally, in each genotype for stem length, large (grain area: 22.9–24.6 mm^2), medium (grain area: 20.1–22.8 mm^2), and small grain (grain area: 17.9–20.0 mm^2) were segregated in the ratio of 11 large grain:6 medium grain:0 small grain in the short stem phenotype, 4 large grain:23 medium grain:9 small grain in the partially sterile phenotype, and 10 large grain:60 medium grain:34 small grain for the long stem phenotype. In other words, as is the case in the F_2 generation, while large grain genotypes segregated at a high rate in the *d60* homozygous semi-dwarf plants, the small grain genotypes segregated at a high rate in the long stem plants; this indicates a linkage relationship between *GW2* and *d60*. The recombinant value was calculated from the segregation ratio of 11 *GW2* homozygous:6 *GW2gw2*:0 *gw2* homozygous plants in the semi-dwarf phenotype, as follows: ((0 × 2 small grain + 6 heterozygous specimens)/17 × 2 total semi-dwarf specimens) × 100 = 17.6 (Figure 4F). The grain weight of the *GW2d60* homozygous large-grain semidwarf isogenic line (BC_4F_2) increased by approximately 31% over that of Koshihikari, and the stem length decreased by 26% (Table 1, Figure 5). The robust and high yielding isogenic Koshihikari line developed in this study by integration with *d60* and *GW2* is a new rice variety could bring about the "New Green Revolution" and overcome difficulties in this age of climate change and globalization.

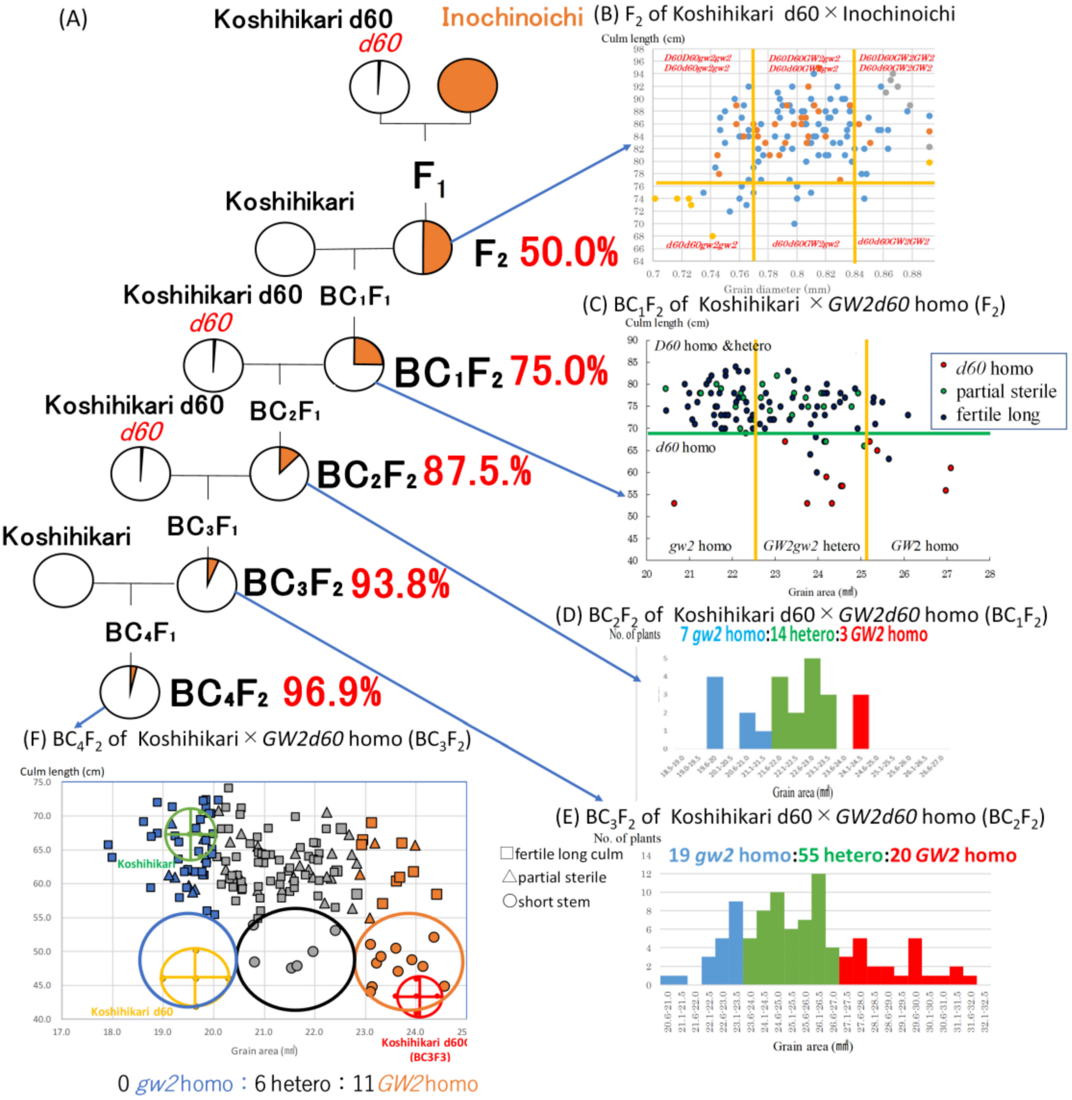


Figure 4. Linkage relationship between *GW2* and *d60* identified by the deviated segregation ratio through the continuous backcross process by using Koshihikari or Koshihikari *d60* as recurrent parents. (**A**) Backcross procedure combining *d60* and *GW2*. (**B**) Distribution of grain size and culm length in the F_2 of Koshihikari d60 × Inochinoichi. (**C**) Deviated segregation for grain size in each *d60/D60* genotype in BC_1F_2. (**D**,**E**) Mendelian segregation for grain size in the *d60* homo background in BC_2F_2 and BC_3F_2. (**F**) Recombination value (17.6) between *GW2* and *d60* was by segregation ratio in *GW2* allele in d60 homozygous in BC_4F_2.

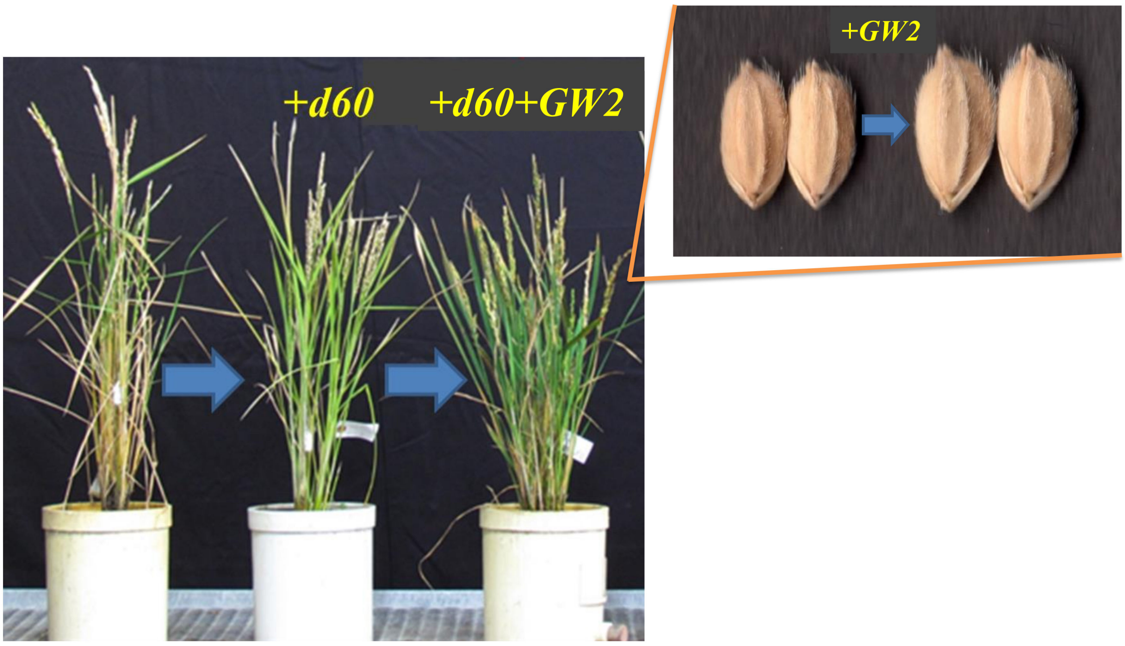


Figure 5. Phenotypic alteration of isogenic Koshihikari by integration of *d60* and *GW2*, derived from "Inochinoichi".

3. Discussion

Crops in Japan and the world are being damaged by climate change caused by global warming. In 2017, the Japanese government put an innovation policy into place to contribute to the world through the development of high-yield crops for a "New Green Revolution." The rice variety "Inochinoichi" is highly rated at both the consumer and producer levels, but the genes that control its large grain size have not been elucidated, so the buried genetic resources never have been used in breeding. In order to identify the genes that control large grain characteristics in Inochinoichi, we crossed a large grain rice variety with Koshihikari and, then, conducted a backcross with Koshihikari, which showed that the large grain size is controlled by a single gene. At the same time, we conducted a linkage analysis of a large grain gene using SSR markers and SNP markers across the 12 rice chromosomes that show polymorphisms between Nipponbare and Inochinoichi. As a result, we detected a linkage between the large grain gene and a DNA marker located 7.7 Mb from the distal end of the short arm of chromosome 2. Furthermore, we established a large grain isogenic line through five continuous back crosses with Koshihikari as the recurrent parent and then analyzed its whole genome using next-generation DNA sequencing. We successfully identified the target gene for large grains integrated in the genetic background of Koshihikari. There was no public information on the consensus sequence of Koshihikari, thus, we conducted a high-coverage whole-genome analysis, by first determining a consensus sequence for Koshihikari. We found that the responsible mutation for the large grain is a single nucleotide deletion located at 8.1 Mb from the distal end of the short arm of chromosome 2, in the *GW2* gene. *GW2* was identified as the causative gene at a QTL involved in grain width in large grain Japonica rice WY3; in Indica rice FAZ1, this allele (Os02g024410) encodes a new RING protein with E3 ubiquitin ligase activity [4]. E3 ubiquitin ligase is involved in the breakdown of proteins in the ubiquitin proteasome pathway [17]. RING-type E3 ubiquitin ligase can control seed development by catalyzing the ubiquitination of expansin-like 1 (EXPLA1), a cell wall-loosening protein that increases cell growth [18]. In contrast, the *GW2* gene in WY3 lost function due to a frame shift caused by a single nucleotide deletion. It has long been known that the size of the awn covering the grain is one of the factors determining grain size [19]. In a FAZ1 near-isogenic line with *GW2* from WY3 rice, the width of the awn was extended by 26.2% because of the increased cell number [4]. In this study, *GW2*, which was identified as a loss of function by the single nucleotide deletion, increased grain weight by 34% in the genetic background of Koshihikari.

In addition, it has been shown, by isolating genes affecting grain size, that cell number of awn is a factor that determines grain size and that grains get larger due to the loss-of-function of such genes [3–8]. *GS3* (Os03g0407400), the first gene to be reported, encodes a transmembrane protein consisting of 232 amino acids [3]. It was shown through a functional complementation test that a loss-of-function by a mutation in the second exon caused an increase in grain size [20]. A gene coding for a new nuclear protein, *qSW5* (Os05g0187500), makes grain width thin in Indica rice Kasalath [5]. The *qSW5* allele in Kasalath reduces cell number in the width direction, which narrows the awn, and in turn suppresses the elongation of endosperm cells, resulting in narrowed grain width. On the contrary, the Nipponbare *qSW5* allele has a 1212 bp deletion in its coding region, which results in a loss-of-function and allows for increase in grain width. *GS5* (Os05g0158500) found in Indica rice Zhenshan97 is a regulatory factor, which encodes serine carboxypeptidase and controls positively for grain size [6]; the difference of *GS5* expression affects grain size. *GW8* derived from a high-yielding rice variety HJX74 is responsible for the QTL involved in grain width [7]; the HJX74 allele also increases grain width. *GW8* is OsSPL16 (Os08g0531600), a gene that codes for a protein that positively controls cell proliferation. The *GW8* allele in HJX74 also increases cell number, enlarging the awn and, consequently, causing increase in grain size. These past studies have shown that loss-of-function mutations such as *GW2*, *GS3*, *GS5*, *qSW5*, and *GW8* cause an increase in cell number and subsequently, cause larger grain size.

TGW6 (Os06g0623700), a gene that increases the thousand kernel weight of Kasalath, has also been isolated [8]. *TWG6* increases the cell number in the endosperm. The *twg6* allele in Nipponbare codes for a protein that hydrolyzes indole-3-acetic acid (IAA)-glucose and synthesizes IAA, which promotes transition into the cell division stage. The Nipponbare *tgw6* allele reduces the cell number in the endosperm as well as grain length. The *TGW6* allele in Kasalath has a loss-of-function due to a single nucleotide deletion in its coding region, which means that grain length suppression via IAA does not occur, and grains elongate into a long phenotype characteristic of Indica rice. The distribution of *TGW6* was investigated in the genetic stock and the Kasalath *TGW6* allele was only found in one line of *Oryza perennis* and four local varieties in Indonesia [8]. This indicates that *TGW6* was not a target of selection and thought to have been discarded during the domestication process. The Japanese large grain rice variety Oochikara is reported to have the same *GW2* allele as those of WY3 and Inochinoichi. However, there is no historical relationship between Inochinoichi and Oochikara through their pedigree. Consequently, *GW2* is thought to be an extremely rare allele.

As discussed earlier, large grain genes that confer Indica rice its characteristics have been identified in Indica or Chinese varieties. However, breeding to enlarge grain size has never been fully explored in the Japanese leading variety Koshihikari. We now have the opportunity to utilize a large grain gene that was ignored during the domestication process to develop a high-yield rice variety. In our study, we showed evidence that an isogenic large grain Koshihikari integrated with *GW2* derived from Japanese rice via five times of backcrosses with Koshihikari has a 34% increased grain size compared to Koshihikari, and a taste score of 80.0, which is comparable to that of Niigata Koshihikari (81.0). Grain size-enlarged Koshihikari has the potential to become advantageously differentiated from US-made Koshihikari.

Rice yields around the world have doubled through the breeding of semi-dwarf varieties, which are representative of the Green Revolution in the mid-20th century, but the increase in yield is now leveling off. Additionally, there has been increased damage from lodging caused by severe weather events like the Western Japan floods and multiple typhoons under the recently intensified climate change; thus there is a need to develop rice plants that are sturdier and more robust. Also, with market liberalization through the Comprehensive and Progressive Agreement for Trans-Pacific Partnership (CPTPP) and negotiations for a Trade Agreement on Goods (TAG), there will soon be international competition in the rice market, so there is a need for low-cost and high-yielding rice. Thus, in order to make a breakthrough in high-yield breeding, which is dependent on a conventional semidwarfing gene *sd1*, we believe that to give rise to the New Green Revolution, semidwarfing should be used as a foundation with integrating/addition of genes related to high-yield including large grain and increased biomass. In this study, we combined the novel semidwarfing gene *d60* and the large grain gene *GW2* in the isogenic background of Koshihikari. In the BC_4F_2 by a cross Koshihikari × Koshihikari d60Gg (BC_3F_2), gametes with both *d60* and the gametic lethal gene *gal* are not viable, so the segregation ratio was (1*D60D60GglGal* + 2*D60D60Galgal* + 1*D60D60galgal* + 2*D60d60GalGal*):2*D60d60Galgal*:1*d60d60GalGal*. Through this genetic process, a linkage between *d60* and *GW2* on chromosome 2 was discovered with a recombination value of 17.6, according to the deviated segregation ratio of the large grain allele, namely 11 *GW2* homozygous:6 *GW2gw2*:0 *gw2* homozygous in the semidwarf *d60Gal* homozygotes. The integration of *GW2* and *d60* resulted in a 20.0% increase in grain size and a 19.2 cm reduction in stem length compared to Koshihikari, which is effective to reduce the lodging risk that accompanies the increased panicle/grain weight. We obtained genetic achievement of the effective integration of genes for large grain and robustness. We made a breakthrough in breeding, which conventionally relied only on a single gene *sd1*, by combining the semidwarfing gene with a gene for a high yields-related factor. Such a combined genotype could secure high yields while providing robustness required to withstand climate change. In other words, this idea could contribute to the New Green Revolution. We have designated the large grain isogenic line with a 34% increased grain weight due to *GW2*, and the large grain semi-dwarf isogenic line due to *GW2* + *d60*, which is capable of stable production with 31% increased grain weight/20 cm (26%) reduction in stem length, as “Koshihikari Suruga Gg” and “Koshihikari Suruga d60Gg”, respectively [21,22]. The two lines have been applied for plant variety registration. The taste and grain quality of these new varieties compared favorably with Niigata Koshihikari.

In China, grain size is being increased through the knockout of *GW2* by genome editing [23]. On the contrary, in the countries under the ratification of the Cartagena Act, including European countries and Japan, there are barriers to the social implementation of genetically modified plants. In our study, we developed a large grain semi-dwarf isogenic variety for stable production that withstands climate change though smart breeding. This was done by identifying the gene responsible for large grain size by NGS and, then, integrating it into the reference Koshihikari genome by continuous backcrossing, to finally construct a targeted gene-integrated isogenic genotype. The variety “Koshihikari Suruga d60Gg” has an epoch-making phenotype as it integrates the large grain gene *GW2*, which increases grain weight by 34%, as well as the semidwarfing gene *d60*, which reduces lodging risk, into the Koshihikari genome. This new variety could potential be a Super Koshihikari that could replace the leading variety Koshihikari which currently has a 36% share but suffers from considerable damage by abnormal weather. Our breakthrough rice plant type that integrates both semidwarfing and large grain phenotype should be a key resource for the New Green Revolution.

4. Materials and Methods

4.1. Genetic Analysis

We focused on the large grain characteristics of Inochinoichi as a genetic resource for high yields. First, we analyzed the mode of inheritance of grain size in the F_2 generation of Koshihikari × Inochinoichi. Furthermore, we conducted a progeny test using 50 F_3 lines derived from 50 randomly selected F_2 plants from Koshihikari × Inochinoichi. We then conducted five times of backcrosses with Koshihikari as a recurrent parent by using a large grain homozygous F3 plant (grain diameter: 0.75 mm) of Koshihikari × Inochinoichi as a nonrecurrent parent. We conducted genetic analysis of large grain in each $BCnF_2$ generation through five times of backcrosses, and the large grain homozygous segregants in each $BCnF_2$ were used as pollen parents for backcrosses with Koshihikari.

In order to develop an isogenic line that is both a semi-dwarf and large grains, the large grain semidwarf segregant in the F2 generation of Koshihikari d60 × Inochinoichi was used as a nonrecurrent parent to backcross with Koshihikari once, then Koshihikari d60 twice, then Koshihikari once again.

Koshihikari d60 is an isogenic Koshihikari integrated with semidwarfing gene *d60* derived from Hokuriku100 by seven times of continuous backcrosses, namely Koshihikari*7//(Koshihikari/Hokuriku 100 F_2) [16]. Genetic analyses of the large grain gene and *d60* were conducted in each backcross generation. The resulting genetically segregating populations were transplanted into Shizuoka University Ohya Field, and phenotypic traits (heading date, stem length, plant type, grain length, and grain width) of all plants were investigated. For grain characteristics, we evaluated the grain diameter at the early generation of F_2, and grain area (grain length/2 × grain width/2 × π) in the near isogenic backgrounds through backcrossing.

4.2. Mapping of Large Grain Gene by DNA Markers

In order to map the large grain gene, 1328 F_2 plants of Nipponbare × Inochinoichi were used. We sampled leaves from 371 small grain homozygous F_2 plants. The leaves were powdered while being frozen by liquid nitrogen using a Precellys 24 high-throughput bead-mill homogenizer (Bertin Technologies, Montigny-le-Bretonneux, France), and then genomic DNA was extracted using the cetyl trimethylammonium bromide (CTAB) method. A linkage analysis of large grain genes was conducted across the 12 rice chromosomes using SSR markers and SNP markers, which are polymorphic between Nipponbare and Inochinoichi. For the PCR reactions used to detect SSR markers, the mixtures were first heated to 95 °C for two minutes to denature the DNA, followed by 35 cycles of denaturing at 95 °C for 30 s, annealing at 50 °C or 55 °C for 30 s, and extension at 72 °C for 30 s. The SSR polymorphisms in the PCR products were analyzed by electrophoresis using a cartridge QIAxcel DNA Screening Kit (2400) in a QIAxcel electrophoresis apparatus (Qiagen, Hilden, Germany) at 5 kV for 10 min. SNP allele-specific TaqMan probes were designed and labeled using the fluorescent dyes FAM or HEX. The real time PCR reaction was used to amplify allele-specific fluorescence, by first heating the material to 95 °C for 30 s to denature the DNA, followed by 40 cycles of denaturing at 95 °C for 15 s, and annealing at 48 °C to 53.5 °C for 30 s.

4.3. Next-generation Sequencing (NGS) Analysis

Whole-genome sequencing of both Koshihikari and a large grain isogenic Koshihikari line (BC_4F_2), which was integrated with large grain gene derived from Inochinoichi by four times of back crosses into the genetic background of Koshihikari, were conducted. The leaves were powdered using a mortar and pestle while being frozen by liquid nitrogen. The DNA was then extracted using the CTAB method. Genomic DNA was fragmented and simultaneously tagged with the Nextera® transposome (Illumina, Rockville, MD, USA) such that the peak size of fragments was approximately 500 bp. Adapter sequences, including the sequencing primers, were synthesized in both ends via PCR. After the size selection of DNA fragments using magnetic beads, the DNA library was prepared following qualitative check by Bioanalyzer 2100 system (Agilent Technologies, Inc., Palo Alto, USA), and quantitative measurement by Qubit® Fluorometer (Life Technologies; Thermo Fisher Scientific, Inc., Waltham, MA, USA). The sequencing data were gained with paired-end reads using a HiSeq next-gen sequencer. The read sequences obtained were mapped using Burrows-Wheeler Aligner (BWA) software to the Nipponbare genome as a reference.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/20/21/5442/s1. Figure S1. Complementary gamete lethal between semidwarfing gene *d60* and gamete lethal gene *gal*.

Author Contributions: Conceptualization—M.T.; Methodology—M.T.; Investigation—M.T., S.Y., Y.U.; Resources—M.T.; Writing—Original draft preparation—M.T.; Writing—Review and editing—M.T.; Project administration, M.T.; Funding acquisition—M.T.

Funding: This work is founded by the Adaptable and Seamless Technology Transfer Program (A-STEP), Industry-Academia Joint Promotion Stage, High Risk Challenge Type grant by Japan Science and Technology Agency (JST) to Motonori Tomita, project ID14529973, entitled "Development of super high yield/large grain/early and late ripening rice suitable to the era of globalization and global warming by Next-Generation Sequencer/genome-wide analysis", since 2014 to 2018.

Acknowledgments: The author thanks Rieko Kamioka; Tomoaki Tsuchiya; Hideumi Ebata for their assistance of phenotypic investigation.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Khush, G.S. Green revolution: Preparing for the 21st century. *Genome* **1999**, *42*, 646–655. [CrossRef] [PubMed]
2. Chauhan, B.S.; Jabran, K.; Mahajan, G. *Rice Production Worldwide*; Springer International Publishing AG: Cham, Switzerland, 2017; p. 547.
3. Fan, C.; Xing, Y.; Man, H.; Lu, T.; Han, B.; Xu, C.; Zhang, Q. *GS3*, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. *Theor. Appl. Genet.* **2006**, *112*, 1164–1171. [CrossRef] [PubMed]
4. Song, X.J.; Huang, W.; Shi, M.; Zhu, M.Z.; Lin, H.X. A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nature Genet.* **2007**, *39*, 623–630. [CrossRef] [PubMed]
5. Shomura, A.; Izawa, T.; Ebana, K.; Ebitani, T.; Kanegae, H.; Konishi, S.; Yano, M. Deletion in a gene associated with grain size increased yields during rice domestication. *Nature Genet.* **2008**, *40*, 1023–1028. [CrossRef] [PubMed]
6. Li, Y.; Fan, C.; Xing, Y.; Jiang, Y.; Luo, L.; Sun, L.; Shao, D.; Xu, C.; Li, X.; Xiao, J.; et al. Natural variation in *GS5* plays an important role in regulating grain size and yield in rice. *Nature Genet.* **2011**, *43*, 1266–1269. [CrossRef] [PubMed]
7. Wang, S.; Wu, K.; Yuan, Q.; Liu, X.; Liu, Z.; Lin, X.; Zeng, R.; Zhu, H.; Dong, G.; Qian, Q.; et al. Control of grain size, shape and quality by *OsSPL16* in rice. *Nature Genet.* **2012**, *44*, 950–954. [CrossRef] [PubMed]
8. Ishimaru, K.; Hirotsu, N.; Madoka, Y.; Murakami, N.; Hara, N.; Onodera, H.; Kashiwagi, T.; Ujiie, K.; Shimizu, B.; Onishi, A.; et al. Loss of function of the IAA-glucose hydrolase gene *TGW6* enhances rice grain weight and increases yield. *Nature Genet.* **2013**, *45*, 707–711. [CrossRef] [PubMed]
9. Huang, X.; Kurata, N.; Wei, X.; Wang, Z.X.; Wang, A.; Zhao, Q.; Zhao, Y.; Liu, K.; Lu, H.; Li, W.; et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **2012**, *490*, 497–501. [CrossRef] [PubMed]
10. Japan Ministry of Agriculture, Forestry and Fisheries. TPP related information. Available online: http://www.maff.go.jp/j/kanbo/tpp/ (accessed on 19 September 2019).
11. Japan Meteorological Agency. Heisei 30 July 7 heavy rain and floods. Available online: https://www.data.jma.go.jp/obd/stats/data/bosai/report/2018/20180713/20180713.html (accessed on 19 September 2019).
12. Japan Meteorological Agency. 2018 (Heisei 30) Typhoons (Quick Report). Available online: https://www.jma.go.jp/jma/press/1812/21f/typhoon2018.pdf (accessed on 19 September 2019).
13. Japan Ministry of Agriculture, Forestry and Fisheries. Damage situation by Heisei 30 July 7 heavy rain and floods. Available online: http://www.maff.go.jp/j/saigai/ooame/20180628.html (accessed on 19 September 2019).
14. Japan Meteorological Agency. Abnormal Weather Risk Map. Available online: http://www.data.jma.go.jp/cpdinfo/riskmap/heavyrain.html (accessed on 19 September 2019).
15. Japan Meteorological Agency. Long-term changes in the number and frequency of short-term heavy rains observed at AMeDAS. Available online: http://www.jma.go.jp/jma/kishou/info/heavyraintrend.html (accessed on 19 September 2019).
16. Tomita, M. Combining two semidwarfing genes *d60* and *sd1* for reduced height in 'Minihikari', a new rice germplasm in the 'Koshihikari' genetic background. *Genet. Res. Camb.* **2012**, *94*, 235–244. [CrossRef] [PubMed]
17. Stone, S.L.; Hauksdóttir, H.; Troy, A.; Herschleb, J.; Kraft, E.; Callis, J. Functional analysis of the RING-type ubiquitin ligase family of Arabidopsis. *Plant. Physiol.* **2005**, *137*, 13–30. [CrossRef] [PubMed]
18. Choi, B.S.; Kim, Y.J.; Markkandan, K.; Koo, Y.J.; Song, J.T.; Seo, H.S. *GW2* functions as an E3 Ubiquitin ligase for rice Expansin-Like 1. *Int. J. Mol. Sci.* **2018**, *19*, 1904. [CrossRef] [PubMed]
19. Takeda, K.; Saito, K.; Yamazaki, K.; Mikami, T. Environmental response of yielding capacity in isogenic lines for grain size of rice. *Japan. J. Breed.* **1987**, *37*, 309–317. [CrossRef]

20. Takano-Kai, N.; Jiang, H.; Kubo, T.; Sweency, M.; Matsumoto, T.; Kanamori, H.; Padhukasaharam, B.; Bustamante, C.; Yoshimura, A.; Doi, K.; et al. Evolutionary history of *GS3*, a gene conferring grain length in rice. *Genetics* **2009**, *182*, 1323–1334. [CrossRef] [PubMed]
21. Official Gazette No. 7080, 14 Aug. 2017, Ministry of Agriculture, Forestry and Fisheries. Plant varietal registration application No. 32364, "Koshihaikri Suruga Gg" 2017, Tokyo, Japan. Available online: http://www.hinshu2.maff.go.jp/vips/cmm/apCMM111.aspx?SHUTSUGAN_NO=32364&LANGUAGE=English (accessed on 14 August 2017).
22. Official Gazette No. 7080, 14 Aug. 2017, Ministry of Agriculture, Forestry and Fisheries. Plant varietal registration application No. 32365, "Koshihikari Suruga d60Gg" 2017, Tokyo, Japan. Available online: http://www.hinshu2.maff.go.jp/vips/cmm/apCMM111.aspx?SHUTSUGAN_NO=32365&LANGUAGE=English (accessed on 14 August 2017).
23. Mishra, R.; Joshi, R.K.; Zhao, K. Genome editing in rice: Recent advances, challenges, and future implications. *Front Plant Sci.* **2018**, *9*, 1361. [CrossRef] [PubMed]

International Journal of
Molecular Sciences

Article

Haplotype Networking of GWAS Hits for Citrulline Variation Associated with the Domestication of Watermelon

Vijay Joshi [1,†], Suhas Shinde [2,†], Padma Nimmakayala [2,*], Venkata Lakshmi Abburi [2], Suresh Babu Alaparthi [2], Carlos Lopez-Ortiz [2], Amnon Levi [3], Girish Panicker [4] and Umesh K. Reddy [2,*]

[1] Department of Horticultural Sciences, Texas A&M University, and Texas Texas A&M AgriLife Research and Extension Center, Uvalde, TX 78801, USA; Vijay.Joshi@ag.tamu.edu

[2] Department of Biology and Gus R. Douglass Institute, West Virginia State University, Institute, WV 25112, USA; suhas.shinde@wvstateu.edu (S.S.); vabburi@wvstateu.edu (V.L.A.); salaparthi@wvstateu.edu (S.B.A.); carlos.ortiz@wvstateu.edu (C.L.-O.)

[3] U.S. Department of Agriculture-Agricultural Research Service, U.S. Vegetable Laboratory, Charleston, SC 29414, USA; Amnon.Levi@ars.usda.gov

[4] Center for Conservation Research, Alcorn State University, 1000 ASU Drive, Lorman, MS 39096, USA; panicker@alcorn.edu

[*] Correspondence: padma@wvstateu.edu (P.N.); ureddy@wvstateu.edu (U.K.R.); Tel.: +1-304-766-3258 (P.N.); +1-304-766-3066 (U.K.R.)

[†] Authors contributed equally.

Received: 17 September 2019; Accepted: 26 October 2019; Published: 29 October 2019

Abstract: Watermelon is a good source of citrulline, a non-protein amino acid. Citrulline has several therapeutic and clinical implications as it produces nitric oxide via arginine. In plants, citrulline plays a pivotal role in nitrogen transport and osmoprotection. The purpose of this study was to identify single nucleotide polymorphism (SNP) markers associated with citrulline metabolism using a genome-wide association study (GWAS) and understand the role of citrulline in watermelon domestication. A watermelon collection consisting of 187 wild, landraces, and cultivated accessions was used to estimate citrulline content. An association analysis involved a total of 12,125 SNPs with a minor allele frequency (MAF)>0.05 in understanding the population structure and phylogeny in light of citrulline accumulation. Wild egusi types and landraces contained low to medium citrulline content, whereas cultivars had higher content, which suggests that obtaining higher content of citrulline is a domesticated trait. GWAS analysis identified candidate genes (*ferrochelatase* and *acetolactate synthase*) showing a significant association of SNPs with citrulline content. Haplotype networking indicated positive selection from wild to domesticated watermelon. To our knowledge, this is the first study showing genetic regulation of citrulline variation in plants by using a GWAS strategy. These results provide new insights into the citrulline metabolism in plants and the possibility of incorporating high citrulline as a trait in watermelon breeding programs.

Keywords: citrulline; genome-wide association study; haplotype; watermelon; *acetolactate synthase*; *ferrochelatase*

1. Introduction

Non-protein amino acids present in legumes, fruits, seeds, and nuts are ubiquitous in the human diet. Besides containing several health-promoting bioactive compounds, fruits accumulate substantial amounts of free non-protein amino acids. With a diversity of phytochemicals such as carotenoids, flavonoids, and triterpenoids, watermelon (*Citrullus lanatus* var. *vulgaris*) fruits also accumulate a substantial amount of a non-protein amino acid—citrulline. Scientific studies have demonstrated several health benefits of citrulline, such as anti-atherosclerotic (hardening of the arteries) effects, reduction of aortic blood pressure and stiffness in individuals with hypertension and cardiovascular diseases, improving lipid profiles by lowering cholesterol, lowering inflammation, and increasing athletic performance [1–5]. Additionally, many recent reviews have covered the clinical impacts of citrulline administration on human health in detail [6–8]. Consumption of watermelon has been shown to increase plasma arginine concentrations in adults [9,10]. Watermelons are readily available in most parts of the world, including dry and hot areas of the African continent, where most fruits would not thrive. As watermelon is a nutrient-dense fruit, it is recommended as part of a healthy meal plan as per the U.S. Department of Agriculture's MyPlate guidelines. In plants, citrulline has been suggested to have a functional role in the nitrogen transport and maintenance of cellular osmolarity during abiotic stresses in plants [11,12]. Members of the Cucurbitaceae family are generally considered to contain relatively large amounts of free citrulline, although watermelon accumulates the highest quantities [13]. Citrulline content in watermelons is spatially and developmentally regulated, with the highest values occurring at fruit maturity [14–18]. Unlike its investigation in plants, citrulline regulation has been extensively studied in the mammalian, prokaryotic, and yeast systems [19]. In the absence of functional nitric oxide synthase (NOS), citrulline in plants is synthesized as a metabolic intermediate during arginine biosynthesis by using carbamoyl phosphate and ornithine. Several studies reported the presence of genotypic variation for citrulline content in a selected set of cultivated watermelon varieties [11,16,18,20,21]. A moderate to high range of heritability [21,22] for citrulline content within cultivated watermelons implies its possible genetic improvement with selective breeding.

A complex compartmentalized network of genes coordinates several metabolic pathways to regulate amino acid metabolism in plants. To understand the molecular regulation and genetic inheritance of amino acids, these networks can be unraveled with the availability of whole-genome sequences and other functional resources. Over the last 2 decades, genome-wide association studies (GWASs) have continued to be the favorite tool to identify causal genetic loci of quantitative or qualitative traits in diverse germplasm collections exploiting evolutionarily conserved recombination events. Several GWAS studies have successfully identified candidate genes by using primary metabolites such as amino acid profiles [23–25]. As a naturally rich source of citrulline and with the advent of a newly available genome [26], watermelon could serve as an excellent model to study the evolutionary, biochemical, and molecular determinants of citrulline metabolism and regulation. We previously analyzed genome-wide diversity in watermelon by using a large set of SNPs from accessions collected around the world [27,28] to identify genetic loci controlling traits such as fruit firmness, trichome density and length, fruit length, width, rind thickness and soluble solids [29].

However, we lack information on the genetic basis for variability in citrulline content in watermelon fruits. Our studies with GWAS in watermelon will allow us to estimate population structure and linkage disequilibrium (LD), connecting the variation in the genome with the citrulline content in watermelon germplasm. To understand the evolutionary significance of citrulline in watermelon domestication and its possible genetic regulation, this current study involved (1) characterizing genotypic and phenotypic diversity for citrulline content in a watermelon diversity panel representing wild types, landraces and cultivars; (2) identifying candidate genes significantly associated with citrulline variation; and (3) validating the role of selected genes showing significant association by quantitative real-time-PCR analysis.

2. Results

2.1. Phenotypic and Geographic Variation in Citrulline Content

The current study examined the collection of 144 *Citrullus. lanatus* var. *vulgaris* (sweet watermelons) accessions, 34 semi-wild types (hereafter called landraces), and 9 accessions belonging to *Citrullus. mucasospermus* (egusi) (Table S1). Our collection contained plant introduction (PI) accessions from

Africa, Asia, Europe, North America, and South America. African types were from Algeria, Botswana, Egypt, Ethiopia, Ghana, Kenya, Liberia, Mali, Nigeria, Senegal, South Africa, Sudan, Zaire, Zambia, and Zimbabwe. The detailed distribution and quantitative variation of citrulline content of accessions used in this study are in Figure S1 and Table S1. The mean, range, and distribution of citrulline content in cultivars, landraces, and egusi types are shown in Figure 1A. The means of citrulline content were 11.08, 7.35, and 0.8 mg/g in cultivars, landraces, and egusi types respectively, which indicates high citrulline content as a feature of cultivars. The average citrulline content in the accessions from North and South America was significantly higher ($p \leq 0.0001$) than the accessions from Africa (Figure 1B). Among the selected accessions, mean citrulline content was 10.0 ± 0.04 mg/g (ranging from 0.10 to 47.3 mg/g). The dispersion of free citrulline across the accessions was leptokurtic (kurtosis = 3.27) and asymmetrical (skewness = 1.8). Accessions PI 559993, PI 426625, PI 560020, and PI 526238 had the lowest free citrulline in flesh and Garrisonian, Cole's Early, and PI 442826 had the highest. Restricted maximum likelihood/best linear unbiased prediction (REML/BLUP) [30,31], variance component estimation revealed significant variation in citrulline content within accessions. The estimates of heritability based on REML analysis were high (83%) for citrulline content, whereas the genetic gain at 5% selection intensity was 9.5%.

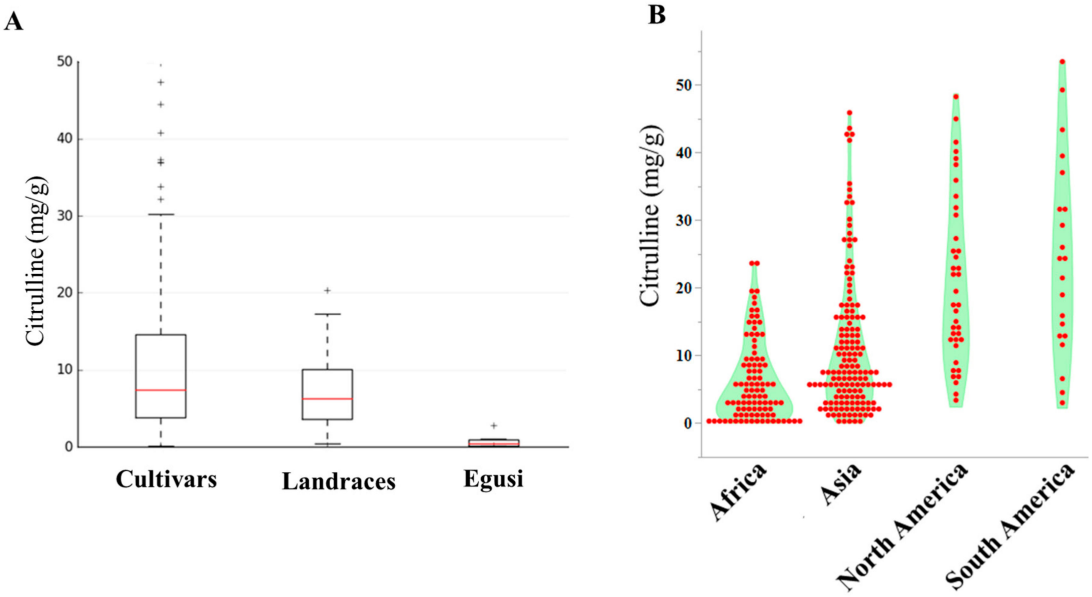


Figure 1. Citrulline content in watermelon accessions. (**A**) Box plots showing the range, mean, and distribution of citrulline content in cultivars, landraces, and egusi types. (**B**) Violin-scaled contour map showing world geographic variation in citrulline content across the accessions.

2.2. Population Structure of Various Accessions Based on Citrulline Content

To examine divergence across accessions during evolution, analysis of population structure and phylogenetic relationships and PCA were carried out [27]. Using the SNP dataset, we constructed a PCA with the first and second principal components PC1 (25.24) and PC2 (4.81) that separated egusi, landraces, and sweet watermelons (Figure 2A). This PCA also separated low, medium, and high citrulline content types (Figure 2B). Sweet watermelon types contained the highest citrulline content as compared with their ancestral progenitors. Egusi types and landraces contained low to medium citrulline content, whereas cultivars showed increased content, which indicates that high citrulline content is a domesticated trait.

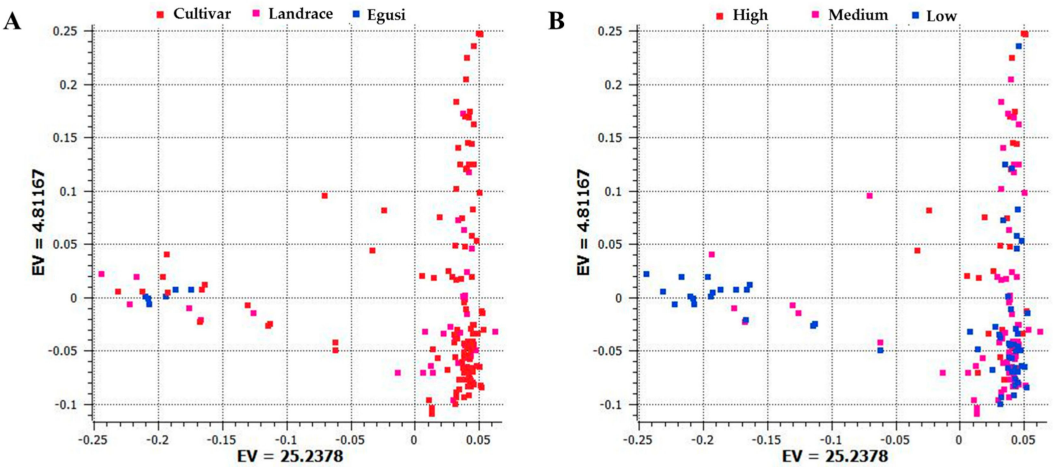


Figure 2. Principal component analysis (PCA) based on the first two components showing the distribution of (**A**) cultivars, landraces, and egusi types; (**B**) and low, medium, and high citrulline content in 187 watermelon accessions by using 1410 single nucleotide polymorphisms (SNPs) generated by genotyping by sequencing. Each dot represents an accession. EV indicates the percentage of explained variance.

Using the final SNP dataset, we constructed an unrooted neighbor-joining (NJ) tree to infer phylogenetic relationships and understand the distribution of citrulline content across the *C. lanatus* accessions (Figure 3A,B). One group (colored blue in the tree) represents the entire egusi collections. The pink- and red-colored clades represent sweet watermelons and landraces (most from South Africa), many with a hard rind and white flesh resembling an intermediate between egusi and sweet watermelon. This study indicated that most low to medium citrulline-content types are egusi and landraces.

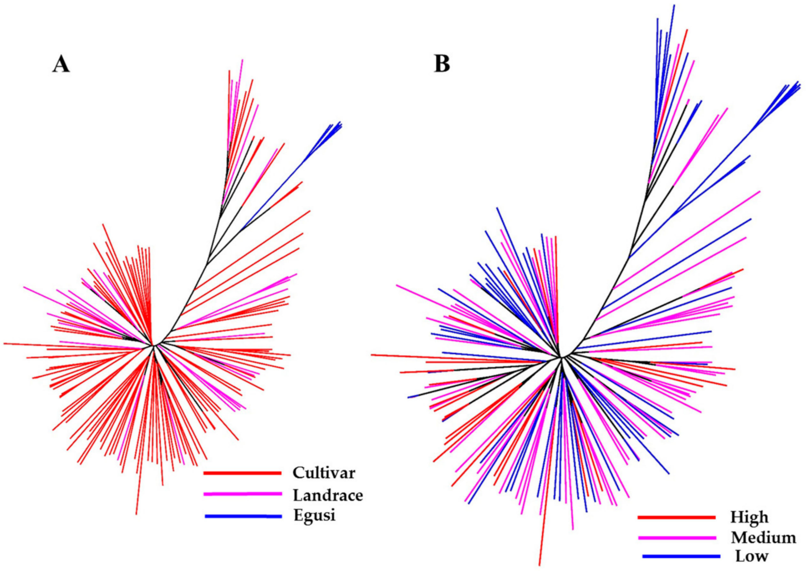


Figure 3. Genetic relationship between a set of 187 watermelon accessions. Neighbor-joining (NJ) tree constructed with 1410 high-quality SNPs explains most of the genetic structure of watermelon germplasm by (**A**) type and (**B**) citrulline content. Accessions in blue-, pink-, and red-colored clades are egusi-types, sweet watermelons, and landraces, respectively. In (**B**), blue, pink, and red clades represent low, medium, and high citrulline content, respectively.

2.3. Genome-wide Association Study to Locate Quantitative Trait Loci for Citrulline Content

We used a GWAS with 12,125 SNPs to identify alleles that affect citrulline content (Figure 4); individual SNP associations along with the details of major and minor allele frequencies and magnitude of associations are in Table 1 and detailed annotations for all associated SNP markers are in Table S2. We found 12 SNPs associated with citrulline content (Tables 1 and 2). Significantly associated SNPs for citrulline content were S02_33508197, S02_33508131, S02_28460679, S04_19161720, S04_10803195, S04_19161725, S06_30930976, S06_30991451, S07_12838412, S07_6258382, S09_9172194, and S10_19726131 and were found in ferrochelatase, F-box/LRR-repeat protein 2, Golgi SNAP receptor complex member 2, DNA polymerase I/DNA polymerase I, acetolactate synthase, BAG family molecular chaperone regulator 1, TLC ATP/ADP transporter, protein of unknown function, and phototropic-responsive NPH3 family protein genes, respectively. Biological roles, molecular processes, and the cellular location of these genes are in Table 2.

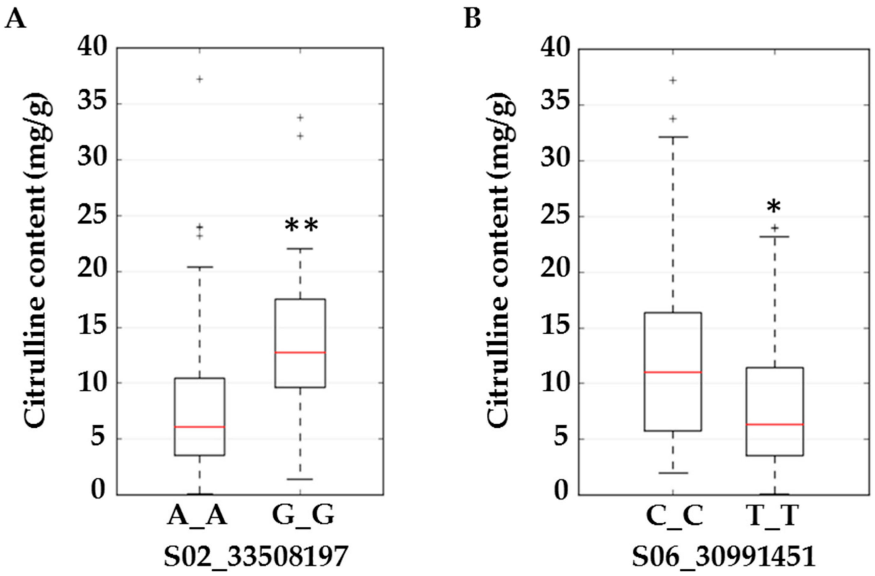


Figure 4. Boxplots for citrulline content (mg/g) in flesh tissue at SNP S02_33508197 located in the intron of *ferrochelatase* (**A**) and S06_30991451 located in an exon of *acetolactate synthase* (**B**). Significant differences (based on the Kruskal–Wallis test) with $p \leq 0.01$ and $p \leq 0.05$ are marked with two (**) and one (*) asterisks respectively.

We selected *ferrochelatase* (*FC*) and *acetolactate synthase* (*ALS*) genes for further validation. Allelic effects of SNPs located in these genes are in Figure 4. S02_33508197 is located in the intron of *FC*. Allele frequencies for S02_33508197 were 0.83 for AA and 0.17 for the minor allele GG. Average citrulline content was 5 mg/g for the AA-containing genotype and 12.5 mg/g for the GG-containing genotypes (Figure 4A). S06_30991451 is located in an exon of *ALS* and a non-synonymous mutation causing N→S with allele frequencies for TT and CC of 0.79 and 0.21, respectively. Allelic effects for major (TT) and minor (CC) alleles can be noted from the box plot (Figure 4B). Strong LD is noted around the associated SNPs in Figure 5A,C. LD around these two genes was further confirmed from a robust set of 1250 accessions (Figure 5B,D) in a recently published study [26].

Table 1. The significant non-synonymous SNPs associated with citrulline content in watermelon flesh.

Marker	Locus1	*p*-Value	−log10	Regression Beta	Beta Standard Error	FDR	Minor Allele Frequency
S02_33508197	ClCG02G018770	0.00	3.18	6.95	1.99	0.10	0.17
S02_33508131	ClCG02G018770	0.00	3.34	6.42	1.78	0.10	0.19
S02_28460679	ClCG02G014160	0.00	3.33	6.57	1.82	0.09	0.19
S04_19161720	ClCG04G005470	0.00	4.29	6.79	1.61	0.09	0.25
S04_10803195	ClCG04G002830/ClCG04G002840	0.00	3.06	6.11	1.79	0.12	0.21
S04_19161725	ClCG04G005470	0.00	4.29	6.79	1.61	0.05	0.25
S06_30930976	ClCG06G017840	0.00	4.21	7.28	1.75	0.04	0.21
S06_30991451	ClCG06G017910	0.00	3.36	6.17	1.71	0.15	0.21
S07_12838412	ClCG07G006720	0.00	3.44	−4.70	1.28	0.16	0.49
S07_6258382	ClCG07G004850	0.00	3.25	−4.85	1.37	0.10	0.42
S09_9172194	ClCG09G009500	0.00	3.34	6.28	1.74	0.14	0.21
S10_19726131	ClCG10G008990	0.00	3.24	5.72	1.62	0.09	0.28

Abbreviations: FDR- false discovery rate.

Table 2. Gene ontology classification of the genes associated with significant SNPs.

Marker	Locus and SNP Location	Gene Annotation	Molecular Function	Biological Process	Cellular Component	Ma/Mi	Amino acid change
S02_33508197	ClCG02G018770-Intron	Ferrochelatase	Ferrochelatase activity	Heme biosynthesis	Cytoplasm	A/G	-
S02_33508131	ClCG02G018770-Intron	Ferrochelatase	Ferrochelatase activity	Heme biosynthesis	Cytoplasm	G/A	-
S02_28460679	ClCG02G014160-Exon	F-box/LRR-repeat protein 2	Protein binding	Protein destabilization	Nucleus	T/C	G→G
S04_19161720	ClCG04G005470-3′UTR	Golgi SNAP receptor complex 2	SNAP receptor activity	Transport	Golgi apparatus	C/T	F→F
S04_10803195	ClCG04G002830/ ClCG04G002840 Intergenic	DNA polymerase I	DNA binding	Regulation of transcription	Nucleus	C/A	-
S04_19161725	ClCG04G005470-3′UTR	Golgi SNAP receptor complex 2	SNAP receptor activity	Transport	Golgi apparatus	T/G	R→M
S06_30930976	ClCG06G017840-Intron	SAP domain-containing protein	DNA binding	Regulation of translation	Nucleus	C/A	
S06_30991451	ClCG06G017910-Exon	Acetolactate synthase	Valine biosynthesis	BCAA biosynthesis	Chloroplast	T/C	N→S
S07_12838412	ClCG07G006720-Exon	BAG family molecular chaperone regulator 1	Protein binding	Defense response to fungus,	Plasmodesma	A/C	L→W
S07_6258382	ClCG07G004850-Intron	TLC ATP/ADP transporter	ATP:ADP antiporter activity	Transport	Membrane	C/A	-
S09_9172194	ClCG09G009500-Intron	Protein of unknown function	–	–	–	T/A	-
S10_19726131	ClCG10G008990-Exon	Phototropic-responsive NPH3 family protein	Protein binding	Protein ubiquitination	–	G/C	W→C

Abbreviations: LLR- leucine-rich repeat; SAP- after SAF-A/B, Acinus and PIAS; BAG- B-cell lymphoma 2 associated athanogene 1; TLC- thin layer chromatography; NPH3- nonphototropic hypocotyl 3; SNAP- soluble NSF (N-ethylmaleimide-sensitive fusion protein) attachment protein; Ma/Mi- Major/Minor allele. The arrow (→) indicates amino acid change. 2.4. Functional Validation of Associated Genes.

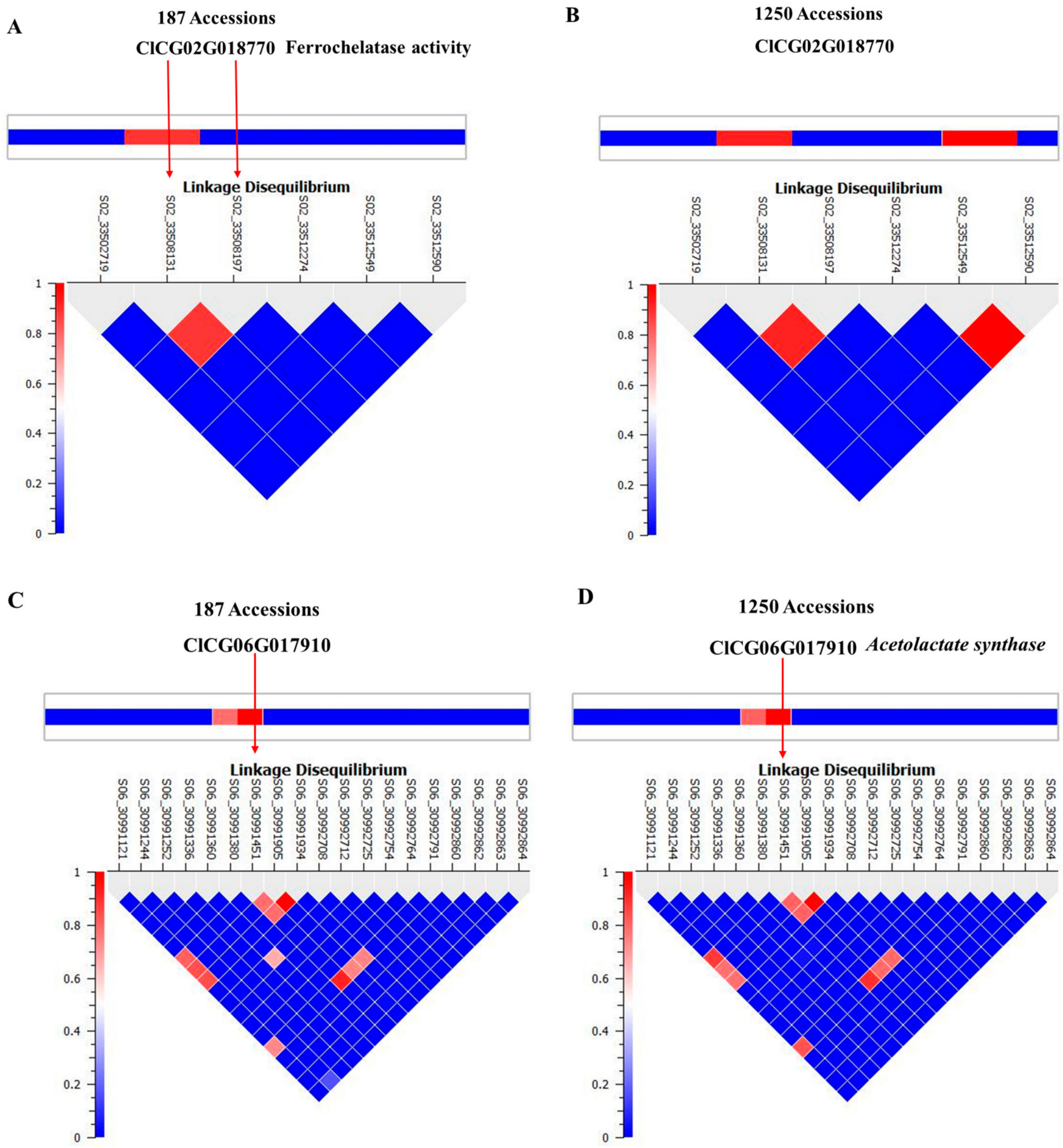


Figure 5. Linkage disequilibrium structure and implicated genomic regions for SNPs aligned with: (**A**,**B**) *ferrochelatase*; (**C**,**D**) *acetolactate synthase*.

The watermelon genome has a single copy of *ALS* gene (a large subunit) and two genes that encode *ALS* small subunits (ClCG09G014670 and ClCG03G010140, putative *ALS* small subunits 1 and 2, respectively). The ALS enzyme facilitates the first step in the biosynthesis of branched-chain amino acids (BCAAs; valine, leucine, and isoleucine) in microbes and plants. Moreover, ALS enzyme is inhibited by a group of imidazolinone and sulfonylurea herbicides [32–34], thereby preventing biosynthesis of BCAAs. We used real-time quantitative PCR to validate the association of the ALS gene(s) in selected high and low citrulline-content watermelon accessions (Figure 6A). The relative expression of *ALS* mRNA was significantly upregulated in flesh tissues of high citrulline-content watermelon accessions and downregulated in PI560020, with low citrulline content (Figure 6B). In summary, the expression of *ALS* showed strong association with citrulline content in watermelon.

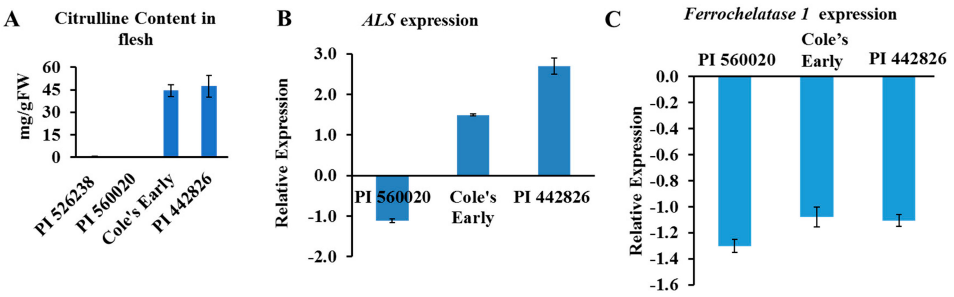


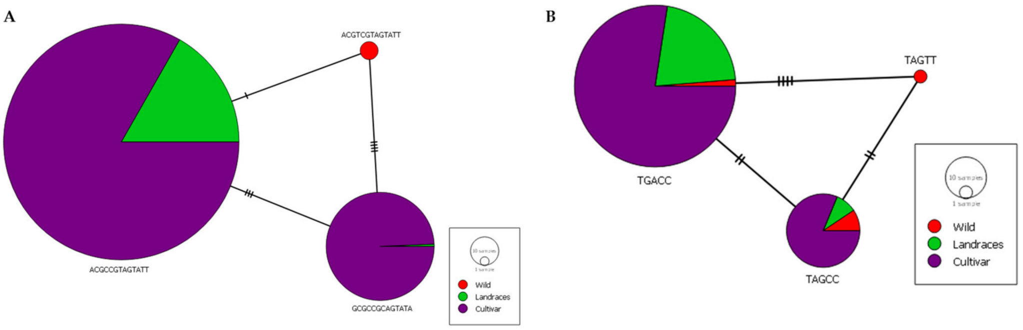
Figure 6. (**A**) Citrulline content and expression of candidate genes in flesh of selected watermelon accessions. Data are means ± SD (*n* = 3). (**B**) Expression of *acetolactate synthase* (*ALS*) and (**C**) *Ferrochelatase 1* genes in flesh. PI 526238 accession, with low citrulline content, was used as a calibrator for relative expression in (**A**). Gene expression was normalized to that of *Actin 8*. Data are means ± SD (*n* = 3).

The watermelon genome database search revealed two genes coding for FC enzyme: ClCG02G018770 (appeared in GWAS) and ClCG08G016940. The BLAST search against the *Arabidopsis thaliana* genome revealed that the watermelon ClCG08G016940 and ClCG02G018770 genes are orthologues of AT5G26030 and AT2G30390, labeled *FC1* and *FC2*, respectively.

The mRNA abundance for *FC1* was quantified by qRT-PCR, and we wondered whether *FC1* transcript level was associated with citrulline level in watermelon. Reduced *FC1* transcript abundance was seen in all three accessions (Figure 6C). However, the reason as to why the *FC1* expression was low in accessions with both high and low citrulline content is elusive. We hypothesize that FC1 may act in a spatiotemporal and growth dependent manner. Our data suggest that the *ALS* and *FC* genes may have a strong association with citrulline accumulation. However, future analysis is required to characterize the functions of FC and ALS proteins in citrulline-BCAA metabolism and NO turnover in watermelon.

2.4. Haplotyping and Network Analysis of Acetolactate Synthase and Ferrochelatase

With the four segregating sites in the LD block in *ALS*, we could build a network of haplotypes for egusi (wild), landraces, and cultivated watermelon (Figure 7A). Haplotype ACGTCGTAGTATT had undergone a single nucleotide change to form a landrace haplotype (ACGCCGTAGTATT). GCGCCGCAGTATA has three segregating sites as compared with the landrace haplotype and four segregating sites as compared with the wild haplotype. We noted Tajima's D as 0.88 and nucleotide diversity of 0.07, indicating a high degree of positive selection around this gene. Contrastingly, for the other causal gene *FC*, we noted reduced nucleotide diversity of 0.01 and negative Tajima's D (−0.221) indicating purifying selection in evolution (Figure 7B).



Chromosome No.:	6
Nucleotide Diversity:	0.0668965
Number of segregating sites:	4
Tajima's D:	0.883926

Chromosome No.:	2
Nucleotide Diversity:	0.00120651
Number of segregating sites:	4
Tajima's D:	-0.221883

Figure 7. Haplotyping and network analysis of (**A**) *acetolactate synthase* and (**B**) *ferrochelatase* on chromosomes 6 and 2, respectively.

3. Discussion

The current research aimed to analyze citrulline content evolution by using GBS-generated SNPs across the global collections of sweet watermelon, landraces, and egusi types. Watermelons were cultivated in the northeast of Africa around 5000 years ago [35]. Pictures on Egyptian frescos and seeds found in the grave of King Tutankhamun indicated that the flesh of watermelons was still white and tasted bland [36]. These bland vegetable-kind watermelons were cultivated for thousands of years before sweet watermelons arose [37]. Depictions of cut watermelons throughout the centuries (Giovanni Stanchi 1608 to c. 1675) deviated from the modern watermelons. When explored through flesh of landraces, one can find the evidence for these anthropological records. Because of this, in the current study, landraces are named as semi-wild as the sweet watermelons are quite different. Modern watermelon provides a large amount of water and nutrients, such as sugars, carotenoids, lycopene, minerals, and amino acids, including citrulline. Although citrulline is the most abundant free amino acid in watermelon [12], because of the crop's commercial value, genetic mapping efforts have mostly focused on sugar content [38,39]. During domestication, dessert watermelons were mostly selected for qualitative traits such as sweetness, flesh color, and rind pattern [40]. Although primitive dessert watermelon landraces, with naturally low sugar content, were expected to be valuable sources of bioactive compounds [41], our study demonstrates that citrulline content is, in general, lower in landraces and egusi types. Modern cultivars from the Americas that are rich in citrulline may have acquired the trait inadvertently during the selection for sweetness or the trait might have an unknown role in adaptation. A positive correlation of sugar content with citrulline was demonstrated in modern watermelon cultivars [18,21,22]. The distinctness of American, European and Asian ecotypes has been described in watermelons [29,42,43], but an independent parallel phenotypic evolution for citrulline during domestication cannot be ruled out [44].

3.1. Genetic Characterization of ALS and FC Locus Haplogroups

We genotyped 187 watermelon accessions with a mix of cultivars, landraces and egusi types predominantly collected from Africa, Europe, North America, and Asia with ~12K SNPs. A subset of this data allowed for estimating genomic diversity across the LD around the associated locus among various groups, constructing an NJ tree, resolving population structure, estimating chromosome-wise LD patterns and understanding the extent of population differentiation and haplotype networking in terms of citrulline content in watermelons. This study revealed high nucleotide diversity and Tajima's D for the *ALS* locus, which showed a strong association with citrulline content. In this study sweet watermelon in South Africa, and most landraces of South Africa clustered with sweet watermelon, which strengthens the argument that the Kalahari Desert could be the center of origin [45–47]. Alternatively, egusi types of northeast Africa could be the progenitors of landrace watermelon types because they share segregating sites in the both haplotype networks, thereby indicating a stepwise evolutionary pattern and probably is the second event of domestication. Similar to our observation in this study of haplotype networking in the *ALS* and *FC* loci between egusi and sweet watermelon, Chomicki and Renner [48] reclassified egusi (var. mucasospermus) and sweet watermelons (var. vulgaris) as two different species naming egusi as *C. mucasospermus*. Knowledge of causal genes underlying domestication traits and distribution of functionally diverse alleles in different watermelon populations may be required to better demonstrate the evolutionary route of dessert watermelons [26]. The current SNP-based analysis revealed citrulline accumulation as an important step for the domestication process of sweet watermelons. In this study, we identified 105 private SNPs segregating in sweet watermelons that was not found in related wild species. Such SNPs are valuable because they can be of adaptive importance and would be of immense use for generating passport information for enhancing nutritional traits such as citrulline.

3.2. Plausible Role of Candidate Genes Identified in Citrulline Biosynthesis

In eukaryotic cells, FC proteins are located in the mitochondrial inner membrane and facilitate the final step in the heme biosynthetic pathway, inserting a ferrous iron into protoporphyrin IX to yield the heme [49]. In animals, arginine can be a substrate for citrulline biosynthesis via the action of NOS [50–52]. However, in plants, arginine-dependent NOS-like proteins are not functionally validated; therefore, the synthesis of citrulline directly from arginine in plants remains debated [18,53,54]. In animals, NOS proteins are soluble hemoproteins that facilitate the conversion of L-arginine to citrulline and NO [55]. NO, a crucial signaling molecule also called a gaseous hormone in plants and animals regulates many physiological and biochemical processes and stress responses in plants (reviewed by Domingos et al. [56]). In animals, NO preferentially binds to hemoproteins in both ferrous and ferric states [57,58], which indicates a strong association of heme synthesis by FC and the arginine–citrulline cycle. We observed that both low and high citrulline content accessions showed reduced *FC1* expression. However, the reason as to why the *FC1* transcript levels were low in accessions with both high and low citrulline content is elusive. We hypothesize that *FC1* expression may be regulated in a spatio-temporal and growth-dependent manner. Further analysis of *FC1* expression in watermelon flesh at various fruit developmental stages may reveal the association of citrulline content and *FC1* transcript abundance.

The ALS complex consists of three pairs of subunits, a large subunit responsible for catalysis, and a small subunit for feedback inhibition. All three genes encoding the putative ALS large subunit (ClCG06G017910), ALS small subunit_1 (ASsu1; ClCG09G014670) and ALS small subunit_2 (ASsu2; ClCG03G010140) proteins were found in the watermelon genome. There are no known direct metabolic links between citrulline and BCAA synthesis (BCAAs) in the canonical or non-canonical systems. Because of the sub-optimal levels of free citrulline in most plants, limited information is available on its relationship with other amino acids. Nonetheless, some studies suggest the possibility of unknown links between citrulline/arginine and BCAA pathways; for example, (1) arginine, a catabolic product of citrulline, is significantly and positively correlated with BCAAs in soybean seeds [59]; (2) levels of BCAAs were increased along with those of citrulline, arginine, and ornithine during conditional downregulation of *target of rapamycin* gene in *Arabidopsis* [60]; and (3) citrulline level was increased in leaves of the rice ALS mutant [61].

4. Materials and Methods

4.1. Plant Materials, Citrulline Extraction, and Quantification

Selfed seeds of 187 National Plant Germplasm System collections of *C. lanatus* var. *vulgaris* and *C. lanatus* var. *mucasospermus* were grown in the greenhouse. The authors are grateful to R. Jarret, Plant Genetic Resources Conservation Unit, USDA-ARS (Griffin, GA, USA) for providing the seeds of germplasm accessions. Citrulline content in the flesh determined by using three biological replicates for each accession (Table S1). For determining citrulline content, ~20 mg lyophilized flesh tissue for each biological replicate was homogenized by using 3 mm Demag stainless steel balls (Abbott Ball Co., CT, USA) and a Harbil model 5G-HD paint shaker, suspended in 20 mM cold HCl and centrifuged at 14,609× *g* for 20 min at 4 °C. The supernatant was filtered by using 0.45 uM 96-well filters (Pall Life Sciences, NY, USA). The filtrate was derivatized with an AccQ-Fluor reagent kit (Waters Corp., Milford, MA, USA) per the manufacturer's protocol. UPLC-ESI-MS/MS analysis involved using a Waters Acquity H-class UPLC system coupled to a Waters Xevo TQ mass spectrometer with an electrospray ionization (ESI) probe, Waters ACQUITY UPLC Fluorescence (FLR) detector and Water's AccQ•Tag Ultra column. The mobile phase consisted of water (0.1% formic acid v/v) (A) and acetonitrile (0.1% formic acid v/v) (B). The column heater was set to 60 °C, and the mobile phase flow rate was maintained at a constant rate of 0.6 mL/min. By using the Waters IntelliStart software, multiple reaction monitoring transitions for citrulline were optimized. The ESI source was operated at 150 °C, with desolvation temperature 450 °C, desolvation gas flow rate 900 L/h and capillary voltage 3.2 kV. Multiple reaction monitoring was performed in the positive mode. Instrument monitoring and data acquisition, integration, and quantification involved using Water's MassLynx software.

4.2. Phylogenetic and Population Genomic Analyses

Phenotypic data were analyzed by using JMP software (JMP Pro 14 (SAS Institute, Cary, NC, SAS http://www.jmp.com/en_us/home.html)). The variance components were estimated by using REML-BLUP analysis. The broad-sense heritability was estimated as $H^2 = \delta^2{}_g/(\ \delta^2{}_g + \delta^2{}_{e/r})$, where $\delta^2{}_g$ and $\delta^2{}_e$ are estimated genotypic and error variances, respectively [62]. The genetic gain was calculated as $Gs = K \times H^2 \times (\delta^2{}_p)^{-1/2}$, where K is the selection intensity at 5% ($k = 2.056$), H^2 is heritability in a broad sense and $(\delta^2{}_p)^{-1/2}$ is the phenotypic standard deviation. Totals of 11,485 SNPs [27] and 16,292 SNPs [26] were combined and further filtered using MAF ≥ 0.01 to identify 12,125 SNPs that have a call rate of 70% in our GWAS panel. For population stratification analysis, we used more stringent cut off for SNP selection based on MAF of 0.1 and 90% call rate. Further informative SNPs were selected by removing SNPs located in a LD block by 50% cut off and also SNPs that deviated from Hardy–Weinberg equilibrium were discarded. A neighbor-joining (NJ) tree for watermelon accessions was constructed by using TASSEL 5.0 [63]. Archaeopteryx [64] was used to visualize and analyze the tree. For analyzing population structure, we used the principal components, or eigenvectors, of principal component analysis (PCA), and corresponding eigenvalues were estimated by using the EIGENSTRAT algorithm [65] with the SNP & Variation Suite (SVS v8.8.1; Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com).

4.3. Association Analysis

For GWAS, the population structure Q matrix was replaced by the PC matrix. The PC matrix and identity by descent (IBD) was calculated from LD-pruned SNPs in SVS v8.1.5. GWAS involved a multiple-locus mixed linear model developed by the EMMAX method and implemented in SVS v8.1.5. We used a PC matrix (first two vectors) and the IBD matrix to correct for population stratification. Manhattan plots for associated SNPs were visualized in GenomeBrowse v1.0 (Golden Helix, Inc) (Figure S2). The SNP *p*-values from GWAS underwent false discovery rate (FDR) analysis.

4.4. Haplotype Network Analysis

To evaluate the haplotype frequency, we first specified the haplogroups and the frequency with which a determined haplotype appeared in each haplogroup. Then, we determined the genetic constitution as the proportion of SNPs that belong to a haplogroup. The haplotype frequency analysis involved estimating haplotype diversity. The number of haplotypes and the haplotype diversity values were calculated by using DnaSP 5.10 (http://www.ub.edu/dnasp/DnaSP_OS.html) [66]. Haplotype networking and LD calculation were performed by using publicly available SNP data for 1209 *C. lanatus* var. *vulgaris* and 51 *C. mucasospermus* (egusi) [21]. Adjacent SNP pairs within a chromosome underwent LD analysis by using expectation maximization [67]. All LD plots as well as LD measurements and haplotype frequency calculations involved use of SVS v8.8.1. This collection consisted 220 PIs from Africa, 352 from Asia, 497 from Europe, 121 from North America, and 16 from South America [26].

4.5. Total RNA Extraction and qRT-PCR

Total RNA was extracted from watermelon flesh (three biological replicates) by using TRIzol^{TM} Reagent (Life Technologies, Carlsbad, CA, USA). On-column DNase I treatment was used to remove genomic DNA. The RNA concentrations and quality were determined by using NanoDrop 2000 (Thermofisher, Waltham, MA, USA). The first strand cDNAs were synthesized with 1 µg total RNA in a 20 µL reaction mixture with SuperScript^{TM} II Reverse Transcriptase (Thermo Fisher Scientific, USA) per the manufacturer's instructions. cDNAs were then diluted by adding 180 µL sterile nuclease-free water. qRT-PCR was performed with PowerUp^{TM} SYBR Green Master Mix (Applied Biosystems, Thermo Fisher Scientific, USA) and the StepOnePlus^{TM} Real-Time PCR System (Applied Biosystems). PCR was programmed at 95 °C for 30 s, followed by 95 °C for 10 s and 60 °C for 60 s for 40 cycles. Gene expression was normalized to *Actin 8* expression (housekeeping reference), and transcript abundance of target genes was calculated with the PI 526238 accession, with significantly low citrulline content, as a calibrator. The $2^{-\Delta\Delta CT}$ method was used to estimate relative transcript level [68]. The primers for qRT-PCR analysis are in Table S3.

5. Conclusions

This study demonstrated a wide variation in watermelon germplasm for citrulline content. Identifying such variability and introgressing it into high-yield lines will be key to long-term watermelon breeding for improved nutraceutical content. We found moderate to high (83%) broad-sense heritability, which indicates the possibility of successful citrulline enhancement with introgression breeding. Here we provide new insights into the domestication and population genetic structure based on citrulline content in geographically spread-out watermelon accessions. The SNP loci associated with citrulline content identified in this study would benchmark the efforts of development of molecular markers to enhance nutritional quality in elite watermelon cultivars.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/20/21/5392/s1.

Author Contributions: U.K.R., P.N., A.L. and V.J. designed the research program. S.B.A. collected the samples. U.K.R., V.L.A., P.N. analyzed the data and conducted the GWAS. V.J. performed metabolic analysis. S.S. performed gene expression analysis. C.L.-O. performed haplotyping and network analysis. G.P. performed field evaluation. U.K.R., S.S., P.N., A.L., and V.J. wrote the manuscript. All authors read and approved the manuscript.

Funding: This work is supported by the USDA-NIFA 2017-38821-26434, USDA-NIFA 2018-38821-27734, USDA-NIFA [Hatch Project No TEX09647/project accession no. 1011513] and National Watermelon Promotion Board.

Acknowledgments: We thank Diwas Silwal for assistance in sample processing for metabolic analysis.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

ALS	Acetolactate synthase
ASsu1	Acetolactate synthase small subunit_1
ASsu1	Acetolactate synthase small subunit_2
BCAAs	branched-chain amino acid synthesis
FDR	false discovery rate
FC	Ferrochelatase
GWAS	genome-wide association studies
IBD	identity by descent
LD	linkage disequilibrium
NJ	neighbor-Joining
PI	Plant Introduction
SNP	single nucleotide polymorphism

References

1. Alsop, P.; Hauton, D. Oral nitrate and citrulline decrease blood pressure and increase vascular conductance in young adults: a potential therapy for heart failure. *Eur. J. Appl. Physiol.* **2016**, *116*, 1651–1661. [CrossRef] [PubMed]
2. Moinard, C.; Maccario, J.; Walrand, S.; Lasserre, V.; Marc, J.; Boirie, Y.; Cynober, L. Arginine behaviour after arginine or citrulline administration in older subjects. *Br. J. Nutr.* **2016**, *115*, 399–404. [CrossRef] [PubMed]
3. Morita, M.; Hayashi, T.; Ochiai, M.; Maeda, M.; Yamaguchi, T.; Ina, K.; Kuzuya, M. Oral supplementation with a combination of L-citrulline and L-arginine rapidly increases plasma L-arginine concentration and enhances NO bioavailability. *Biochem. Biophys. Res. Commun.* **2014**, *454*, 53–57. [CrossRef] [PubMed]

4. Bahri, S.; Zerrouk, N.; Aussel, C.; Moinard, C.; Crenn, P.; Curis, E.; Chaumeil, J.-C.; Cynober, L.; Sfar, S. Citrulline: from metabolism to therapeutic use. *Nutrition* **2013**, *29*, 479–484. [CrossRef]
5. Schellekens, G.A.; de Jong, B.A.W.; van den Hoogen, F.H.J.; van de Putte, L.B.A.; van Venrooij, W.J. Pillars Article: Citrulline is an Essential Constituent of Antigenic Determinants Recognized by Rheumatoid Arthritis–specific Autoantibodies. *J. Clin. Invest.* **1998**, *101*, 273–281. [CrossRef]
6. Allerton, T.D.; Proctor, D.N.; Stephens, J.M.; Dugas, T.R.; Spielmann, G.; Irving, B.A. l-Citrulline Supplementation: Impact on Cardiometabolic Health. *Nutrients* **2018**, *10*, 921. [CrossRef]
7. Papadia, C.; Osowska, S.; Cynober, L.; Forbes, A. Citrulline in health and disease. Review on human studies. *Clin. Nutr.* **2018**, *37*, 1823–1828. [CrossRef]
8. Breuillard, C.; Bonhomme, S.; Couderc, R.; Cynober, L.; De Bandt, J.P. In vitro anti-inflammatory effects of citrulline on peritoneal macrophages in Zucker diabetic fatty rats. *Br. J. Nutr.* **2015**, *113*, 120–124. [CrossRef]
9. Collins, J.K.; Wu, G.; Perkins-Veazie, P.; Spears, K.; Claypool, P.L.; Baker, R.A.; Clevidence, B.A. Watermelon consumption increases plasma arginine concentrations in adults. *Nutrition* **2007**, *23*, 261–266. [CrossRef]
10. Wu, G.; Collins, J.K.; Perkins-Veazie, P.; Siddiq, M.; Dolan, K.D.; Kelly, K.A.; Heaps, C.L.; Meininger, C.J. Dietary supplementation with watermelon pomace juice enhances arginine availability and ameliorates the metabolic syndrome in Zucker diabetic fatty rats. *J. Nutr.* **2007**, *137*, 2680–2685. [CrossRef]
11. Akashi, K.; Miyake, C.; Yokota, A. Citrulline, a novel compatible solute in drought-tolerant wild watermelon leaves, is an efficient hydroxyl radical scavenger. *FEBS Lett.* **2001**, *508*, 438–442. [CrossRef]
12. Joshi, V.; Fernie, A.R. Citrulline metabolism in plants. *Amino Acids* **2017**, *49*, 1543–1559. [CrossRef] [PubMed]
13. Fish, W.W.; Bruton, B.D. Quantification of L-citrulline and other physiologic amino acids in watermelon and selected cucurbits. In Proceedings of the Cucurbitaceae 2010, Charleston, SC, USA, 14–18 November 2010; pp. 152–154.
14. Rimando, A.M.; Perkins-Veazie, P.M. Determination of citrulline in watermelon rind. *J. Chromatogr. A* **2005**, *1078*, 196–200. [CrossRef] [PubMed]
15. Davis, A.R.; Fish, W.; Levi, A.; King, S.; Wehner, T.; Perkins-Veazie, P. L-citrulline levels in watermelon cultivars from three locations. *Cucurbit. Genet. Coop. Rpt.* **2010**, *33*, 36–39. [CrossRef]
16. Davis, A.R.; Webber, C.L.; Fish, W.W.; Wehner, T.C.; King, S.; Perkins-Veazie, P. L-citrulline levels in watermelon cultigens tested in two environments. *HortScience* **2011**, *46*, 1572–1575. [CrossRef]
17. Akashi, K.; Mifune, Y.; Morita, K.; Ishitsuka, S.; Tsujimoto, H.; Ishihara, T. Spatial accumulation pattern of citrulline and other nutrients in immature and mature watermelon fruits. *J. Sci. Food Agric.* **2017**, *97*, 479–487. [CrossRef]
18. Joshi, V.; Joshi, M.; Silwal, D.; Noonan, K.; Rodriguez, S.; Penalosa, A. Systematized biosynthesis and catabolism regulate citrulline accumulation in watermelon. *Phytochemistry* **2019**, *162*, 129–140. [CrossRef]
19. Curis, E.; Nicolis, I.; Moinard, C.; Osowska, S.; Zerrouk, N.; Bénazeth, S.; Cynober, L. Almost all about citrulline in mammals. *Amino Acids* **2005**, *29*, 177. [CrossRef]
20. Yokota, A.; Kawasaki, S.; Iwano, M.; Nakamura, C.; Miyake, C.; Akashi, K. Citrulline and DRIP-1 protein (ArgE homologue) in drought tolerance of wild watermelon. *Ann. Bot.* **2002**, *89*, 825–832. [CrossRef]
21. Wehner, T.C.; Naegele, R.P.; Perkins-Veazie, P. Heritability and Genetic Variance Components Associated with Citrulline, Arginine, and Lycopene Content in Diverse Watermelon Cultigens. *HortScience* **2017**, *52*, 936–940. [CrossRef]
22. Fall, L.A.; Perkins-Veazie, P.; Ma, G.; McGregor, C. QTLs associated with flesh quality traits in an elite × elite watermelon population. *Euphytica* **2019**, *215*, 30. [CrossRef]
23. Angelovici, R.; Lipka, A.E.; Deason, N.; Gonzalez-Jorge, S.; Lin, H.; Cepela, J.; Buell, R.; Gore, M.A.; Dellapenna, D. Genome-wide analysis of branched-chain amino acid levels in Arabidopsis seeds. *Plant Cell* **2013**, *25*, 4827–4843. [CrossRef] [PubMed]
24. Angelovici, R.; Batushansky, A.; Deason, N.; Gonzalez-Jorge, S.; Gore, M.A.; Fait, A.; DellaPenna, D. Network-Guided GWAS Improves Identification of Genes Affecting Free Amino Acids. *Plant Physiol.* **2017**, *173*, 872–886. [CrossRef] [PubMed]
25. Peng, Y.; Liu, H.; Chen, J.; Shi, T.; Zhang, C.; Sun, D.; He, Z.; Hao, Y.; Chen, W. Genome-Wide Association Studies of Free Amino Acid Levels by Six Multi-Locus Models in Bread Wheat. *Front. Plant Sci.* **2018**, *9*, 1196. [CrossRef]

26. Wu, S.; Wang, X.; Reddy, U.; Sun, H.; Bao, K.; Gao, L.; Mao, L.; Patel, T.; Ortiz, C.; Abburi, V.L.; et al. Genome of 'Charleston Gray', the principal American watermelon cultivar, and genetic characterization of 1,365 accessions in the U.S. National Plant Germplasm System watermelon collection. *Plant Biotechnol. J.* **2019**, *0*, 1–13. [CrossRef]
27. Nimmakayala, P.; Levi, A.; Abburi, L.; Abburi, V.L.; Tomason, Y.R.; Saminathan, T.; Vajja, V.G.; Malkaram, S.; Reddy, R.; Wehner, T.C.; et al. Single nucleotide polymorphisms generated by genotyping by sequencing to characterize genome-wide diversity, linkage disequilibrium, and selective sweeps in cultivated watermelon. *BMC Genom.* **2014**, *15*, 767. [CrossRef]
28. Reddy, U.K.; Nimmakayala, P.; Levi, A.; Abburi, V.L.; Saminathan, T.; Tomason, Y.R.; Vajja, G.; Reddy, R.; Abburi, L.; Wehner, T.C.; et al. High-resolution genetic map for understanding the effect of genome-wide recombination rate on nucleotide diversity in watermelon. *G3* **2014**, *4*, 2219–2230. [CrossRef]
29. Reddy, U.K.; Abburi, L.; Abburi, V.L.; Saminathan, T.; Cantrell, R.; Vajja, V.G.; Reddy, R.; Tomason, Y.R.; Levi, A.; Wehner, T.C. A genome-wide scan of selective sweeps and association mapping of fruit traits using microsatellite markers in watermelon. *J. Hered.* **2014**, *106*, 166–176. [CrossRef]
30. Patterson, H.D.; Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **1971**, *58*, 545–554. [CrossRef]
31. Henderson, C.R. Sire evaluation and genetic trends. *J. Anim. Sci.* **1973**, *1973*, 10–41. [CrossRef]
32. Chaleff, R.S.; Mauvais, C. Acetolactate synthase is the site of action of two sulfonylurea herbicides in higher plants. *Science* **1984**, *224*, 1443–1445. [CrossRef] [PubMed]
33. Shaner, D.L.; Anderson, P.C.; Stidham, M.A. Imidazolinones: potent inhibitors of acetohydroxyacid synthase. *Plant Physiol.* **1984**, *76*, 545–546. [CrossRef] [PubMed]
34. McCourt, J.A.; Pang, S.S.; King-Scott, J.; Guddat, L.W.; Duggleby, R.G. Herbicide-binding sites revealed in the structure of plant acetohydroxyacid synthase. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 569–573. [CrossRef] [PubMed]
35. Paris, H.S. Origin and emergence of the sweet dessert watermelon, *Citrullus lanatus*. *Ann. Bot.* **2015**, *116*, 133–148. [CrossRef] [PubMed]
36. Paris, H.S.; Daunay, M.C.; Janick, J. Medieval iconography of watermelons in Mediterranean Europe. *Ann. Bot.* **2013**, *112*, 867–879. [CrossRef]
37. Vergauwen, D.; De Smet, I. Watermelons versus Melons: A Matter of Taste. *Trends Plant Sci.* **2019**, *24*, 973–976. [CrossRef]
38. Sandlin, K.; Prothro, J.; Heesacker, A.; Khalilian, N.; Okashah, R.; Xiang, W.; Bachlava, E.; Caldwell, D.G.; Taylor, C.A.; Seymour, D.K.; et al. Comparative mapping in watermelon [*Citrullus lanatus* (Thunb.) Matsum. et Nakai]. *Theor. Appl. Genet.* **2012**, *125*, 1603–1618. [CrossRef]
39. Hashizume, T.; Shimamoto, I.; Hirai, M. Construction of a linkage map and QTL analysis of horticultural traits for watermelon [*Citrullus lanatus* (THUNB.) MATSUM & NAKAI] using RAPD, RFLP and ISSR markers. *Theor. Appl. Genet.* **2003**, *106*, 779–785. [CrossRef]
40. Gusmini, G.; Wehner, T.C. Foundations of yield improvement in watermelon. *Crop Sci.* **2005**, *45*, 141–146. [CrossRef]
41. Davis, A.R.; Levi, A.; Tetteh, A.; Wehner, T.; Russo, V.; Pitrat, M. Evaluation of watermelon and related species for resistance to race 1W powdery mildew. *J. Am. Soc. Hortic. Sci.* **2007**, *132*, 790–795. [CrossRef]
42. Guo, S.; Zhang, J.; Sun, H.; Salse, J.; Lucas, W.J.; Zhang, H.; Zheng, Y.; Mao, L.; Ren, Y.; Wang, Z.; et al. The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat. Genet.* **2013**, *45*, 51–58. [CrossRef] [PubMed]
43. Ren, Y.; Zhao, H.; Kou, Q.; Jiang, J.; Guo, S.; Zhang, H.; Hou, W.; Zou, X.; Sun, H.; Gong, G.; et al. A high resolution genetic map anchoring scaffolds of the sequenced watermelon genome. *PLoS ONE* **2012**, *7*, e29453. [CrossRef] [PubMed]
44. Gross, B.L.; Olsen, K.M. Genetic perspectives on crop domestication. *Trends Plant Sci.* **2010**, *15*, 529–537. [CrossRef]
45. Hancock, J.F. *Plant Evolution and the Origin of Crop Species*; CABI: Wallingford, OX, UK, 2012.
46. Meyer, R.S.; DuVal, A.E.; Jensen, H.R. Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *New Phytol.* **2012**, *196*, 29–48. [CrossRef]
47. Vavilov, N.I. *Proiskhozhdenie i geografiya kul'turnykh rastenii (Origin and Geography of Cultivated Plants)*; Cambridge University Press: Cambridge, UK, 1987; ISBN 13:978-0521111591.

48. Chomicki, G.; Renner, S.S. Watermelon origin solved with molecular phylogenetics including Linnaean material: another example of museomics. *New Phytol.* **2015**, *205*, 526–532. [CrossRef]
49. Ferreira, G.C.; Franco, R.; Lloyd, S.G.; Moura, I.; Moura, J.J.; Huynh, B.H. Structure and function of ferrochelatase. *J. Bioenerg. Biomembr.* **1995**, *27*, 221–229. [CrossRef] [PubMed]
50. Noble, M.A.; Munro, A.W.; Rivers, S.L.; Robledo, L.; Daff, S.N.; Yellowlees, L.J.; Shimizu, T.; Sagami, I.; Guillemette, J.G.; Chapman, S.K. Potentiometric analysis of the flavin cofactors of neuronal nitric oxide synthase. *Biochemistry* **1999**, *38*, 16413–16418. [CrossRef]
51. Stuehr, D.; Pou, S.; Rosen, G.M. Oxygen reduction by nitric-oxide synthases. *J. Biol. Chem.* **2001**, *276*, 14533–14536. [CrossRef]
52. Förstermann, U.; Sessa, W.C. Nitric oxide synthases: regulation and function. *Eur. Heart J.* **2011**, *33*, 829–837. [CrossRef]
53. Gas, E.; Flores-Perez, U.; Sauret-Gueto, S.; Rodriguez-Concepcion, M. Hunting for plant nitric oxide synthase provides new evidence of a central role for plastids in nitric oxide metabolism. *Plant Cell* **2009**, *21*, 18–23. [CrossRef]
54. Fröhlich, A.; Durner, J. The hunt for plant nitric oxide synthase (NOS): is one really needed? *Plant Sci.* **2011**, *181*, 401–404. [CrossRef] [PubMed]
55. Hamza, I.; Dailey, H.A. One ring to rule them all: trafficking of heme and heme synthesis intermediates in the metazoans. *Biochim. Biophys. Acta* **2012**, *1823*, 1617–1632. [CrossRef] [PubMed]
56. Domingos, P.; Prado, A.M.; Wong, A.; Gehring, C.; Feijo, J.A. Nitric oxide: A multitasked signaling gas in plants. *Mol. Plant* **2015**, *8*, 506–520. [CrossRef] [PubMed]
57. Henry, Y.; Guissani, A. Interactions of nitric oxide with hemoproteins: roles of nitric oxide in mitochondria. *Cell. Mol. Life Sci.* **1999**, *55*, 1003–1014. [CrossRef]
58. Kim-Shapiro, D.B.; Gladwin, M.T. Heme Protein Metabolism of NO and Nitrite. In *Nitric Oxide*; Ignarro, L.J., Freeman, B.A., Eds.; Academic Press, London Wall: London, UK, 2017; pp. 85–96. [CrossRef]
59. Assefa, Y.; Purcell, L.C.; Salmeron, M.; Naeve, S.; Casteel, S.N.; Kovacs, P.; Archontoulis, S.; Licht, M.; Below, F.; Kandel, H.; et al. Assessing Variation in US Soybean Seed Composition (Protein and Oil). *Front. Plant Sci.* **2019**, *10*, 298. [CrossRef]
60. Caldana, C.; Li, Y.; Leisse, A.; Zhang, Y.; Bartholomaeus, L.; Fernie, A.R.; Willmitzer, L.; Giavalisco, P. Systemic analysis of inducible target of rapamycin mutants reveal a general metabolic switch controlling growth in A rabidopsis thaliana. *Plant J.* **2013**, *73*, 897–909. [CrossRef]
61. Endo, M.; Shimizu, T.; Fujimori, T.; Yanagisawa, S.; Toki, S. Herbicide-resistant mutations in acetolactate synthase can reduce feedback inhibition and lead to accumulation of branched-chain amino acids. *Food Nutr. Sci.* **2013**, *4*, 522–528. [CrossRef]
62. Nyquist, W.E.; Baker, R. Estimation of heritability and prediction of selection response in plant populations. *Crit. Rev. Plant Sci.* **1991**, *10*, 235–322. [CrossRef]
63. Bradbury, P.J.; Zhang, Z.; Kroon, D.E.; Casstevens, T.M.; Ramdoss, Y.; Buckler, E.S. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **2007**, *23*, 2633–2635. [CrossRef]
64. Han, M.V.; Zmasek, C.M. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinf.* **2009**, *10*, 356. [CrossRef]
65. Patterson, N.; Price, A.L.; Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2006**, *2*, e190. [CrossRef] [PubMed]
66. Rozas, J.; Ferrer-Mata, A.; Sanchez-DelBarrio, J.C.; Guirao-Rico, S.; Librado, P.; Ramos-Onsins, S.E.; Sanchez-Gracia, A. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol. Biol. Evol.* **2017**, *34*, 3299–3302. [CrossRef] [PubMed]
67. Excoffier, L.; Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid LDpopulation. *Mol. Biol. Evol.* **1995**, *12*, 921–927. [CrossRef] [PubMed]
68. Livak, K.J.; Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* **2001**, *25*, 402–408. [CrossRef] [PubMed]

International Journal of
Molecular Sciences

MDPI

Review

Insights on Calcium-Dependent Protein Kinases (CPKs) Signaling for Abiotic Stress Tolerance in Plants

Rana Muhammad Atif [1,2,*,†], Luqman Shahid [1,†], Muhammad Waqas [1], Babar Ali [1], Muhammad Abdul Rehman Rashid [1,3], Farrukh Azeem [4], Muhammad Amjad Nawaz [5], Shabir Hussain Wani [6] and Gyuhwa Chung [7,*]

[1] Department of Plant Breeding and Genetics, University of Agriculture, Faisalabad 38000, Pakistan; luqmanshahid73@gmail.com (L.S.); bhuttawaqas@ymail.com (M.W.); babar1292ali@gmail.com (B.A.); rashidpbg@hotmail.com (M.A.R.R.)

[2] Center for Advanced Studies in Agriculture and Food Security, University of Agriculture, Faisalabad 38040, Pakistan

[3] Industrial Crops Research Institute, Yunnan Academy of Agricultural Sciences, Kunming 650200, China

[4] Department of Bioinformatics and Biotechnology, Government College University, Faisalabad 38040, Pakistan; azeuaf@hotmail.com

[5] Education Scientific Center of Nanotechnology, Far Eastern Federal University, 690950 Vladivostok, Russia; amjad_ucauos@yahoo.com

[6] Mountain Research Centre for Field Crops, Sher-e-Kashmir University of Agricultural Sciences and Technology of Kashmir, Srinagar 190001, India; shabirhussainwani@gmail.com

[7] Department of Biotechnology, Chonnam National University, Chonnam 59626, Korea

* Correspondence: dratif@uaf.edu.pk (R.M.A.); chung@chonnam.ac.kr (G.C.)

† These authors contributed equally to this work.

Received: 28 June 2019; Accepted: 17 October 2019; Published: 24 October 2019

Abstract: Abiotic stresses are the major limiting factors influencing the growth and productivity of plants species. To combat these stresses, plants can modify numerous physiological, biochemical, and molecular processes through cellular and subcellular signaling pathways. Calcium-dependent protein kinases (CDPKs or CPKs) are the unique and key calcium-binding proteins, which act as a sensor for the increase and decrease in the calcium (Ca) concentrations. These Ca flux signals are decrypted and interpreted into the phosphorylation events, which are crucial for signal transduction processes. Several functional and expression studies of different CPKs and their encoding genes validated their versatile role for abiotic stress tolerance in plants. CPKs are indispensable for modulating abiotic stress tolerance through activation and regulation of several genes, transcription factors, enzymes, and ion channels. CPKs have been involved in supporting plant adaptation under drought, salinity, and heat and cold stress environments. Diverse functions of plant CPKs have been reported against various abiotic stresses in numerous research studies. In this review, we have described the evaluated functions of plant CPKs against various abiotic stresses and their role in stress response signaling pathways.

Keywords: calcium-dependent protein kinases; calcium signaling; ABA; drought; salinity

1. Introduction

Plants have several adaptive features to cope with biotic and abiotic stresses under challenging environmental situations. Plants respond to these stresses by inducing the expression of stress-responsive genes through a complex signaling pathway. The expression of these stress-responsive genes is induced upon changes in calcium ion (Ca^{2+}) concentrations, due to various biotic and abiotic stimuli [1,2], which enable plant adaptations in a wide range of stressed environments.

Calcium (Ca) as a ubiquitous secondary messenger regulates the stress signaling mechanism in plants. Changes in Ca^{2+} concentration are sensed by several calcium-binding proteins, especially calcium-dependent protein kinases [3]. The calcium-dependent abiotic and biotic stress signaling mechanisms are most commonly dominated by calcium-dependent protein kinases, which play a pivotal role in the regulation of plant responsiveness to salt, drought, and cold and heat stresses as well as other environmental factors. Ca^{2+} is involved in abscisic acid (ABA)-dependent biotic and abiotic stress signals in various plant species [4,5]. The calcium-dependent protein kinases phosphorylate the ABA-responsive element-binding factors (ABFs). ABA regulation by Ca^{2+} is associated with plant defense systems through induction of antioxidants [6], including reactive oxygen species (ROS) [2], and other enzymes like superoxide dismutase (SOD), catalase 3 (CAT3), ascorbate peroxidase (APX), glutathione peroxidase (GPX), and glutathione reductase (GR) [6,7]. It is also involved in the induction of some nonenzymatic antioxidants like ascorbic acid, α-tocopherol, carotenoids, and glutathione and controls multiple abiotic stress response processes [6,8–10]. This review will provide insight into the role of calcium-dependent protein kinases (CPKs) in abiotic stress tolerance in different plant species.

2. CPK Enzymes and Related Kinases

Several calcium-binding protein families have been identified in plants, which are potentially involved in the regulation of calcium-dependent abiotic stress response mechanisms. These Ca^{2+} sensors decode and transmit complex information, present in the form of calcium signal, to the phosphorylation events and regulate stress-responsive genes through protein interactions [11]. These Ca^{2+} signal-decoding groups include calcium-dependent protein kinases (CDPKs or CPKs), calmodulins (CaMs), calmodulin-like protein kinases (CMLs), calcineurin β-like proteins (CBLs), and Ca^{2+}/calmodulin-dependent protein kinase (CCaMK) [12,13]. Among all these kinases, CPKs, CMLs, and CBLs have only been discovered in plants and some protozoans, while CaMs are highly conserved among all eukaryotes [11,14]. CaMs, CBLs, and CMLs are small proteins that function as calcium signal communicators through binding to downstream effectors (EFs) [15,16]. CaMs evolved from CMLs, which are considered as the most primitive calcium-binding proteins [13]. Among all these, CPKs were identified in plants as well as green algae, oomycetes, and in some protozoans [17], but they are not present in animals. CPKs, through direct binding with Ca^{2+}, have a predominant regulatory role for the Ca-sensing protein families [17].

3. CPK Family in Plants

CPKs are considered as the versatile player for the regulation of abiotic stress management in plants [17]. In 1984, the very first plant CPKs were identified in *Pisum sativum* [18]. These proteins were initially purified from soybeans in 1987. A CPK encoding gene was cloned from *Arabidopsis thaliana* in 1991, which opened new ways for CPK gene cloning in several other plant species [11,19,20]. The presence of CPKs in almost all parts of the plant demonstrates that these kinases have a high potential for regulating various signal transduction pathways and have a significant influence on plant growth and development [17,21–23].

3.1. CPK Distribution and Localization in Plants

CPKs show a widespread distribution in different plant species. The whole-genome sequencing of plant species (e.g., *Arabidopsis* [24]) enables researchers to conduct genome-wide identifications of variable CPK encoding genes. These studies identified 34 CPK-encoding genes in the genome of *Arabidopsis thaliana*, 20 in *Triticum aestivum* (wheat), and 31 in *Oryza sativa* (rice) [20,25,26]. *Solanum lycopersicum* (tomato), which is a model plant of the *Solanaceae* family, has 29 CPK-encoding genes [27]. Genome-wide exploration of some other plants such as *Zea mays* (maize), *Hordeum vulgare* (barley), *Cucumis melo* (melon), *Populus trichocarpa* (poplar), *Gossypium raimondii* (cotton), *Manihot esculenta* (cassava), and *Vitis vinifera* (grapevine) revealed the presence of 40, 28, 18, 30, 41, 27, and 19 CPK-encoding genes, respectively [28–34] (Table 1). Mostly, CPK-encoding genes are expressed in leaves, meristems, roots, and flowers, while some are expressed only in specific tissues [23,35,36].

Table 1. Genome-wide identification of calcium-dependent protein kinases (CPKs) among various plant species.

Sr. #	Common Name	Botanical Name	No. of CPKs	Genome Size (Mb)	Reference
1	Algae	*Volvox carteri*	6	131.2	[37]
2	Apple	*Malus domestica*	28	881.3	[37]
3	Arabidopsis	*Arabidopsis thaliana*	34	135	[20]
4	Banana	*Musa acuminata*	44	523	[38]
5	Barley	*Hordeum vulgare*	27	667	[39]
6	Barley	*Hordeum vulgare*	28	667	[31]
7	Barrel clover	*Medicago truncatula*	11	360	[37]
8	Black cottonwood	*Populus trichocarpa*	28	422.9	[37]
9	Poplar	*Populus trichocarpa*	30	500	[34]
10	Butcher	*Micromonas pusilla*	22	2	[13,37]
11	Cacao tree	*Theobroma cacao*	17	346	[13,37]
12	Canola	*Brassica napus*	25	1130	[40]
13	Cassava	*Manihot esculenta*	26	532.5	[30]
14	Caster bean	*Ricinus communis*	15	400	[37]
15	Castor bean	*Ricinus communis*	15	400	[13,37]
16	Chinese liquorice	*Glycyrrhiza uralensis*	23	379	[41]
17	Chlamydomonas	*Chlamydomonas reinhardtii*	14	111.1	[13,37]
18	Clementine	*Citrus clememtina*	26	301.4	[37]
19	Cocoa tree	*Theobroma cacao*	17	346	[37]
20	Columbine	*Aquilegia coerulea*	16	306.5	[13,37]
21	Cotton	*Gossypium raimondii*	41	880	[28]
22	Cotton	*Gossypium hirsutum*	98	2250–2430	[42]
23	Cucumber	*Cucumis sativus*	19	323.99	[43]
24	Cucumber	*Cucumis sativus*	18	203	[37]
25	Finger Millet	*Eleusine coracana*	4	1593	[44]
26	Flax	*Linum usitatissimum*	47	318.3	[37]
27	Flooded gum	*Eucalyptus grandis*	22	691	[37]
28	Foxtail Millet	*Setaria italic*	27	405.7	[37]
29	Foxtail Millet	*Setaria italic*	29	405.7	[45]
30	Foxtail millet	*Setaria italica*	27	405.7	[13,37]
31	Grape	*Vitis vinifera*	19	500	[29]
32	Grapevine	*Vitis amurensis*	17	500	[46]
33	Grapevine	*Vitis amurensis*	13	500	[47]
34	Green algae	*Coccomyxa subellipsoidea*	2	49	[13,37]
35	Green algae	*Ostreococcus lucimarinus*	3	13.2	[13,37]
36	Green bean	*Phaseolus vulgaris*	25	521.1	[37]
37	Linseed	*Linum usitatissimum*	47	318.3	[13,37]
38	Maize	*Zea mays*	35	2500	[48]
39	Maize	*Zea mays*	40	2500	[49]
40	Maize	*Zea mays*	47	2500	[37]

Table 1. *Cont.*

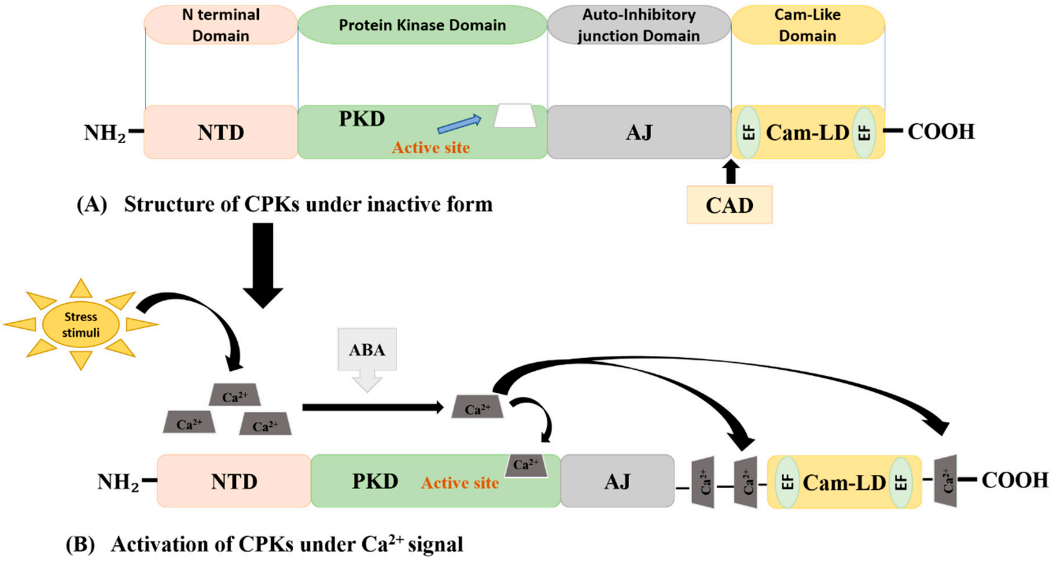
Sr. #	Common Name	Botanical Name	No. of CPKs	Genome Size (Mb)	Reference
41	Melon	*Cucumis melo*	18	375	[32]
42	Monkey flower	*Mimulus guttatus*	25	321.7	[37]
43	Mustard	*Brassica rapa*	49	283.8	[37]
44	Norway spruce	*Picea abies*	11	1960	[37]
45	Oilseed rape	*Brassica rapa*	49	283.8	[13,37]
46	Orange	*Citrus sinensis*	24	319	[13,37]
47	Papaya	*Carica papaya*	15	135	[13,37]
48	Papaya	*Carica papaya*	15	135	[37]
49	Peach	*Prunus persica*	17	227.3	[37]
50	Pepper	*Capsicum annuum*	31	407.5	[50]
51	Pigeon Pea	*Cajanus cajan*	23	852	[51]
52	Potato	*Solanum tubersum*	21	800	[37]
53	Potato	*Solanum tuberosum*	23	800	[52]
54	Purple false brome	*Brachypodium distachyon*	27	272	[37]
55	Purple false brome	*Brachipodium distachyon*	27	272	[37]
56	Red Shepherd's Purse	*Capsella rubella*	32	134.8	[37]
57	Rice	*Oryza sativa*	29	430	[53]
58	Rice	*Oryza sativa*	22	430	[54]
59	Rice	*Oryza sativa*	30	372	[37]
60	Rubber tree	*Hevea brasiliensis*	30	1332	[55]
61	Salt cress	*Thellungiella halophile*	31	238.5	[13,37]
62	Shepherd's Purse	*Capsella rubella*	32	134.8	[37]
63	Sorghum	*Sorghum bicolor*	28	697.5	[37]
64	Soybean	*Glycine max*	39	1115	[56]
65	Soybean	*Glycine max*	50	1115	[57]
66	Soybean	*Glycine max*	39	1115	[58]
67	Soybean	*Glycine max*	41	978	[13,37]
68	Spikemosses	*Selaginella moellendorffii*	11	212.5	[13,37]
69	Spreading earthmoss	*Physcomitrella patens*	25	480	[13,37]
70	Sweet orange	*Citrus sinensis*	24	319	[37]
71	Switchgrass	*Panicum virgatum*	53	1358	[37]
72	Tobacco	*Nicotiana tabacum*	15	323.75	[59]
73	Tomato	*Solanum lycopersicum*	29	900	[60]
74	Tomato	*Solanum lycopersicum*	28	900	[37]
75	Tomato	*Solanum lycopersicum*	29	900	[61]
76	Wheat	*Triticum aestivum*	20	2125	[26]
77	Wild Strawberry	*Fragaria vesca*	14	240	[37]

Similarly, CPKs are also found in pollens, embryonic cells, guard cells, xylem, and meristem [36]. These Ca-dependent functional proteins are involved in biological functioning in cellular and subcellular compartments. Numerous CPKs of *Arabidopsis* are membrane-localized. It is considered that the myristylation causes CPKs to target the membrane [62]. This cellular and subcellular localization indicates a significant role of CPKs in several signaling transduction pathways under stress stimuli.

3.2. CPK Domain Organization and Calcium Ion Signal Decryption

On account of specific abiotic stress stimuli, the plant activates distinct physiological and biochemical response pathways. These stimuli are perceived by some protein and nonprotein elements. Protein elements include enzymes, transcription factors, and disparate receptors, while nonproteins comprise some secondary messengers such as calcium ion cyclic nucleotides, hydrogen ions, lipids, and active oxygen species [17,63]. Among them, Ca is a crucial secondary messenger involved in the signal transduction in all eukaryotes. It regulates the cell polarity and is essential for the regulation of stress-responsive cellular processes, cell morphogenesis, as well as plant growth and development [3,11,64,65]. These calcium signals are recognized by several protein kinases (CPKs), which regulate the response of downstream factors.

The CPK-encoding protein commonly has four functional domains, viz., calcium-binding domain (CBD), N terminus variable domain (NTD), protein kinase domain (PKD), and autoinhibitory junction (AJ), but many CPKs also contain an amino-terminal domain with varying sequence lengths, which is a source of functional diversity in the CPK family [62]. Sometimes, the C-terminus variable domain (CTD) also considered as a distinct domain instead of NTD. Different plant species contain varying numbers of CPK genes that are functionally important. The CBD contains four loops where calcium ions directly bind, called EF-hands, and are 20 amino acids in length [20,66–68]. The PKD domain has a characteristic serine/threonine phosphorylation site, which responds during regulation of CBD and AJ through Ca signals [68,69]. Among the number of CPK proteins, the majority of them have a myristylation site upstream from their N-terminal variable domain, showing that no CPKs appear in the form of membrane integral proteins [23]. The N-terminus of CPKs has a greater percentage of proline, glutamine, serine, and threonine (PEST) sequences, which carry out swift proteolytic degradation. There is an auto-inhibitory domain adjacent to the conserved domains, having a pseudo-substrate domain activity, and can cause inhibition of the regulatory pathways [68]. The variation in the length of CPK genes is due to the NTD, CT domain, and EF hand of the calcium-binding domain. Ca^{2+} through binding with the EF-hand motif, carries out the phosphorylation of the CPK substrate by removing autoinhibition of kinase activity [22,70]. The highly conserved calmodulin-like domain regulates all the activities of the CPKs by binding the four Ca^{2+} ions to four EF hands at its downstream end. Proteomics of most of the CPKs show that the autophosphorylation of proteins at serine and threonine through a calcium-dependent manner regulate the kinase activity (Figure 1).



(A) Structure of CPKs under inactive form

(B) Activation of CPKs under Ca^{2+} signal

Figure 1. Structure and activation process of plant CPKs. (**A**) CPK domain structure under the inactive state, (**B**) activation of CPKs after the binding of Ca^{2+} to the active site of the protein kinase domain (PKD), the autoinhibitory junction (AJ), and calmodulin-like domain (CaM-like domain, CaM-LD).

CPKs are monomolecular Ca-signaling protein kinases that regulate protein phosphorylation. In response to extrinsic and intrinsic cues, the variation in Ca^{2+} concentration, also called "Ca^{2+} signatures", is recognized, interpreted, and transduced to the downstream toolkit by a group of Ca^{2+}-binding proteins. Phosphorylation events cause the activation of CPKs.

3.3. Functional Characterization of Plant CPKs

CPKs are differentially involved in diverse and indispensable functions in various plant species. CPKs show their role against biotic and abiotic stress tolerance upon interaction with specific calcium signals. With respect to abiotic stresses, CPKs are involved in drought [71], salinity [72], and heat [73] and cold [74] stress response signaling by regulating the ABA-responsive transcriptional factors and ion channel regulation [75]. Some *Arabidopsis* CPKs (e.g., *CPK13*) are also involved in potassium ion (K^+) channel regulation and other ion transportation in guard cells [11]. CPKs are also a major participant for providing pathogen-related immunity to plants. In several plant species, CPKs enhance the resistance against fungal elicitors [1,76,77], bacterial invasions [78], and many other pathogen-related diseases [60,79]. Some CPKs are involved in the regulation of the jasmonic acid (JA)-dependent pathway during insect and plant interaction and indirectly regulate plant resistance against insects [80]. The crucial role of CPKs have also been reported in various growth and developmental processes in plants. CPK-encoding genes (*AtCPK28*) in *Arabidopsis* play a positive role in stem elongation and contribute to secondary growth by interacting with the gibberellic acid (GA) pathway [81,82]. Similarly, some CPKs regulate pollen tube growth [83], latex biosynthesis [55,84], higher biomass accumulation [85], wounding and herbivory attack [80,86], germination and seedling growth [87], early maturity [88,89], pigmentation and fruit development [90], and several other metabolic and developmental pathways [91]. Still, the role and functionality of various CPK-encoding genes against biotic and abiotic stresses are veiled.

4. Role of CPKs in Abiotic Stress Tolerance

CPKs are recognized as a key Ca sensor group of protein kinase, having a multigene family in the whole plant kingdom [55,92]. The functions of these CPKs are completely dependent on Ca^{2+} signatures. Most of CPK functionality has been identified only in vitro, which is why only specific stress response-associated functions are known [93]. CPKs are not only involved in ion channel regulation but also respond to multiple stress-related pathways through interactions with other distant transcription factors through phosphorylation. Several loss-of-function and gain-of-function studies have confirmed the role of CPKs in abiotic stress tolerance. The cytosolic Ca^{2+} concentration fluxes, induced by various environmental stresses, viz., heat [47], cold [94] light [95], drought [96,97], salt [72,98], and osmotic [99] and pathogen-related factors [100], activate the plant's transcriptional and metabolic activities [101]. Expression analyses and genome-wide studies have discovered the CPKs transcript activity, protein, and substrate recognition in different plant parts [93]. CPKs are also involved in the ABA-dependent abiotic stress signaling in various plant species. Several CPK genes are involved in the regulation of ABA signaling pathways in plants. Transient gene expression analyses in protoplasts of maize show that *CPK11* (closely related to *AtCPK4* and *AtCPK11*) acts upstream of mitogen-activated proteins (MPK5) and is required for the activation of defense functions and antioxidant enzyme activity by regulating the expression of MPK5 genes. Similarly, *CPK11* induced by hydrogen peroxide (H_2O_2) regulates and controls the activity of SOD and APX production induced by the ABA signaling pathway [102,103]. CPK activity confirmed by global expression analyses, shows that several CPK members are expressed differentially under varying ABA, salinity, drought, and heat and cold levels [93]. The change in the expression of CPK genes indicates the role of CPKs in plant adaptation against abiotic stress environments.

4.1. CPK-Mediated Drought Response Signaling

Drought stress is a major destructive factor affecting plant growth and development. It decreases water potential in plants as a result, where ABA accumulation controls the opening and closing of stomata, which leads to a lower photosynthetic activity [104]. It decreases the biomass and grain yield in plants. Under drought, plants adopt several conformational changes in the cell. These include ABA-dependent stomatal movement through regulation of guard cells, osmotic adjustments through the accumulation of osmolytes, regulating the oxidative damage by ROS homeostasis, and so on [93,105]. Changes in cytosolic Ca^{2+} concentrations due to water deficiency initiates CPK activity, resulting in the release of ABA in the cell [97]. ABA induces the injection of a calcium chelator (i.e., 1,2-bis (2-aminophenoxy) ethane-*N*,*N*,*N*′,*N*′-tetra acetic acid; BAPTA), into the guard cell, which causes the closing of the stomata and, eventually, control of the transpiration process. Several plant CPKs are involved in drought stress-response mechanisms through an ABA-dependent manner. The CPK-encoding gene (*CPK10*) of *Arabidopsis* and an identified interacting heat shock protein (HSP1) lead to a drought-sensitive genotype. *CPK10* T-DNA insertional mutants show sensitivity to drought stress as compared to the wild types. *AtCPK9* and *AtCPK10* are involved in Ca^{2+}-dependent ABA-mediated stomatal regulation through interaction with *AtCPK33* [106]. The light-induced *Arabidopsis* encoding gene (CPK13) is involved in inhibiting stomatal opening and contributes to the drought stress responsiveness [11]. Some drought-responsive CPKs also have some associated functions. In rice, for example, *OsCPK9* controls both drought stress tolerance and spikelet fertility through an ABA-dependent manner. Results of overexpression of *OsCPK9* (*OsCPK9*-OX) induces stomatal closure through osmotic adjustment and increases the pollen viability and spikelet fertility under polyethylene glycol (PEG-6000)-induced drought stress [71]. Another CPK-encoding gene from the wild grapevine (*CPK20*) acts as a regulator for drought and its associated with heat/cold responsive pathways. Expression of these genes studied in transgenic *Arabidopsis* reveals that *VaCPK20* overexpression exhibits a high level of tolerance to drought and cold stress through regulation of stress responder genes, viz., ABA-responsive element binding factor 3 (ABF3) or sodium/hydrogen exchanger 1 (NHX1), and cold regulator gene (*COR47*) [107]. While a CPK-encoding gene of broad bean (*VfCPK1*) reported being highly expressed in leaf epidermal peels, it is not considered a tissue-specific gene and is only expressed under drought stress [108]. This CPK-encoding gene shows no relationship with both high (37 °C) and low (4 °C) temperatures. The increase in the number of transcripts of *VfCPK1* under drought stress only plays a role in the up-regulation of ABA-responsive genes and other kinases that are involved in the signal transduction pathway [108].

Some CPKs are involved in the regulation of antioxidant production and osmolyte homeostasis to combat drought stress. *AtCPK8* regulates the movement of the stomatal guard cell and H_2O_2 homeostasis in response to cellular Ca^{2+}. An *Arabidopsis* T-DNA insertion mutant of *CPK8* was found to be more sensitive to drought stress as compared to the wild-type plant, which reveals their drought response functionality [97]. CPKs phosphorylate some interactional proteins and perform interactive functioning in plants. Under drought stress, *AtCPK8* with an interacting protein CAT3 controls the Ca^{2+}-dependent ABA-mediated regulation of stomatal guard cells. The CPK8 mutant was more sensitive to drought stress, while overexpressing CPK8 in transgenic plants exhibited tolerance [97,109]. *CaCPK1* activity increases the chickpea responsiveness to drought stress, and its activity is ubiquitous in all tissues of the plant [110]. The activation of drought-responsive CPK-encoding genes is also triggered by various biochemical pathways. A rice CPK-encoding gene (*OsCPK1*) specifically activated by sucrose starvation was involved in mechanism to prevent drought stress injury during germination by negatively regulating the expression of GA biosynthesis and activating the expression of a 14-3-3 protein ‘GF14c’ [111].

Some closely related CPK-encoding isoforms show functional diversity in response to drought stress. For example, functional divergence is present between two closely homologous (*TaCPK7* and *TaCPK12*) genes of wheat [112]. Functional analysis of *TaCPK7* and *TaCPK12* reveals that *TaCPK7* responded to H_2O_2, drought, salt, and low temperature, while T*aCPK12* responded only through the ABA signaling pathway [112]. Several transgenic studies have been conducted to characterize the functions of CPKs in different plant species in relation to drought stress response signaling in plants. The *ZoCDPK1* genes from ginger overexpressed in tobacco (*Nicotiana tabacum*) conferred drought as well as salinity tolerance by improving the photosynthesis and growth of the plant [113]. Enhanced expression of *ZoCDPK1* under drought and JA treatment was observed, but no variation was found in expression because of low-temperature stress and abscisic acid treatment. *ZoCDPK1* induces the expression of stress-responsive genes (i.e., early responsive to dehydration stress (*ERD1*) and responsive to dehydration (*RD21A*)). In ginger, it controls the stress signaling pathway and works in a CTR/DRE-independent manner [113]. Expression of CPK encoding genes of maize studied in *Arabidopsis* shows that *ZmCPK4* is involved in resistance to drought stress through ABA-regulated stomatal regulation. *ZmCPK4* induced by H_2O_2 and ABA treatment shows that there might be an association between mitogen-activated protein kinase (MAPKs) members and *ZmCPK4* in the upregulation of ABA-regulatory components, especially ABA-insensitive (ABI5), ABF3, and Ras-associated binding protein (RAB18) [87]. The functions of several drought-responsive CPK-encoding genes are summarized in Table 2. (Details of all the genes are given in Table S1)

Table 2. Various functions of CPKs in biotic and abiotic stresses in different plant species.

Sr. #	Specie Name	Gene	Function	Reference
1	*Arabidopsis thaliana*	*AtCPK1*	Cellular homeostasis, resistance fungal elicitor.	[76,78,114–116]
2		*AtCPK3*	Salt resistance.	[117,118]
3		*AtCPK4*	Regulate ABA-regulatory transcription factors (e.g., ABF, ABF4, drought resistance).	[98]
4		*AtCPK5*	Regulate immunity responses, ROS-dependent cell-to-cell communication.	[78]
5		*AtCPK6*	Drought tolerance, ABA-dependent osmotic adjustment.	[119]
6		*AtCPK8*	Drought tolerance through interaction with protein CAT3.	[97,109]
7		*AtCPK9*	Regulate the ABA-dependent signaling pathway interacting with *CPK33.*	[75]
8		*AtCPK10*	Drought responsiveness, ABA-mediated stomatal movements.	[106]
9		*AtCPK11*	Phosphorylation of AtDi19, ABA signaling.	[120]
10		*AtCPK12*	Seed germination, activation of ABA regulators.	[72,121]
11		*AtCPK16*	Root-gravitropism phosphorylate *AtACS7.*	[122]
12		*AtCPK21*	Hyperosmotic adjustments.	[123]
13		*AtCPK23*	Salt stress, drought stress.	[124]
14		*AtCPK27*	Salinity resistance, H_2O_2 and ionic homeostasis.	[125]
15		*AtCPK28*	Vascular development, stem elongation, ethylene synthesis, lignin deposition.	[81,82]
16		*AtCPK32*	ABA-regulatory gene activation.	[126]
17		*AtCPK33*	Regulates flowering, biosynthesis of florigen and flowering locus T protein.	[127]
18	*Cicer areitinum* (Chickpea)	*CaCPK1*	Salt stress, drought stress, phytohormones, and defense signaling pathways.	[110]
19		*CaCPK2*		

Table 2. *Cont.*

Sr. #	Specie Name	Gene	Function	Reference
20	*Capsicum annuum* (Peppers)	*CaCPK3*	Pathogen resistance, defense functioning (i.e., regulates jasmonic and salicylic acid), ethephon.	[79]
21	*Fragaria* x *ananassa* (Garden strawberry)	*FaCPK1*	low-temperature tolerance, fruit ripening.	[128]
22	*Medicago sativa* (Alfalfa)	*MsCPK3*	Heat stress resistance, embryogenesis.	[129]
23	*Oryza sativa* (Rice)	*OsCPK1*	Drought stress, seed germination, and GA biosynthesis.	[111]
24		*OsCPK4*	Microbial-associated immunity, OsRLCK176 degradation.	[130]
25		*OsCDPK5*	Fungal attacks phosphorylate OsERG1 and OsERG3.	[131]
26		*OsCPK9*	Drought stress tolerance, ABA sensitivity spikelet fertility.	[71]
27		*OsCPK10*	*Pseudomonas syringae pv* resistance, SA and JA regulator.	[132]
28		*OsCPK12*	Salt tolerance, blast disease resistance, induce ROS production, leaf senescence,	[1,133]
29		*OsCDPK13*	Regulate cold, salt, dehydration responses.	[134]
30		*OsCPK17*	Cold stress interacts with sucrose synthase and plasma membrane intrinsic proteins.	[135]
31		*OsCPK21*	Salt tolerance, ABA pathway activation.	[136]
32		*OsCPK24*	Cold stress tolerance, inhibition of OsGrx10.	[74]
33		*OsCPK31*	Starch accumulation, early grain filling.	[137]
34	*Nicotiana tabacum* (Tobacco)	*NtCPK1*	Signaling localization for repression of shoot growth, GA biosynthesis.	[138]
35		*NtCPK2*	Biotic stress immunity.	[139]
36		*NtCPK32*	Pollen tube growth interacts with CNGC18.	[83]
37	*Hevea brasiliensis* (Rubber tree)	*HbCDPK1*	Latex biosynthesis, rubber production.	[84]
38	*Panax ginseng* (Chinese ginseng)	*PgCDPK1a*	Regulate ginseng growth.	[85]
39	*Phalaenopsis amabilis* (Moth orchid)	*PaCPK1*	Cold stress sensitivity, wounding, pathogen attack.	[86]
40	*Triticum aestivum* (Wheat)	*TaCDPK1*	Regulate metabolic and developmental pathways.	[91]
41		*TaCPK7*	Drought stress, salt stress, ABA signaling pathway.	[112]
42		*TaCPK12*		
43	*Zingiber officinale* (Ginger)	*ZoCDPK1*	Salinity and drought stress tolerance.	[113]
44	*Zea mays* (Maize)	*ZmCPK1*	Cold stress regulates ZmERF3 expression.	[33]
45		*ZmCPK4*	Upregulate ABA-regulatory components (i.e., ABI5, ABF3 and RAB18) with MAPKs.	[87]
46		*ZmCPK11*	Superoxide dismutase and ascorbate peroxidase production, ABA pathway.	[103]

Table 2. *Cont.*

Sr. #	Specie Name	Gene	Function	Reference
47	*Vigna radiata* (Mung bean)	*VrCPK1*	Salt stress tolerance.	[140]
48	*Vicia faba* (Broad bean)	*VfCPK1*	Drought stress resistance.	[108]
49	*Solanum lycopersicum* (Tomato)	*SlCDPK2*	Flowering.	[141]
50		*SlCDPK10*	*Xanthomonas oryzae* pv. *oryzae* and *Pseudomonas syringae* resistance.	
51		*SlCDPK18*	*Xanthomonas oryzae* pv. *oryzae* and *Pseudomonas syringae* resistance.	[60]
52	*Solanum tuberosum* (Potato)	*StCPK1*	Tuber formation.	[142]
53		*StCPK4*	Fungal pathogen resistance, ROS production.	[143]
54		*StCDPK5*	Blight resistance and susceptibility, ROS defense functioning.	[100]
55		*StCDPK7*	Resistance against *Phytophthora infestans*.	[77]
56	*Nicotiana attenuate* (Coyote tobacco)	*NaCDPK4*	Wound-induced jasmonic acid (JA) accumulation, insect resistance.	[80]
57		*NaCDPK5*		
58	*Camellia sinensis* (Tea plant)	*CsCDPK20*	High-temperature stress resistance.	[144]
59		*CsCDPK26*		
60	*Hordeum vulgare* (Barley)	*HvCPK3*	Resistance against powdery mildew.	[145]
61		*HvCPK4*		
62	*Brassica napus* (Oilseed rape)	*BnaCPK2*	ROS accumulation, cell death.	[2]
63	*Musa acuminate* (Banana)	*MaCDPK7*	Heat-induced fruit ripening, chilling, stress tolerance.	[146]
64		*MaCDPK2*	Sensitive to Foc-TR4 infection, biotic stress tolerance.	[147]
65		*MaCDPK4*	Sensitive to Foc-TR4 infection, biotic stress tolerance.	
66		*MaCDPK3*	Responsive for drought, cold, and salinity.	
67	*Vitis amurensis* (Grapevine)	*VaCPK1*	Salt stress, heat-responsiveness, stilbene bio-synthesis.	[89,148]
68		*VaCPK26*	Salt stress, Stilbene bio-synthesis, through the induced expression of stilbene synthase (STS) genes.	[89,148]
69		*VaCPK20*	Drought stress, cold stress.	[107]
70		*VaCPK21*	Salt stress signaling.	[149]
71	*Pharbitis nil* (Picotee)	*PnCPK1*	Seed germination, seedling growth, flowering, regulation of light-dependent pathways, embryogenesis.	[90]
72	*Populus euphratica* (Desert poplar)	*PeCPK10*	Drought and cold stress tolerance, ABA-responsive genes regulator.	[150]
73	*Cucumis melo* (Hami melon)	*HmCDPK2*	Resistance against Penicillium infection.	[151]

4.2. CPKs-Mediated Salt Response Signaling

Salt stress is also a major abiotic factor limiting plant growth and global agricultural productivity. Salinity, mostly due to the accumulation of sodium Na^+ and chloride Cl^- ions, causes an ion imbalance that leads the plants toward oxidative stress [152]. These ions also induce the toxicity of other ions in plants. Salts also increases the production of ROS in plants. Several studies have presented the functioning of CPK-encoding genes in plants against salt stresses. In *Arabidopsis*, *AtCPK27* genes were found in favor of plant adaptation against salt stress [125]. Disruption in the expression of *CPK27* in a T-DNA insertional mutant shows salt hypersensitivity at early growth stages in Arabidopsis. *CPK27* regulated H_2O_2 and ionic homeostasis. *AtCPK3* functions in guard cell movement through osmotic adjustment and ion channel regulation during salt accumulation [11,117,118]. The overexpression of AtCPK*3* also increases ABA sensitivity and salt hypersensitivity, affecting the seedling growth and stomatal regulation [98,117]. *AtCPK6* belongs to a subclass of the CPK gene family in *Arabidopsis* whose expression is induced under salt-stressed conditions. *AtCPK6* and other kinases are activated because of cytoplasmic Ca^{2+} elevation in the calcium-dependent pathway, which depends on ABA. These kinases combined with *AtCPK6* trigger the salt and osmotic stress tolerance. Overexpression of *AtCPK6* in *Arabidopsis* increases the drought and salt tolerance in transgenic plants. RT-PCR analyses showed an increase in the expression of salt-regulated genes in plants, in which the *AtCPK6* gene was over-expressed [119].

OsCPK12 positively modulates salt stress tolerance, and it is associated with decreases in the resistance against blast disease by increasing the sensitivity to ABA and inducing the accumulation of ROS in rice [1]. In *Arabidopsis*, *AtCPK27* was found to be favorable for plant adaptation against salt stress. Disruption in the expression of *CPK27* in T-DNA insertional mutant shows salt hypersensitivity at early growth stages. Under salt stress, *CPK27* regulates H_2O_2 and ionic homeostasis and makes plants resistant to salt stress (Figure 2) [125].

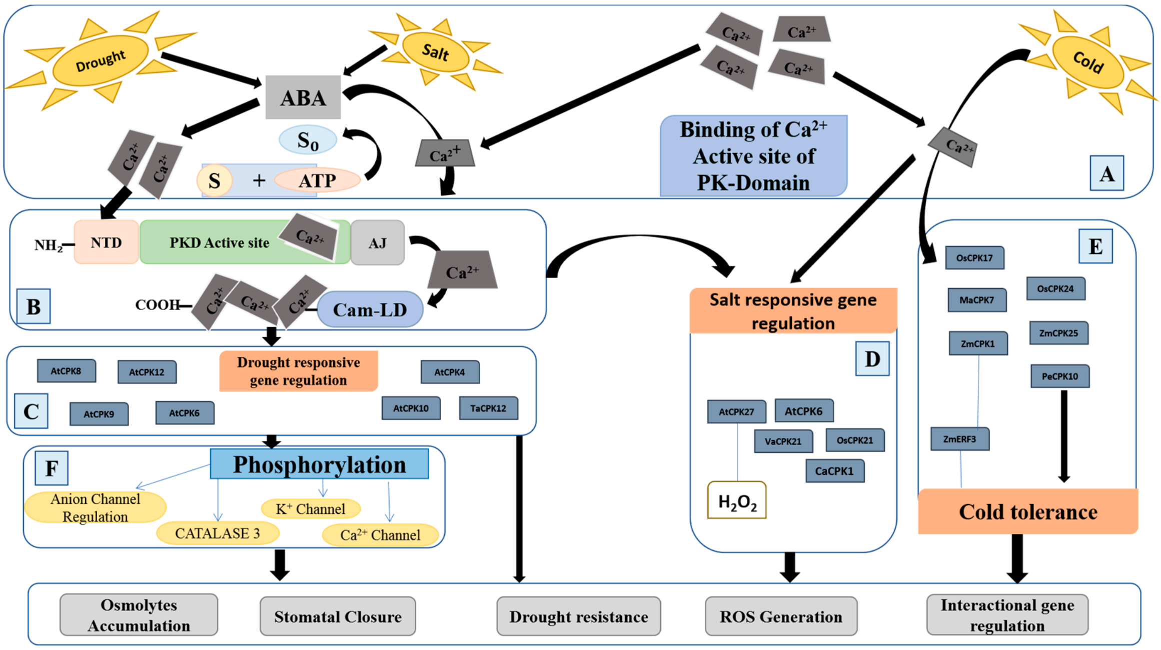


Figure 2. Role of different CPKs under various abiotic stresses; (**A**) Ca^{2+}-dependent ABA-mediated drought and salt stress signal recognition by CPKs; (**B**) Ca^{2+} binding at the active site of protein kinase domain (PKD); (**C**) some drought-responsive genes involved in metabolite regulation and signal transduction pathways; (**D**) some salt-responsive genes and their role in antioxidant production (i.e., H_2O_2), as well as ROS detoxification; (**E**) some cold stress-responsive genes and their interaction genes activation; and (**F**) phosphorylation events controlling the anion channel regulation, K^+-inward channel regulation, Ca^{2+}-concentration, and channel regulation in the cell, and ABA-mediated CATALASE 3 regulation in plant cells.

OsCPK21 genes regulate the ABA-dependent salt stress signaling pathway. The high survival rate of transgenic rice seedlings developed by a mini scale, full-length cDNA over-expresser (FOX) gene hunting system was found due to the overexpression of *OsCPK21*-FOX under salt stress. In these plants, many salt-induced and ABA-regulating genes were expressed more as compared to wild-type plants.

Overexpression of *OsCPK21* increases exogenous ABA and enhances salt tolerance by regulating and inducing the salt tolerance genes [136].

VaCPK21 gene up-regulation is positively involved in salt stress-response signaling mechanisms in grapevines. Overexpression of this gene in transgenic *Arabidopsis* and *V. amurensis* callus cell lines shows that under the salt stress, *VaCPK21* acts as a regulator for genes that respond to salt stress (i.e., *AtRD26*, kinase-like protein (*AtKIN1), AtRD29B*, *AtNHX1*, catalase (*AtCAT1*), copper superoxide dismutase (*AtCSD1*), cold regulator (*AtCOR15* and *AtCOR15*)), and are found functionally important for salt stress tolerance [149]. Similarly, *CaCPK1* and *CaCPK2* activities are enhanced during high salt stress in leaves of chickpea plants. These isoforms play a role in the regulation of phytohormones and defense signaling pathways [110].

4.3. CPK-Dependent Cold and Heat Stress Signaling

Several CPK-encoding genes are differentially expressed under cold and heat treatments, but their exact molecular response mechanism is still unknown. *OsCPK17* was reported to be important for the cold stress response by targeting the sucrose synthase and plasma membrane intrinsic proteins in rice [135]. *OsCPK24* causes inhibition of glutaredoxin (OsGrx10) to sustain higher glutathione levels and phosphorylation, through the Ca^{2+} signaling pathway, and responds positively to cold stress tolerance in rice [74]. *MaCDPK7* was found as a positive regulator of heat-induced fruit ripening and chilling stress tolerance in bananas [146].

PeCPK10 provides cold and drought stress tolerance through ABA-induced stomatal closing in *P. euphratica.* Its constitutive expression regulates ABA-responsive genes (i.e., *RD29B* and *COR15A*) that regulate the cellular functioning. Transgenic *Arabidopsis* with over-expressed *PeCPK10* showed lower water loss under drought stress and tolerance against freezing. Expression analyses reveal that *PeCPK10* localizes in cytoplasm quickly in response to changes in Ca^{2+} concentrations and regulates the stomata guard cells, while nuclear-localized *PeCPK10* only regulates the transcriptional factors [150]. *CPK16* and *CPK32* in grapevine plants positively regulate stilbene (a phenolic secondary metabolite) biosynthesis and CPK30 individually involved in both cold and drought tolerance [153]. In maize, *ZmCPK1* and *ZmCPK25* gene expressions were increased or decreased, respectively, upon exposure to cold stress. *ZmCPK1* is negatively related with the regulation of the cold stress signaling mechanism. Studies of transgenic *Arabidopsis* also show that *ZmCPK1* inversely regulates the expression of ethylene response factor (*ZmERF3*) genes and impairs cold stress tolerance [33]. *CsCDPK20* and *CsCDPK26* act as regulatory factors for heat stress-responsive genes and control positive heat stress signaling in the tea plant [144].

4.4. Role of CPKs in ROS Detoxification

Drought, salt, and heat stress triggers ROS production in plants, which must be detoxified by the plant to prevent itself from oxidative stress. Mitochondria, chloroplasts, and peroxisomes are the central organelles for ROS accumulation [105,154]. ABA-induced ROS production in plants is reported to be dependent on nicotinamide adenine dinucleotide phosphate hydrogen (NADPH) oxidase [105], which plays a vital role in oxidative bursting and activating plant defense responses [155,156]. Plant CPKs have been reported to regulate ROS production [2]. For instance, *StCPK4* functions in the phosphorylation of NADPH oxidase and indirectly regulates ROS accumulation [143]. In *B. napus*, *BnaCPK2* controls the activity of the respiratory burst oxidase homolog protein D (RbohD) during cell death and ROS production [2]. Arabidopsis *CPK32* interacts with ABF4 in the ABA signaling pathway [126]. *AtCPK6* from *Arabidopsis* decreases ROS production by reducing lipid peroxidation and confers drought stress [119]. Likewise, *OsCPK12* promotes salt stress tolerance in rice through decreasing ROS accumulation [1]. The other CPKs and ROS responses are summarized in Table 2.

5. Functional Interaction of CPKs with Other Kinases in Abiotic Stress Signaling

CPK crosstalk and several interactions have been revealed in molecular regulatory pathways by functional studies. CPKs are not only involved in specific stress responses but also in multiple stress-related pathways by interacting with other distant proteins and regulating phosphorylation events. In *Arabidopsis*, *CPK28* supports the turnover and phosphorylation of plasma membrane-related receptor-like cytoplasmic kinase (botrytis-induced kinase 1, BIK1), an important convergent substrate of multiple pattern recognition receptor (PRR) complexes for plant immunity [36]. *AtCPK8* regulates and phosphorylates CAT3. It is involved in Ca^{2+}-dependent ABA and H_2O_2-induced guard cell regulation and provides drought resistance [97,109]. Molecular responses of *AtCPK1* studied by using real-time PCR (RT-PCR) show that the investigated gene expressions, viz., pyrroline-5-carboxylate synthetase 1(*P5CS1*), galactinol synthase 1(*GOLS1), RD22* (dehydration-responsive protein)*, RD29A*, C-repeat binding factor (CBF4), and *KIN2* (kinases), were upregulated by *ATCPK1* and conferred salinity stress tolerance [157]. Further, *AtCPK1* in loss-of-function and gain-of-function mutants were studied. It provides salt and drought stress resistance by up and down-regulation of stress responder genes, viz., zinc finger protein (*ZAT10*), *APX2*, *COR15A*, and *RD29A* [157]. *AtCPK12* phosphorylates several salt stress response-related proteins during regulatory functioning [72]. Another grapevine gene (*VaCPK21*) transgenically expressed in *Arabidopsis* interacts with several salt stress-related genes (i.e., *AtRD29*, *AtRD26*, *AtKIN1*, *AtNHX1*, *AtCSD1*, *AtCAT1*, *AtCOR15*, and *AtCOR47*). Likewise, *VaCPK20* responds to cold and drought stress tolerance by regulating *COR47*, *NHX1*, *KIN1*, or *ABF3* in transgenic *Arabidopsis* [107,149].

In vivo interaction validated by co-immunoprecipitation assays (Co-IP) revealed that *OsCPK4*, a dual-face protein, was involved in the regulation of the stability of cytoplasmic kinase (*CPK176*) in rice. *OsCPK4* plays a vital role in the negative regulation of receptor-like *OsCPK176* accumulation. *OsCPK4* and *OsCPK176* phosphorylation events provide pattern-triggered immunity [130]. *OsCPK17* phosphorylates the sucrose-phosphate synthase (*OsSPS4*) and plasma membrane intrinsic proteins (*OsPIP2;1* and *OsPIP2;6*) (aquaporin), which are essential in sugar metabolism and membrane channel activity against cold stress responses in rice [135]. Moreover, *OsCPK24* is involved in the phosphorylation of glutathione-dependent thioltransferase and inhibition of *OsGRX10* to maintain a higher level of glutathione. This regulatory pathway induces the overall cold stress responsiveness in rice [74]. The plant CPK-encoding genes also induce the regulation of other stress-responsive genes, viz., *AtRBOHF*, *AtRBOHD*, *AtABI1*, *AtRAB18*, *AtRD29B*, *AtHSP101*, *AtHSP70*, *Arabidopsis* heat stress transcription factor A2 (*AtHSFA2*), *AtP5CS2*, proline transporter (*AtProT1*), *AtPOD*, and *AtAPX1* for drought, salt, heat and cold stresses [11]. In tea plants, *CsCDPK20* and *CsCDPK26* have an interactive function for thermo-tolerance [144]. BnaCPK2 interacts with NADPH oxidase-like RbohD and controls ROS accumulation and cell death in oilseed rape [2]. In *Arabidopsis*, *CPK9* controls the ABA ion channel regulation through a Ca^{2+}-dependent manner. Overexpression studies revealed that CPK9 and CPK33 mutually controlled the regulation of guard cells and stomatal movement [75]. *CPK16* and *CPK32* in grapevine plants positively regulate stilbene (a phenolic secondary metabolite) biosynthesis and *CPK30* individually involved in both drought and cold tolerance [153]. Moreover, *VaCPK1* and *VaCPK26* genes are also involved in the same regulatory pathway ([89]. The overexpression of VaCPK29 up-regulates stress-responsive genes (i.e., dehydration elements (DREs) *AtABF3*, *AtDREB1A*, *AtDREB2A*, *AtRD29A*, and *AtRD29B*), which provide resistance to heat as well as osmotic stress [73]. Under in vitro conditions, post-transcriptionally miR390-regulated *StCDPK1* controls the downstream auxin efflux carrier of PIN-proteins (*StPIN4*), which are involved in potato tuber development [142].

Arabidopsis CPKs interact and phosphorylate the basic leucine zipper domain (bZIP) transcription factor FD and have a crucial role in florigen complex formation, which induces late flowering in plants [127]. Biochemical analyses show that the cold-induced marker gene (*Zmerf3*), which is a type II ethylene response factor, is suppressed by *ZmCPK1* in maize. It is supposed that the *ZmCPK1* directly phosphorylates the ERF3 protein and, as a result, inactivates ERF and has a negative role in the cold stress response [33]. *ZmCPK11* controls the upstream *ZmMPK5*, which is involved in ABA-dependent defense-related signaling in maize. CPK-encoding genes also have several interactive functions concerning plant growth and development. In *Xenopus* oocytes, AtCPK32 potentially regulates the cyclic nucleotide-gated ion channel regulating gene (*CNGC18*). AtCPK32 stimulation of CNGC18 regulates pollen tube depolarization in *Arabidopsis* [83]. Constitutively active *OsCDPK1* in gain and loss-of-function transgenic rice targets the G-box factor 14-3-3c protein (GF14c). The expression of this protein causes the biosynthesis of GA and improves drought tolerance in rice seedlings [111]. AtCPK28 seems to be a regulatory component for the control of stem length and vascular development in *Arabidopsis*. The mutant of CPK28 (i.e., cpk28) was involved in the altered expression of NAC transcriptional regulators, such as NST1 and NST3, as well as gibberellin-3-beta-oxigenase 1 (GA3ox1), a regulator of gibberellic acid homeostasis [81]. After ABA treatment, the dual functioning *OsCPK9*-OX in rice increases the transcript levels of drought and spikelet fertility-responsive genes, viz., *OsRSUS*, *Rab21*, *Osbzip66*, and *OsNAC45*. The results confirmed by quantitative reverse transcription polymerase chain reaction (qRT-PCR) demonstrate that *OsCPK9* in interacting with these genes switches on the molecular regularization of ABA and stress-associated pathways [71]. The *ZoCDPK1* gene from ginger promotes the expression of drought and salinity stress associated genes, viz., RD2A (dehydration responsive protein 2A) and ERD1 (early responsive to dehydration stress 1) in tobacco. This DRE/CRT-independent regulatory pathway improves photosynthesis and plant growth as well [113]. Constitutive expression of calcium-dependent protein kinase of *Populus euphratica* (*PeCPK10*) regulates (*RD29B* and *COR15A*) cold and drought genes [150]. This cross-talk between CPK isoforms and the interactive partners increases the complexities among the signaling pathways.

6. Conclusions

The multifaceted role of CPKs in plants is consequential for abiotic stress tolerance in plants. Regardless of the reported functional detail on CPK-encoding genes, there are many other important isoforms identified whose expression profiles and involvement in abiotic stress signal transduction pathways in plants are still not clearly known. Future research is required to extend and identify the remaining CPK-encoding genes, their interactional regulators, and their functional exploration with respect to abiotic stress responses. These research studies are helpful to improve the plant's adaptation under unpredictable environments and to minimize threats to the world's food security.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/20/21/5298/s1.

Author Contributions: Conceptualization, R.M.A. and L.S.; Validation, G.C., M.W., M.A.R.R., F.A., and S.H.W.; Investigation, L.S., M.W., and F.A.; Resources, M.W., B.A., M.A.R.R., M.A.N., and S.H.W.; Writing—original draft preparation, R.M.A. and L.S.; Writing—review and editing, M.A.R.R., L.S. and M.A.N.; Visualization, G.C., M.W., M.A.R.R.; Supervision, L.S.; Project administration, R.M.A.; Funding acquisition, R.M.A.

Funding: This research was funded by the Punjab Agricultural Research Board (PARB) of Pakistan under the project grant # PARB-938.

Acknowledgments: The authors wish to acknowledge the Center for Advance Studies in Agriculture and Food Security (CAS-AFS) at the University of Agriculture, Faisalabad Pakistan for providing the research opportunity to the Master's Candidates of Plant Breeding and Genetics.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

ABA	Abscisic acid
AJ	Autoinhibitory junction
APX	Ascorbate peroxidase
AtABI1	*Arabidopsis thaliana* ABA-insenstive-1
AtCDPK/AtCPK	*Arabidopsis thaliana*-calcium dependent protein
AtCSD1	*Arabidopsis thaliana* copper superoxide dismutase 1
AtHSP101 and 70	*Arabidopsis thaliana* heat shock protein 101 and 70
AtProT1	*Arabidopsis thaliana* proline transporter 1
AtRBOHD	*Arabidopsis thaliana* respiratory burst oxidase protein D
AtRBOHF	*Arabidopsis thaliana* respiratory burst oxidase protein F
BaCDPK/BaCPK	*Brassica napus* calcium dependent protein kinase
BIK1	Botrytis-induced kinase 1
bZIP	Basic leucine zipper domain
Ca	Calcium
Ca2+	Calcium ion
CaCDPK/CaCPK	*Cicer arietinum* calcium dependent protein kinase
CaMs	Calmodulins
CAT	Catalase
CAT3	Catalase-3
CBD	Calcium binding domain
CBF4	C-repeat binding factor 4
CBLs	Calcineurin β-like proteins
CCaMK	Calcium/Calmodulin-dependent protein kinase
CDPKs/CPKs	Calcium dependent protein kinases
CMLs	Calmodulin-like protein Kinase
CNGC18	Cyclic nucleotide-gated ion channel 18
Co-IP	Co-immunoprecipitation assay
COR	Cold regulator
CT	C-terminus
CTR	C-repeat
DRE	Dehydration elements
EF	Elongation Factor
ERD1	Early responsive to dehydration stress 1
ERF3	Ethylene response factor 3
FaCDPK/FaCPK	*Fragaria* x *ananassa* calcium dependent protein kinase
GA	Gibberellic acid
GA3ox1	Gibberellin-3-betaoxigenase 1
GOLS1	Galactinol synthase 1
GPX	Glutathione peroxidase
GR	Glutathione reductase
H_2O_2	Hydrogen peroxide
HSF	Heat stress transcription factor
HSP	Heat shock protein
HvCDPK/HvCPK	*Hordeum vulgare* calcium dependent protein kinase
JA	Jasmonic acid
K^+	Potassium ion
LeCDPK/LeCPK	*Solanum lycopersicum* calcium dependent protein kinase
MaCDPK/MaCPK	*Musa acuminate* calcium dependent protein kinase
MPK5	Mitogen-activated protein kinase 5
MsCPK	*Medicago sativa* calcium dependent protein kinase
NaCDPK/NaCPK	*Nicotiana attenuate* calcium dependent protein kinase

NADPH	Nicotinamide Adenine Dinucleotide Phosphate Hydrogen
NHX	Sodium/Hydrogen exchanger
NST	NAC-transcription factors
NtCDPK/NtCPK	*Nicotiana tabacum* calcium dependent protein kinase
N-VD	N-terminus variable domain
OsCPK/OsCDPK	*Oryza sativa* calcium dependent protein kinase
OsGrx10	*Oryza sativa* glutaredoxin 10
OX	Overexpression
P5CS1	Pyrroline-5-carboxylate synthetase 1
PaCDPK/PaCPK	*Phalaenopsis amabilis* calcium dependent protein kinase
PeCDPK/PeCPK	*Populus euphratica* calcium dependent protein kinase
PEG	Polyethylene glycol
PEST	Proline, glutamine, serine and threonine
PgCDPK/PgCPK	*Panax ginseng* calcium dependent protein kinase
PIP	Plasma membrane intrinsic protein
PKD	Protein kinase domain
PnCDPK/PnCPK	*Populus euphratica* calcium dependent protein kinase
PRR	Pattern recognition receptor
qRT-PCR	Quantitative reverse transcription Polymerase chain reaction
RAB18	Ras-associated binding protein 18
RbohD	Respiratory burst oxidase homolog protein D
RD2A	Dehydration responsive protein 2A
RD29A	Dehydration responsive protein 29A
ROS	Reactive oxygen species
RT-PCR	Real-time PCR
SiCDPK/SiCPK	*Setaria italic* calcium dependent protein kinase
SOD	Superoxide dismutase
StCDPK/StCPK	*Solanum tuberosum* calcium dependent protein kinase
TaCDPK/TaCPK	*Triticum aestivum* calcium dependent protein kinase
T-DNA	Transfer DNA
VaCDPK/VaCPK	*Vitis amurensis* calcium dependent protein kinase
VaCPK/VaCDPK	*Vitis amurenssis* calcium dependent protein kinase
VfCPK/VfCDPK	*Vicia faba* calcium dependent protein kinase
VrCDPK/VrCPK	*Vigna radiata* calcium dependent protein kinase
ZmCDPK1/ZmCPK1	*Zea mays* calcium dependent protein kinase 1
ZoCDPK/ZoCPK	*Zingiber officinale* calcium dependent protein kinase

References

1. Asano, T.; Hayashi, N.; Kobayashi, M.; Aoki, N.; Miyao, A.; Mitsuhara, I.; Ichikawa, H.; Komatsu, S.; Hirochika, H.; Kikuchi, S. A rice calcium-dependent protein kinase OsCPK12 oppositely modulates salt-stress tolerance and blast disease resistance. *Plant J.* **2012**, *69*, 26–36. [CrossRef] [PubMed]
2. Wang, W.; Zhang, H.; Wei, X.; Yang, L.; Yang, B.; Zhang, L.; Li, J.; Jiang, Y.-Q. Functional characterization of calcium-dependent protein kinase (CPK) 2 gene from oilseed rape (*Brassica napus* L.) in regulating reactive oxygen species signaling and cell death control. *Gene* **2018**, *651*, 49–56. [CrossRef] [PubMed]
3. Sanders, D.; Pelloux, J.; Brownlee, C.; Harper, J.F. Calcium at the crossroads of signaling. *Plant Cell* **2002**, *14*, S401–S417. [CrossRef] [PubMed]
4. Li, A.; Wang, X.; Leseberg, C.H.; Jia, J.; Mao, L. Biotic and abiotic stress responses through calcium-dependent protein kinase (CDPK) signaling in wheat (*Triticum aestivum* L.). *Plant Sig. Behav.* **2008**, *3*, 654–656. [CrossRef]
5. Asano, T.; Hayashi, N.; Kikuchi, S.; Ohsugi, R. CDPK-mediated abiotic stress signaling. *Plant Sig. Behav.* **2012**, *7*, 817–821. [CrossRef]
6. Miao, Y.; Lv, D.; Wang, P.; Wang, X.-C.; Chen, J.; Miao, C.; Song, C.-P. An *Arabidopsis* glutathione peroxidase functions as both a redox transducer and a scavenger in abscisic acid and drought stress responses. *Plant Cell* **2006**, *18*, 2749–2766. [CrossRef]

7. Ahmad, P.; Jaleel, C.A.; Salem, M.A.; Nabi, G.; Sharma, S. Roles of enzymatic and nonenzymatic antioxidants in plants during abiotic stress. *Crit. Rev. Biotechnol.* **2010**, *30*, 161–175. [CrossRef]
8. Jiang, M.; Zhang, J. Water stress-induced abscisic acid accumulation triggers the increased generation of reactive oxygen species and up-regulates the activities of antioxidant enzymes in maize leaves. *J. Exp. Bot.* **2002**, *53*, 2401–2410. [CrossRef]
9. Hu, X.; Jiang, M.; Zhang, A.; Lu, J. Abscisic acid-induced apoplastic H_2O_2 accumulation up-regulates the activities of chloroplastic and cytosolic antioxidant enzymes in maize leaves. *Planta* **2005**, *223*, 57. [CrossRef]
10. Baba, A.I.; Rigó, G.; Andrási, N.; Tietz, O.; Palme, K.; Szabados, L.; Cséplő, Á. *Striving Towards Abiotic Stresses: Role of the Plant CDPK Superfamily Members*; Springer: Berlin, Germany, 2019; pp. 99–105.
11. Shi, S.; Li, S.; Asim, M.; Mao, J.; Xu, D.; Ullah, Z.; Liu, G.; Wang, Q.; Liu, H. The *Arabidopsis* calcium-dependent protein kinases (CDPKs) and their roles in plant growth regulation and abiotic stress responses. *Int. J. Mol. Sci.* **2018**, *19*, 1900. [CrossRef]
12. Ray, S.D. *Decrypting Calcium Signaling in Plants: The Kinase Way*; Springer: Berlin, Germany, 2015; pp. 119–174.
13. Mohanta, T.K.; Yadav, D.; Khan, A.L.; Hashem, A.; Abd Allah, E.F.; Al-Harrasi, A. Molecular Players of EF-hand Containing Calcium Signaling Event in Plants. *Int. J. Mol. Sci.* **2019**, *20*, 1476. [CrossRef] [PubMed]
14. Reddy, A.S.; Ali, G.S.; Celesnik, H.; Day, I.S. Coping with stresses: Roles of calcium-and calcium/calmodulin-regulated gene expression. *Plant Cell* **2011**, *23*, 2010–2032. [CrossRef] [PubMed]
15. Luan, S.; Kudla, J.; Rodriguez-Concepcion, M.; Yalovsky, S.; Gruissem, W. Calmodulins and calcineurin B–like proteins: Calcium sensors for specific signal response coupling in plants. *Plant Cell* **2002**, *14*, S389–S400. [CrossRef] [PubMed]
16. Hashimoto, K.; Kudla, J. Calcium decoding mechanisms in plants. *Biochimie* **2011**, *93*, 2054–2059. [CrossRef]
17. Valmonte, G.R.; Arthur, K.; Higgins, C.M.; MacDiarmid, R.M. Calcium-dependent protein kinases in plants: Evolution, expression and function. *Plant Cell Physiol.* **2014**, *55*, 551–569. [CrossRef] [PubMed]
18. Hetherington, A.; Trewavas, A. Activation of a pea membrane protein kinase by calcium ions. *Planta* **1984**, *161*, 409–417. [CrossRef]
19. Harmon, A.C.; Gribskov, M.; Harper, J.F. CDPKs–A kinase for every Ca^{2+} signal? *Trends Plant Sci.* **2000**, *5*, 154–159. [CrossRef]
20. Cheng, S.-H.; Willmann, M.R.; Chen, H.-C.; Sheen, J. Calcium signaling through protein kinases. The *Arabidopsis* calcium-dependent protein kinase gene family. *Plant Physiol.* **2002**, *129*, 469–485. [CrossRef]
21. Rudd, J.J.; Franklin-Tong, V.E. Unravelling response-specificity in Ca^{2+} signalling pathways in plant cells. *New Phytol.* **2001**, *151*, 7–33. [CrossRef]
22. Boudsocq, M.; Sheen, J. CDPKs in immune and stress signaling. *Trends Plant Sci.* **2013**, *18*, 30–40. [CrossRef]
23. Simeunovic, A.; Mair, A.; Wurzinger, B.; Teige, M. Know where your clients are: Subcellular localization and targets of calcium-dependent protein kinases. *J. Exp. Bot.* **2016**, *67*, 3855–3872. [CrossRef] [PubMed]
24. Initiative, A.G. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **2000**, *408*, 796. [CrossRef] [PubMed]
25. Ray, S.; Agarwal, P.; Arora, R.; Kapoor, S.; Tyagi, A.K. Expression analysis of calcium-dependent protein kinase gene family during reproductive development and abiotic stress conditions in rice (*Oryza sativa* L. ssp. *indica*). *Mol. Genet. Genom.* **2007**, *278*, 493–505. [CrossRef] [PubMed]
26. Li, A.-L.; Zhu, Y.-F.; Tan, X.-M.; Wang, X.; Wei, B.; Guo, H.-Z.; Zhang, Z.-L.; Chen, X.-B.; Zhao, G.-Y.; Kong, X.-Y. Evolutionary and functional study of the CDPK gene family in wheat (*Triticum aestivum* L.). *Plant Mol. Biol.* **2008**, *66*, 429–443. [CrossRef] [PubMed]
27. Wang, N.; Xia, E.-H.; Gao, L.-Z. Genome-wide analysis of WRKY family of transcription factors in common bean, *Phaseolus vulgaris*: Chromosomal localization, structure, evolution and expression divergence. *Plant Gene* **2016**, *5*, 22–30. [CrossRef]
28. Liu, W.; Li, W.; He, Q.; Daud, M.K.; Chen, J.; Zhu, S. Genome-wide survey and expression analysis of calcium-dependent protein kinase in *Gossypium raimondii*. *PLoS ONE* **2014**, *9*, e98189. [CrossRef]
29. Zhang, K.; Han, Y.-T.; Zhao, F.-L.; Hu, Y.; Gao, Y.-R.; Ma, Y.-F.; Zheng, Y.; Wang, Y.-J.; Wen, Y.-Q. Genome-wide identification and expression analysis of the CDPK gene family in grape, *Vitis* spp. *BMC Plant Biol.* **2015**, *15*, 164. [CrossRef]
30. Hu, W.; Hou, X.; Xia, Z.; Yan, Y.; Wei, Y.; Wang, L.; Zou, M.; Lu, C.; Wang, W.; Peng, M. Genome-wide survey and expression analysis of the calcium-dependent protein kinase gene family in cassava. *Mol. Genet. Genom.* **2016**, *291*, 241–253. [CrossRef]

31. Yang, Y.; Wang, Q.; Chen, Q.; Yin, X.; Qian, M.; Sun, X.; Yang, Y. Genome-wide survey indicates diverse physiological roles of the barley (*Hordeum vulgare* L.) calcium-dependent protein kinase genes. *Sci. Rep.* **2017**, *7*, 5306. [CrossRef]
32. Zhang, H.; Wei, C.; Yang, X.; Chen, H.; Yang, Y.; Mo, Y.; Li, H.; Zhang, Y.; Ma, J.; Yang, J. Genome-wide identification and expression analysis of calcium-dependent protein kinase and its related kinase gene families in melon (*Cucumis melo* L.). *PLoS ONE* **2017**, *12*, e0176352. [CrossRef]
33. Weckwerth, P.; Ehlert, B.; Romeis, T. Zm CPK 1, a calcium-independent kinase member of the *Zea mays* CDPK gene family, functions as a negative regulator in cold stress signalling. *Plant Cell Environ.* **2015**, *38*, 544–558. [CrossRef] [PubMed]
34. Zuo, R.; Hu, R.; Chai, G.; Xu, M.; Qi, G.; Kong, Y.; Zhou, G. Genome-wide identification, classification, and expression analysis of CDPK and its closely related gene families in poplar (*Populus trichocarpa*). *Mol. Biol. Rep.* **2013**, *40*, 2645–2662. [CrossRef] [PubMed]
35. Ye, S.; Wang, L.; Xie, W.; Wan, B.; Li, X.; Lin, Y. Expression profile of calcium-dependent protein kinase (CDPKs) genes during the whole lifespan and under phytohormone treatment conditions in rice (*Oryza sativa* L. ssp. *indica*). *Plant Mol. Biol.* **2009**, *70*, 311–325. [CrossRef] [PubMed]
36. Monaghan, J.; Matschi, S.; Shorinola, O.; Rovenich, H.; Matei, A.; Segonzac, C.; Malinovsky, F.G.; Rathjen, J.P.; MacLean, D.; Romeis, T. The calcium-dependent protein kinase CPK28 buffers plant immunity and regulates BIK1 turnover. *Cell Host Microbe* **2014**, *16*, 605–615. [CrossRef]
37. Mohanta, T.K.; Mohanta, N.; Mohanta, Y.K.; Bae, H. Genome-wide identification of calcium dependent protein kinase gene family in plant lineage shows presence of novel DxD and DEL motifs in EF-hand domain. *Front. Plant Sci.* **2015**, *6*, 1146. [CrossRef]
38. Li, M.; Hu, W.; Ren, L.; Jia, C.; Liu, J.; Miao, H.; Guo, A.; Xu, B.; Jin, Z. Identification, Expression, and Interaction Network Analyses of the CDPK Gene Family Reveal Their Involvement in the Development, Ripening, and Abiotic Stress Response in Banana. *Biochem. Genet.* **2019**, 1–23. [CrossRef]
39. Fedorowicz-Strońska, O.; Koczyk, G.; Kaczmarek, M.; Krajewski, P.; Sadowski, J. Genome-wide identification, characterisation and expression profiles of calcium-dependent protein kinase genes in barley (*Hordeum vulgare* L.). *J. App. Genet.* **2017**, *58*, 11–22. [CrossRef]
40. Zhang, H.; Liu, W.-Z.; Zhang, Y.; Deng, M.; Niu, F.; Yang, B.; Wang, X.; Wang, B.; Liang, W.; Deyholos, M.K. Identification, expression and interaction analyses of calcium-dependent protein kinase (CPK) genes in canola (*Brassica napus* L.). *BMC Genom.* **2014**, *15*, 211. [CrossRef]
41. Tong, X.; Cao, A.; Wang, F.; Chen, X.; Xie, S.; Shen, H.; Jin, X.; Li, H. Calcium-Dependent Protein Kinase Genes in *Glycyrrhiza Uralensis* Appear to be Involved in Promoting the Biosynthesis of Glycyrrhizic Acid and Flavonoids under Salt Stress. *Molecules* **2019**, *24*, 1837. [CrossRef]
42. Gao, W.; Xu, F.-C.; Guo, D.-D.; Zhao, J.-R.; Liu, J.; Guo, Y.-W.; Singh, P.K.; Ma, X.-N.; Long, L.; Botella, J.R. Calcium-dependent protein kinases in cotton: Insights into early plant responses to salt stress. *BMC Plant Biol.* **2018**, *18*, 15. [CrossRef]
43. Xu, X.; Liu, M.; Lu, L.; He, M.; Qu, W.; Xu, Q.; Qi, X.; Chen, X. Genome-wide analysis and expression of the calcium-dependent protein kinase gene family in cucumber. *Mol. Genet. Genom.* **2015**, *290*, 1403–1414. [CrossRef] [PubMed]
44. Singh, U.M.; Chandra, M.; Shankhdhar, S.C.; Kumar, A. Transcriptome wide identification and validation of calcium sensor gene family in the developing spikes of finger millet genotypes for elucidating its role in grain calcium accumulation. *PLoS ONE* **2014**, *9*, e103963. [CrossRef] [PubMed]
45. Yu, T.-F.; Zhao, W.-Y.; Fu, J.-D.; Liu, Y.-W.; Chen, M.; Zhou, Y.-B.; Ma, Y.-Z.; Xu, Z.-S.; Xi, Y.-J. Genome-wide analysis of CDPK family in foxtail millet and determination of SiCDPK24 functions in drought stress. *Front. Plant Sci.* **2018**, *9*, 651. [CrossRef] [PubMed]
46. Chen, F.; Fasoli, M.; Tornielli, G.B.; Dal Santo, S.; Pezzotti, M.; Zhang, L.; Cai, B.; Cheng, Z.-M. The evolutionary history and diverse physiological roles of the grapevine calcium-dependent protein kinase gene family. *PLoS ONE* **2013**, *8*, e80818. [CrossRef]
47. Dubrovina, A.S.; Kiselev, K.V.; Khristenko, V.S. Expression of calcium-dependent protein kinase (CDPK) genes under abiotic stress conditions in wild-growing grapevine *Vitis amurensis*. *J. Plant Physiol.* **2013**, *170*, 1491–1500. [CrossRef]
48. Ma, P.; Liu, J.; Yang, X.; Ma, R. Genome-wide identification of the maize calcium-dependent protein kinase gene family. *App. Biochem. Biotechnol.* **2013**, *169*, 2111–2125. [CrossRef]

49. Kong, X.; Lv, W.; Jiang, S.; Zhang, D.; Cai, G.; Pan, J.; Li, D. Genome-wide identification and expression analysis of calcium-dependent protein kinase in maize. *BMC Genom.* **2013**, *14*, 433. [CrossRef]
50. He, S. Genome-wide identification and expression analysis of calcium-dependent protein kinase and its closely related kinase genes in *Capsicum annuum*. *Front. Plant Sci.* **2015**, *6*, 737.
51. Wankhede, D.P.; Kumari, M.; Richa, T.; Aravind, J.; Rajkumar, S. Genome wide identification and characterization of Calcium Dependent Protein Kinase gene family in *Cajanus cajan*. *J. Environ. Biol.* **2017**, *38*, 167. [CrossRef]
52. Gromadka, R.; Cieśla, J.; Olszak, K.; Szczegielniak, J.; Muszyńska, G.; Polkowska-Kowalczyk, L. Genome-wide analysis and expression profiling of calcium-dependent protein kinases in potato (*Solanum tuberosum*). *Plant Growth Reg.* **2018**, *84*, 303–315. [CrossRef]
53. Asano, T.; Tanaka, N.; Yang, G.; Hayashi, N.; Komatsu, S. Genome-wide identification of the rice calcium-dependent protein kinase and its closely related kinase gene families: Comprehensive analysis of the CDPKs gene family in rice. *Plant Cell Physiol.* **2005**, *46*, 356–366. [CrossRef] [PubMed]
54. Wan, B.; Lin, Y.; Mou, T. Expression of rice Ca^{2+}-dependent protein kinases (CDPKs) genes under different environmental stresses. *FEBS Lett.* **2007**, *581*, 1179–1189. [CrossRef] [PubMed]
55. Xiao, X.H.; Yang, M.; Sui, J.L.; Qi, J.Y.; Fang, Y.J.; Hu, S.N.; Tang, C.R. The calcium-dependent protein kinase (CDPK) and CDPK-related kinase gene families in *Hevea brasiliensis*—Comparison with five other plant species in structure, evolution, and expression. *FEBS Open Biol.* **2017**, *7*, 4–24. [CrossRef] [PubMed]
56. Liu, F. Calcium-Dependent Protein Kinase Regulates Soybean Serine Acetyltransferase in Response to Oxidative Stress. Ph.D. Thesis, University of Florida, Gainesville, FL, USA, 2002.
57. Hettenhausen, C.; Sun, G.; He, Y.; Zhuang, H.; Sun, T.; Qi, J.; Wu, J. Genome-wide identification of calcium-dependent protein kinases in soybean and analyses of their transcriptional responses to insect herbivory and drought stress. *Sci. Rep.* **2016**, *6*, 18973. [CrossRef] [PubMed]
58. Liu, H.; Che, Z.; Zeng, X.; Zhou, X.; Sitoe, H.M.; Wang, H.; Yu, D. Genome-wide analysis of calcium-dependent protein kinases and their expression patterns in response to herbivore and wounding stresses in soybean. *Funct. Integ. Genom.* **2016**, *16*, 481–493. [CrossRef] [PubMed]
59. Tai, S.-S.; Liu, G.-S.; Sun, Y.-H.; Jia, C. Cloning and expression of calcium-dependent protein kinase (CDPK) gene family in common tobacco (*Nicotiana tabacum*). *Agric. Sci. China* **2009**, *8*, 1448–1457. [CrossRef]
60. Wang, J.-P.; Xu, Y.-P.; Munyampundu, J.-P.; Liu, T.-Y.; Cai, X.-Z. Calcium-dependent protein kinase (CDPK) and CDPK-related kinase (CRK) gene families in tomato: Genome-wide identification and functional analyses in disease resistance. *Mol. Genet. Genom.* **2016**, *291*, 661–676. [CrossRef]
61. Hu, Z.; Lv, X.; Xia, X.; Zhou, J.; Shi, K.; Yu, J.; Zhou, Y. Genome-wide identification and expression analysis of calcium-dependent protein kinase in tomato. *Front. Plant Sci.* **2016**, *7*, 469. [CrossRef]
62. Podell, S.; Gribskov, M. Predicting N-terminal myristoylation sites in plant proteins. *BMC Genom.* **2004**, *5*, 37. [CrossRef]
63. Liese, A.; Romeis, T. Biochemical regulation of in vivo function of plant calcium-dependent protein kinases (CDPK). *Biochim. Biophys. Acta BBA Mol. Cell Res.* **2013**, *1833*, 1582–1589. [CrossRef]
64. White, P.J.; Broadley, M.R. Calcium in plants. *Ann. Bot.* **2003**, *92*, 487–511. [CrossRef] [PubMed]
65. Himschoot, E.; Beeckman, T.; Friml, J.; Vanneste, S. Calcium is an organizer of cell polarity in plants. *Biochim. Biophys. Acta BBA Mol. Cell Res.* **2015**, *1853*, 2168–2172. [CrossRef] [PubMed]
66. Harmon, A.C.; Gribskov, M.; Gubrium, E.; Harper, J.F. The CDPK superfamily of protein kinases. *New Phytol.* **2001**, *151*, 175–183. [CrossRef]
67. Grabarek, Z. Structural basis for diversity of the EF-hand calcium-binding proteins. *J. Mol. Biol.* **2006**, *359*, 509–525. [CrossRef]
68. Klimecka, M.; Muszynska, G. Structure and functions of plant calcium-dependent protein kinases. *Acta Biochim. Pol. Eng. Edi.* **2007**, *54*, 219.
69. Wernimont, A.K.; Artz, J.D.; Finerty, P., Jr.; Lin, Y.-H.; Amani, M.; Allali-Hassani, A.; Senisterra, G.; Vedadi, M.; Tempel, W.; Mackenzie, F. Structures of apicomplexan calcium-dependent protein kinases reveal mechanism of activation by calcium. *Nat. Struct. Mol. Biol.* **2010**, *17*, 596. [CrossRef]
70. Parvathy, S.T. Versatile roles of ubiquitous calcium-dependent protein kinases (CDPKs) in plants. *Indian Soc. Oilseeds Res.* **2018**, *35*, 1–11.

71. Wei, S.; Hu, W.; Deng, X.; Zhang, Y.; Liu, X.; Zhao, X.; Luo, Q.; Jin, Z.; Li, Y.; Zhou, S. A rice calcium-dependent protein kinase OsCPK9 positively regulates drought stress tolerance and spikelet fertility. *BMC Plant Biol.* **2014**, *14*, 133. [CrossRef]
72. Zhang, H.; Zhang, Y.; Deng, C.; Deng, S.; Li, N.; Zhao, C.; Zhao, R.; Liang, S.; Chen, S. The *Arabidopsis* Ca^{2+}-Dependent Protein Kinase CPK12 Is Involved in Plant Response to Salt Stress. *Int. J. Mol. Sci.* **2018**, *19*, 4062. [CrossRef]
73. Dubrovina, A.S.; Kiselev, K.V.; Khristenko, V.S.; Aleynova, O.A. The calcium-dependent protein kinase gene VaCPK29 is involved in grapevine responses to heat and osmotic stresses. *Plant Growth Reg.* **2017**, *82*, 79–89. [CrossRef]
74. Liu, Y.; Xu, C.; Zhu, Y.; Zhang, L.; Chen, T.; Zhou, F.; Chen, H.; Lin, Y. The calcium-dependent kinase OsCPK24 functions in cold stress responses in rice. *J. Integr. Plant Biol.* **2018**, *60*, 173–188. [CrossRef] [PubMed]
75. Chen, D.-H.; Liu, H.-P.; Li, C.-L. Calcium-dependent protein kinase CPK9 negatively functions in stomatal abscisic acid signaling by regulating ion channel activity in *Arabidopsis*. *Plant Mol. Biol.* **2019**, *99*, 113–122. [CrossRef] [PubMed]
76. Coca, M.; San Segundo, B. AtCPK1 calcium-dependent protein kinase mediates pathogen resistance in *Arabidopsis*. *Plant J.* **2010**, *63*, 526–540. [CrossRef] [PubMed]
77. Fantino, E.; Segretin, M.E.; Santin, F.; Mirkin, F.G.; Ulloa, R.M. Analysis of the potato calcium-dependent protein kinase family and characterization of StCDPK7, a member induced upon infection with *Phytophthora infestans*. *Plant Cell Rep.* **2017**, *36*, 1137–1157. [CrossRef] [PubMed]
78. Dubiella, U.; Seybold, H.; Durian, G.; Komander, E.; Lassig, R.; Witte, C.-P.; Schulze, W.X.; Romeis, T. Calcium-dependent protein kinase/NADPH oxidase activation circuit is required for rapid defense signal propagation. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 8744–8749. [CrossRef]
79. Chung, E.; Park, J.M.; Oh, S.-K.; Joung, Y.H.; Lee, S.; Choi, D. Molecular and biochemical characterization of the *Capsicum annuum* calcium-dependent protein kinase 3 (CaCDPK3) gene induced by abiotic and biotic stresses. *Planta* **2004**, *220*, 286–295. [CrossRef]
80. Yang, D.-H.; Hettenhausen, C.; Baldwin, I.T.; Wu, J. Silencing *Nicotiana attenuata* calcium-dependent protein kinases, CDPK4 and CDPK5, strongly up-regulates wound-and herbivory-induced jasmonic acid accumulations. *Plant Physiol.* **2012**, *159*, 1591–1607. [CrossRef]
81. Matschi, S.; Werner, S.; Schulze, W.X.; Legen, J.; Hilger, H.H.; Romeis, T. Function of calcium-dependent protein kinase CPK 28 of *Arabidopsis thaliana* in plant stem elongation and vascular development. *Plant J.* **2013**, *73*, 883–896. [CrossRef]
82. Jin, Y.; Ye, N.; Zhu, F.; Li, H.; Wang, J.; Jiang, L.; Zhang, J. Calcium-dependent protein kinase CPK 28 targets the methionine adenosyltransferases for degradation by the 26S proteasome and affects ethylene biosynthesis and lignin deposition in *Arabidopsis*. *Plant J.* **2017**, *90*, 304–318. [CrossRef]
83. Zhou, L.; Lan, W.; Jiang, Y.; Fang, W.; Luan, S. A calcium-dependent protein kinase interacts with and activates a calcium channel to regulate pollen tube growth. *Mol. Plant* **2014**, *7*, 369–376. [CrossRef]
84. Zhu, J.-H.; Chen, X.; Chang, W.-J.; Tian, W.-M.; Zhang, Z.-L. Molecular characterization of HbCDPK1, an ethephon-induced calcium-dependent protein kinase gene of *Hevea brasiliensis*. *Biosci. Biotechnol. Biochem.* **2010**, *74*, 2183–2188. [CrossRef] [PubMed]
85. Kiselev, K.V.; Turlenko, A.V.; Zhuravlev, Y.N. Structure and expression profiling of a novel calcium-dependent protein kinase gene PgCDPK1a in roots, leaves, and cell cultures of *Panax ginseng*. *Plant Cell Tissue Organ Cult.* **2010**, *103*, 197–204. [CrossRef]
86. Tsai, T.-M.; Chen, Y.-R.; Kao, T.-W.; Tsay, W.-S.; Wu, C.-P.; Huang, D.-D.; Chen, W.-H.; Chang, C.-C.; Huang, H.-J. PaCDPK1, a gene encoding calcium-dependent protein kinase from orchid, *Phalaenopsis amabilis*, is induced by cold, wounding, and pathogen challenge. *Plant Cell Rep.* **2007**, *26*, 1899–1908. [CrossRef] [PubMed]
87. Jiang, S.; Zhang, D.; Wang, L.; Pan, J.; Liu, Y.; Kong, X.; Zhou, Y.; Li, D. A maize calcium-dependent protein kinase gene, ZmCPK4, positively regulated abscisic acid signaling and enhanced drought stress tolerance in transgenic *Arabidopsis*. *Plant Physiol. Biochem.* **2013**, *71*, 112–120. [CrossRef] [PubMed]
88. Kiselev, K.; Dubrovina, A.; Shumakova, O.; Karetin, Y.; Manyakhin, A. Structure and expression profiling of a novel calcium-dependent protein kinase gene, CDPK3a, in leaves, stems, grapes, and cell cultures of wild-growing grapevine *Vitis amurensis* Rupr. *Plant Cell Rep.* **2013**, *32*, 431–442. [CrossRef] [PubMed]

89. Aleynova, O.; Dubrovina, A.; Kiselev, K. Activation of stilbene synthesis in cell cultures of *Vitis amurensis* by calcium-dependent protein kinases VaCPK1 and VaCPK26. *Plant Cell Tissue Organ Cult.* **2017**, *130*, 141–152. [CrossRef]
90. Pawełek, A.; Szmidt-Jaworska, A.; Świeżawska, B.; Jaworski, K. Genomic structure and promoter characterization of the CDPK kinase gene expressed during seed formation in *Pharbitis nil*. *J. Plant Physiol.* **2015**, *189*, 87–96. [CrossRef]
91. Martínez-Noël, G.; Nagaraj, V.J.; Caló, G.; Wiemken, A.; Pontis, H.G. Sucrose regulated expression of a Ca^{2+}-dependent protein kinase (TaCDPK1) gene in excised leaves of wheat. *Plant Physiol. Biochem.* **2007**, *45*, 410–419. [CrossRef]
92. Schulz, P.; Herde, M.; Romeis, T. Calcium-dependent protein kinases: Hubs in plant stress signaling and development. *Plant Physiol.* **2013**, *163*, 523–530. [CrossRef]
93. Singh, A.; Sagar, S.; Biswas, D.K. Calcium dependent protein kinase, a versatile player in plant stress management and development. *Critic. Rev. Plant Sci.* **2017**, *36*, 336–352. [CrossRef]
94. Monroy, A.F.; Dhindsa, R.S. Low-temperature signal transduction: Induction of cold acclimation-specific genes of alfalfa by calcium at 25 °C. *Plant Cell* **1995**, *7*, 321–331. [PubMed]
95. Fallon, K.M.; Shacklock, P.S.; Trewavas, A.J. Detection in vivo of very rapid red light-induced calcium-sensitive protein phosphorylation in etiolated wheat (*Triticum aestivum*) leaf protoplasts. *Plant Physiol.* **1993**, *101*, 1039–1045. [CrossRef] [PubMed]
96. Kiegle, E.; Moore, C.A.; Haseloff, J.; Tester, M.A.; Knight, M.R. Cell-type-specific calcium responses to drought, salt and cold in the *Arabidopsis* root. *Plant J.* **2000**, *23*, 267–278. [CrossRef] [PubMed]
97. Zou, J.-J.; Li, X.-D.; Ratnasekera, D.; Wang, C.; Liu, W.-X.; Song, L.-F.; Zhang, W.-Z.; Wu, W.-H. *Arabidopsis* Calcium-Dependent Protein Kinase8 and CATALASE3 function in abscisic acid-mediated signaling and H_2O_2 homeostasis in stomatal guard cells under drought stress. *Plant Cell* **2015**, *27*, 1445–1460. [CrossRef]
98. Zhu, S.-Y.; Yu, X.-C.; Wang, X.-J.; Zhao, R.; Li, Y.; Fan, R.-C.; Shang, Y.; Du, S.-Y.; Wang, X.-F.; Wu, F.-Q. Two calcium-dependent protein kinases, CPK4 and CPK11, regulate abscisic acid signal transduction in *Arabidopsis*. *Plant Cell* **2007**, *19*, 3019–3036. [CrossRef]
99. Takahashi, K.; Isobe, M.; Muto, S. An increase in cytosolic calcium ion concentration precedes hypoosmotic shock-induced activation of protein kinases in tobacco suspension culture cells. *FEBS Lett.* **1997**, *401*, 202–206. [CrossRef]
100. Kobayashi, M.; Yoshioka, M.; Asai, S.; Nomura, H.; Kuchimura, K.; Mori, H.; Doke, N.; Yoshioka, H. StCDPK5 confers resistance to late blight pathogen but increases susceptibility to early blight pathogen in potato via reactive oxygen species burst. *New Phytol.* **2012**, *196*, 223–237. [CrossRef]
101. Perochon, A.; Aldon, D.; Galaud, J.-P.; Ranty, B. Calmodulin and calmodulin-like proteins in plant calcium signaling. *Biochimie* **2011**, *93*, 2048–2053. [CrossRef]
102. Szczegielniak, J.; Borkiewicz, L.; Szurmak, B.; Lewandowska-Gnatowska, E.; Statkiewicz, M.; Klimecka, M.; Cieśla, J.; Muszyńska, G. Maize calcium-dependent protein kinase (ZmCPK11): Local and systemic response to wounding, regulation by touch and components of jasmonate signaling. *Physiol. Plant.* **2012**, *146*, 1–14. [CrossRef]
103. Ding, Y.; Cao, J.; Ni, L.; Zhu, Y.; Zhang, A.; Tan, M.; Jiang, M. ZmCPK11 is involved in abscisic acid-induced antioxidant defence and functions upstream of ZmMPK5 in abscisic acid signalling in maize. *J. Exp. Bot.* **2012**, *64*, 871–884. [CrossRef]
104. Zandalinas, S.I.; Mittler, R.; Balfagón, D.; Arbona, V.; Gómez-Cadenas, A. Plant adaptations to the combination of drought and high temperatures. *Physiol. Plant.* **2018**, *162*, 2–12. [CrossRef] [PubMed]
105. Foyer, C.H.; Noctor, G. Redox homeostasis and antioxidant signaling: A metabolic interface between stress perception and physiological responses. *Plant Cell* **2005**, *17*, 1866–1875. [CrossRef] [PubMed]
106. Zou, J.-J.; Wei, F.-J.; Wang, C.; Wu, J.-J.; Ratnasekera, D.; Liu, W.-X.; Wu, W.-H. *Arabidopsis* calcium-dependent protein kinase CPK10 functions in abscisic acid-and Ca^{2+}-mediated stomatal regulation in response to drought stress. *Plant Physiol.* **2010**, *154*, 1232–1243. [CrossRef] [PubMed]
107. Dubrovina, A.S.; Kiselev, K.V.; Khristenko, V.S.; Aleynova, O.A. VaCPK20, a calcium-dependent protein kinase gene of wild grapevine *Vitis amurensis* Rupr., mediates cold and drought stress tolerance. *J. Plant Physiol.* **2015**, *185*, 1–12. [CrossRef]
108. Liu, G.; Chen, J.; Wang, X. VfCPK1, a gene encoding calcium-dependent protein kinase from *Vicia faba*, is induced by drought and abscisic acid. *Plant Cell Environ.* **2006**, *29*, 2091–2099. [CrossRef]

109. Ratnasekera, D. A calcium dependent protein kinase involves H_2O_2 mediated guard cell signaling in Aarabidopsis. *Trop. Agri. Res. Exten.* **2015**, *16*, 7–14. [CrossRef]
110. Syam Prakash, S.R.; Jayabaskaran, C. Heterologous expression and biochemical characterization of two calcium-dependent protein kinase isoforms CaCPK1 and CaCPK2 from chickpea. *J. Plant Physiol.* **2006**, *163*, 1083–1093. [CrossRef]
111. Ho, S.-L.; Huang, L.-F.; Lu, C.-A.; He, S.-L.; Wang, C.-C.; Yu, S.-P.; Chen, J.; Yu, S.-M. Sugar starvation-and GA-inducible calcium-dependent protein kinase 1 feedback regulates GA biosynthesis and activates a 14-3-3 protein to confer drought tolerance in rice seedlings. *Plant Mol. Biol.* **2013**, *81*, 347–361. [CrossRef]
112. Geng, S.; Zhao, Y.; Tang, L.; Zhang, R.; Sun, M.; Guo, H.; Kong, X.; Li, A.; Mao, L. Molecular evolution of two duplicated CDPK genes CPK7 and CPK12 in grass species: A case study in wheat (*Triticum aestivum* L.). *Gene* **2011**, *475*, 94–103. [CrossRef]
113. Vivek, P.J.; Tuteja, N.; Soniya, E.V. CDPK1 from ginger promotes salinity and drought stress tolerance without yield penalty by improving growth and photosynthesis in *Nicotiana tabacum*. *PLoS ONE* **2013**, *8*, e76392. [CrossRef]
114. Huang, K.; Peng, L.; Liu, Y.; Yao, R.; Liu, Z.; Li, X.; Yang, Y.; Wang, J. *Arabidopsis* calcium-dependent protein kinase AtCPK1 plays a positive role in salt/drought-stress response. *Biochem. Biophys. Res. Commun.* **2018**, *498*, 92–98. [CrossRef] [PubMed]
115. Brandt, B. Specific Calcium and Abscisic Acid Regulation of Anion Channels in *Arabidopsis* Guard Cells. Ph.D. Thesis, Eberhard Karls University Tübingen, Tübingen, Germany, 2014.
116. Baba, A.; Rigó, G.; Ayaydin, F.; Rehman, A.; Andrási, N.; Zsigmond, L.; Valkai, I.; Urbancsok, J.; Vass, I.; Pasternak, T. Functional Analysis of the *Arabidopsis* thaliana CDPK-Related Kinase Family: AtCRK1 Regulates Responses to Continuous Light. *Int. J. Mol. Sci.* **2018**, *19*, 1282. [CrossRef] [PubMed]
117. Mori, I.C.; Murata, Y.; Yang, Y.; Munemasa, S.; Wang, Y.-F.; Andreoli, S.; Tiriac, H.; Alonso, J.M.; Harper, J.F.; Ecker, J.R. CDPKs CPK6 and CPK3 function in ABA regulation of guard cell S-type anion-and Ca^{2+}-permeable channels and stomatal closure. *PLoS Biol.* **2006**, *4*, e327. [CrossRef] [PubMed]
118. Mehlmer, N.; Wurzinger, B.; Stael, S.; Hofmann-Rodrigues, D.; Csaszar, E.; Pfister, B.; Bayer, R.; Teige, M. The Ca^{2+}-dependent protein kinase CPK3 is required for MAPK-independent salt-stress acclimation in *Arabidopsis*. *Plant J.* **2010**, *63*, 484–498. [CrossRef] [PubMed]
119. Xu, J.; Tian, Y.-S.; Peng, R.-H.; Xiong, A.-S.; Zhu, B.; Jin, X.-F.; Gao, F.; Fu, X.-Y.; Hou, X.-L.; Yao, Q.-H. AtCPK6, a functionally redundant and positive regulator involved in salt/drought stress tolerance in *Arabidopsis*. *Planta* **2010**, *231*, 1251–1260. [CrossRef] [PubMed]
120. Milla, M.A.R.; Townsend, J.; Chang, F.; Cushman, J.C. The *Arabidopsis* AtDi19 gene family encodes a novel type of Cys2/His2 zinc-finger protein implicated in ABA-independent dehydration, high-salinity stress and light signaling pathways. *Plant Mol. Biol.* **2006**, *61*, 13–30. [CrossRef]
121. Zhao, R.; Sun, H.L.; Mei, C.; Wang, X.J.; Yan, L.; Liu, R.; Zhang, X.F.; Wang, X.F.; Zhang, D.P. The *Arabidopsis* Ca^{2+}-dependent protein kinase CPK12 negatively regulates abscisic acid signaling in seed germination and post-germination growth. *New Phytol.* **2011**, *192*, 61–73. [CrossRef]
122. Huang, S.-J.; Chang, C.-L.; Wang, P.-H.; Tsai, M.-C.; Hsu, P.-H.; Chang, I.-F. A type III ACC synthase, ACS7, is involved in root gravitropism in *Arabidopsis thaliana*. *J. Exp. Bot.* **2013**, *64*, 4343–4360. [CrossRef]
123. Franz, S.; Ehlert, B.; Liese, A.; Kurth, J.; Cazalé, A.-C.; Romeis, T. Calcium-dependent protein kinase CPK21 functions in abiotic stress response in *Arabidopsis thaliana*. *Mol. Plant* **2011**, *4*, 83–96. [CrossRef]
124. Ma, S.-Y.; Wu, W.-H. AtCPK23 functions in *Arabidopsis* responses to drought and salt stresses. *Plant Mol. Biol.* **2007**, *65*, 511–518. [CrossRef]
125. Zhao, R.; Sun, H.; Zhao, N.; Jing, X.; Shen, X.; Chen, S. The *Arabidopsis* Ca^{2+}-dependent protein kinase CPK27 is required for plant response to salt-stress. *Gene* **2015**, *563*, 203–214. [CrossRef]
126. Choi, H.-I.; Park, H.-J.; Park, J.H.; Kim, S.; Im, M.-Y.; Seo, H.-H.; Kim, Y.-W.; Hwang, I.; Kim, S.Y. *Arabidopsis* calcium-dependent protein kinase AtCPK32 interacts with ABF4, a transcriptional regulator of abscisic acid-responsive gene expression, and modulates its activity. *Plant Physiol.* **2005**, *139*, 1750–1761. [CrossRef] [PubMed]
127. Kawamoto, N.; Sasabe, M.; Endo, M.; Machida, Y.; Araki, T. Calcium-dependent protein kinases responsible for the phosphorylation of a bZIP transcription factor FD crucial for the florigen complex formation. *Sci. Rep.* **2015**, *5*, 8341. [CrossRef] [PubMed]

128. Llop-Tous, I.; Domínguez-Puigjaner, E.; Vendrell, M. Characterization of a strawberry cDNA clone homologous to calcium-dependent protein kinases that is expressed during fruit ripening and affected by low temperature. *J. Exp. Bot.* **2002**, *53*, 2283–2285. [CrossRef] [PubMed]
129. Davletova, S.; Mészáros, T.; Miskolczi, P.; Oberschall, A.; Török, K.; Magyar, Z.; Dudits, D.; Deák, M. Auxin and heat shock activation of a novel member of the calmodulin like domain protein kinase gene family in cultured alfalfa cells. *J. Exp. Bot.* **2001**, *52*, 215–221. [CrossRef] [PubMed]
130. Wang, J.; Wang, S.; Hu, K.; Yang, J.; Xin, X.; Zhou, W.; Fan, J.; Cui, F.; Mou, B.; Zhang, S. The kinase OsCPK4 regulates a buffering mechanism that fine-tunes innate immunity. *Plant Physiol.* **2018**, *176*, 1835–1849. [CrossRef]
131. Kang, C.H.; Moon, B.C.; Park, H.C.; Koo, S.C.; Chi, Y.H.; Cheong, Y.H.; Yoon, B.-D.; Lee, S.Y.; Kim, C.Y. Rice small C2-domain proteins are phosphorylated by calcium-dependent protein kinase. *Mol. Cells* **2013**, *35*, 381–387. [CrossRef]
132. Fu, L.; Yu, X.; An, C. Overexpression of constitutively active OsCPK10 increases *Arabidopsis* resistance against *Pseudomonas syringae* pv. tomato and rice resistance against *Magnaporthe grisea*. *Plant Physiol. Biochem.* **2013**, *73*, 202–210. [CrossRef]
133. Wang, B.; Zhang, Y.; Bi, Z.; Liu, Q.; Xu, T.; Yu, N.; Cao, Y.; Zhu, A.; Wu, W.; Zhan, X. Impaired function of the calcium-dependent protein kinase, OsCPK12, leads to early senescence in Rice (*Oryza sativa* L.). *Front. Plant Sci.* **2019**, *10*, 52. [CrossRef]
134. Abbasi, F.; Onodera, H.; Toki, S.; Tanaka, H.; Komatsu, S. OsCDPK13, a calcium-dependent protein kinase gene from rice, is induced by cold and gibberellin in rice leaf sheath. *Plant Mol. Biol.* **2004**, *55*, 541–552. [CrossRef]
135. Almadanim, M.C.; Alexandre, B.M.; Rosa, M.T.; Sapeta, H.; Leitão, A.E.; Ramalho, J.C.; Lam, T.T.; Negrão, S.; Abreu, I.A.; Oliveira, M.M. Rice calcium-dependent protein kinase OsCPK17 targets plasma membrane intrinsic protein and sucrose-phosphate synthase and is required for a proper cold stress response. *Plant Cell Environ.* **2017**, *40*, 1197–1213. [CrossRef]
136. Asano, T.; Hakata, M.; Nakamura, H.; Aoki, N.; Komatsu, S.; Ichikawa, H.; Hirochika, H.; Ohsugi, R. Functional characterisation of OsCPK21, a calcium-dependent protein kinase that confers salt tolerance in rice. *Plant Mol. Biol.* **2011**, *75*, 179–191. [CrossRef]
137. Manimaran, P.; Mangrauthia, S.K.; Sundaram, R.; Balachandran, S. Constitutive expression and silencing of a novel seed specific calcium dependent protein kinase gene in rice reveals its role in grain filling. *J. Plant Physiol.* **2015**, *174*, 41–48. [CrossRef]
138. Ishida, S.; Yuasa, T.; Nakata, M.; Takahashi, Y. A tobacco calcium-dependent protein kinase, CDPK1, regulates the transcription factor repression of shoot growth in response to gibberellins. *Plant Cell* **2008**, *20*, 3273–3288. [CrossRef]
139. Romeis, T.; Ludwig, A.A.; Martin, R.; Jones, J.D. Calcium-dependent protein kinases play an essential role in a plant defence response. *EMBO J.* **2001**, *20*, 5556–5567. [CrossRef]
140. Botella, J.R.; Arteca, J.M.; Somodevilla, M.; Arteca, R.N. Calcium-dependent protein kinase gene expression in response to physical and chemical stimuli in mungbean (*Vigna radiata*). *Plant Mol. Biol.* **1996**, *30*, 1129–1137. [CrossRef]
141. Chang, W.-J.; Su, H.-S.; Li, W.-J.; Zhang, Z.-L. Expression profiling of a novel calcium-dependent protein kinase gene, LeCPK2, from tomato (*Solanum lycopersicum*) under heat and pathogen-related hormones. *Biosci. Biotechnol. Biochem.* **2009**, *73*, 24–27. [CrossRef]
142. Santin, F.; Bhogale, S.; Fantino, E.; Grandellis, C.; Banerjee, A.K.; Ulloa, R.M. Solanum tuberosum StCDPK1 is regulated by miR390 at the posttranscriptional level and phosphorylates the auxin efflux carrier StPIN4 in vitro, a potential downstream target in potato development. *Physiol. Plant.* **2017**, *159*, 244–261. [CrossRef]
143. Kobayashi, M.; Ohura, I.; Kawakita, K.; Yokota, N.; Fujiwara, M.; Shimamoto, K.; Doke, N.; Yoshioka, H. Calcium-dependent protein kinases regulate the production of reactive oxygen species by potato NADPH oxidase. *Plant Cell* **2007**, *19*, 1065–1080. [CrossRef]
144. Wang, M.; Li, Q.; Sun, K.; Chen, X.; Zhou, Q.; Li, H.; Zhang, X.; Li, X. Involvement of CsCDPK20 and CsCDPK26 in Regulation of Thermotolerance in Tea Plant (*Camellia sinensis*). *Plant Mol. Biol. Rep.* **2018**, *36*, 176–187. [CrossRef]

145. Freymark, G.; Diehl, T.; Miklis, M.; Romeis, T.; Panstruga, R. Antagonistic control of powdery mildew host cell entry by barley calcium-dependent protein kinases (CDPKs). *Mol. Plant Microbe Interact.* **2007**, *20*, 1213–1221. [CrossRef] [PubMed]
146. Wang, H.; Gong, J.; Su, X.; Li, L.; Pang, X.; Zhang, Z. MaCDPK7, a calcium-dependent protein kinase gene from banana is involved in fruit ripening and temperature stress responses. *J. Hort. Sci. Biotechnol.* **2017**, *92*, 240–250. [CrossRef]
147. Wang, Z.; Li, J.; Jia, C.; Xu, B.; Jin, Z. Molecular cloning and expression analysis of eight calcium-dependent protein kinase (CDPK) genes from banana (*Musa acuminata* L. AAA group, cv. Cavendish). *S. Afr. J. Bot.* **2016**, *104*, 134–141. [CrossRef]
148. Dubrovina, A.; Kiselev, K. The Role of Calcium-Dependent Protein Kinase Genes VaCPK1 and VaCPK26 in the Response of *Vitis amurensis* (in vitro) and *Arabidopsis thaliana* (in vivo) to Abiotic Stresses. *Russian J. Genet.* **2019**, *55*, 319–329. [CrossRef]
149. Dubrovina, A.S.; Kiselev, K.V.; Khristenko, V.S.; Aleynova, O.A. VaCPK21, a calcium-dependent protein kinase gene of wild grapevine *Vitis amurensis* Rupr., is involved in grape response to salt stress. *Plant Cell. Tissue Organ Cult.* **2016**, *124*, 137–150. [CrossRef]
150. Chen, J.; Xue, B.; Xia, X.; Yin, W. A novel calcium-dependent protein kinase gene from *Populus euphratica*, confers both drought and cold stress tolerance. *Biochem. Biophys. Res. Commun.* **2013**, *441*, 630–636. [CrossRef]
151. Ning, M.; Tang, F.; Zhang, Q.; Zhao, X.; Yang, L.; Cai, W.; Shan, C. Effects of Penicillium infection on the expression and activity of CDPK2 in postharvest Hami melon treated with calcium chloride. *Physiol. Mol. Plant Pathol.* **2019**, *106*, 175–181. [CrossRef]
152. AbdElgawad, H.; Zinta, G.; Hegab, M.M.; Pandey, R.; Asard, H.; Abuelsoud, W. High salinity induces different oxidative stress and antioxidant responses in maize seedlings organs. *Front. Plant Sci.* **2016**, *7*, 276. [CrossRef]
153. Dubrovina, A.; Aleynova, O.; Manyakhin, A.; Kiselev, K. The Role of Calcium-Dependent Protein Kinase Genes *CPK16*, *CPK25*, *CPK30*, and *CPK32* in Stilbene Biosynthesis and the Stress Resistance of Grapevine *Vitis amurensis* Rupr. *App. Biochem. Microbiol.* **2018**, *54*, 410–417. [CrossRef]
154. Miller, G.; Suzuki, N.; Ciftci-Yilmaz, S.; Mittler, R. Reactive oxygen species homeostasis and signalling during drought and salinity stresses. *Plant Cell Environ.* **2010**, *33*, 453–467. [CrossRef]
155. Kwak, J.M.; Mori, I.C.; Pei, Z.M.; Leonhardt, N.; Torres, M.A.; Dangl, J.L.; Bloom, R.E.; Bodde, S.; Jones, J.D.; Schroeder, J.I. NADPH oxidase AtrbohD and AtrbohF genes function in ROS-dependent ABA signaling in *Arabidopsis*. *EMBO J.* **2003**, *22*, 2623–2633. [CrossRef] [PubMed]
156. Mittler, R. ROS are good. *Trends Plant Sci.* **2017**, *22*, 11–19. [CrossRef] [PubMed]
157. Tao, X.-C.; Lu, Y.-T. Loss of AtCRK1 gene function in *Arabidopsis thaliana* decreases tolerance to salt. *J. Plant Biol.* **2013**, *56*, 306–314. [CrossRef]

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

International Journal of Molecular Sciences Editorial Office
E-mail: ijms@mdpi.com
www.mdpi.com/journal/ijms

MDPI

Academic Open
Access Publishing

www.mdpi.com

ISBN 978-3-0365-8089-0