Article

Characterisation of Faba Bean (*Vicia faba* L.) Transcriptome Using RNA-Seq: Sequencing, *De Novo* Assembly, Annotation, and Expression Analysis

Shivraj Braich [1], Shimna Sudheesh [1], John W. Forster [1,2] and Sukhjiwan Kaur [1,*]

[1]   Agriculture Victoria, Biosciences Research, AgriBio, the Centre for AgriBioscience, 5 Ring Road, Bundoora, Victoria 3083, Australia; shivi.braich@ecodev.vic.gov.au (S.B.); shimna.sudheesh@ecodev.vic.gov.au (S.S.); john.forster@ecodev.vic.gov.au (J.W.F.)

[2]   School of Applied Systems Biology, La Trobe University, Bundoora, Victoria 3086, Australia

[*]   Correspondence: sukhjiwan.kaur@ecodev.vic.gov.au; Tel.: +61-03-90327117

Abstract: RNA sequencing (RNA-Seq) is a deep sequencing method used for transcriptome profiling. RNA-Seq assemblies have successfully been used for a broad variety of applications, such as gene characterisation, functional genomic studies, and gene expression analysis, particularly useful in the absence of a well-studied genome reference sequence. This study reports on the development of reference unigene sets from faba bean using RNA-Seq. Two Australian faba bean cultivars (Doza and Farah) that differ in terms of disease resistance, breeding habit, and adaptation characteristics, and have been extensively used in breeding programs, were utilised in this study. The *de novo* assembly resulted in a total of 58,962 and 53,275 transcripts with approximately 67 Mbp (1588 bp N50) and 61 Mbp (1629 bp N50) for Doza and Farah, respectively. The generated transcripts have been compared to the protein and nucleotide databases of NCBI, as well as to the gene complements of several related legume species such as *Medicago truncatula*, soybean, and chickpea. Both assemblies were compared to previously-published faba bean transcriptome reference sets for the degree of completeness and utility. Annotation of unigenes has been performed, and patterns of tissue-specific expression identified. The gene complement derived from this comprehensive transcriptome analysis shows that faba bean, despite its complex 13 Gbp genome, compares well to other legumes in expressed gene content. This study in faba bean represents the most comprehensive reference transcriptomes from two different Australian cultivars available to date and it provides a valuable resource for future genomics-assisted breeding activities in this species.

Keywords: *V. faba* cultivars; RNA-Seq; Illumina; *de novo* assembly; BLAST; sequence annotation; tissue-specific gene expression

1. Introduction

Faba bean (*Vicia faba* L.) is a cool-season legume species, producing protein-rich grain not only for human production (particularly in Western Asia and Northern Africa), but also for livestock feed in developed regions, such as Europe and Australia [1,2]. Global cultivation on 2.4 Mha in 2012 produced circa 4 Mt [3], the primary production countries being China, Ethiopia, Morocco, and Australia.

Productivity of faba bean is limited by a number of biotic stresses, including diseases caused by viral, bacterial, and fungal pathogens, and invertebrate pests, such as nematodes. The major fungal diseases are chocolate spot (caused by *Botrytis fabae* Stard.), rust (caused by *Uromyces viciae-fabae* [Pers.] J. Schrött), ascochyta blight (caused by *A. fabae* Sperg.), and downy mildew (caused by *Peronospora viciae* [Berk.] Caspary) [1,4,5]. Constraints due to environmental stresses, such as drought [6], salinity [7], and

cold and frost [8] are also significant. Quality characteristics, such as protein content [9] and absence of anti-nutritional factors, like tannins [10], are also important targets for breeding improvement.

The genus *Vicia* belongs to the Viceae tribe within the Galegoid (cool-season) clade of the sub-family Papilionoideae, which is, in turn, part of the legume family, Fabaceae [11]. *Vicia* contains more than 160 species, with considerable variation of haploid genome size [12,13], corresponding to ca. 7.5-fold range from ca. 1862–14,112 Mbp. The genome size of faba bean is close to the upper boundary, at ca. 13,000 Mb [14]. Differential genome expansions within the *Vicia* genus are apparently largely attributable to amplifications of large retroelement sequences [15], although the basic genetic complement is likely to be conserved.

Faba bean is a facultative outbreeding species, levels of natural cross-pollination varying from 10% to 70% [16]. The fundamental chromosome number is 6, providing a diploid constitution of $2n = 2x = 12$. Development of genetic linkage maps for faba bean has been relatively slow compared to other pulse species, and was initially dominated by construction of low-density genetic maps populated by first-generation molecular marker systems, such as isoenzymes, restriction fragment length polymorphisms (RFLPs), and randomly-amplified polymorphic DNAs (RAPDs) [17,18]. Later studies based on sequence-characterised marker systems, such as intron-targeted amplified polymorphism (ITAPs), simple sequence repeats (SSRs), and sequence-characterised amplification regions (SCARs), permitted the development of more substantial maps, including the basis for comparative analysis with the genomes of other legume species [14,19,20]. Development of large-scale sequence resources, such as expressed sequence tags (ESTs) [21–23] or genome survey sequences (GSSs) [24] allowed further expansion of marker development, particularly for SSRs and single-nucleotide polymorphisms (SNPs), with associated improvement of genetic map resolution [25–29]. Various population-specific genetic maps have been used for detection of quantitative trait loci (QTLs) for agronomic performance, environmental stress tolerance, and disease resistance characters [28,30–35].

Although current resources are sufficient to support simple strategies for genomics-assisted breeding in faba bean, such as marker-assisted back-crossing [1,36], a transition to genomic selection [37], which depends on high-density genome-wide sequence polymorphism information will require a significant expansion of existing sequence collections. Ideally, a reference whole-genome sequence would be used in conjunction with resequencing of selected individuals from training populations. However, for large higher plant genomes, such as that of faba bean, whole genome sequence assembly remains a technically challenging proposition because of the abundance of repetitive DNA sequences. In contrast, sequencing of the transcriptome, which corresponds to the expressed proportion of the genome, provides an attractive option, especially through use of RNA-Seq technology on second-generation DNA sequencing platforms [38].

A number of transcriptome sampling studies have previously been performed for faba bean, with an emphasis on specific developmental stages or environmental conditions, or by using one type of source tissue [29,39–43]. However, there is an incentive to generate a more comprehensive resource through sampling of RNA populations from multiple tissue sources, allowing the construction of a transcriptome atlas, as previously described for grain legume species, such as soybean and field pea [44,45]. Apart from large-scale development of gene-associated molecular genetic markers for the purposes of genomic selection, transcriptome atlases can support gene isolation, identification of differentially-regulated gene sets, and the measurement of gene expression, as well as studies of comparative genomics, and (ultimately) annotation of whole genome sequences [38,46].The current study reports on comprehensive transcriptome assemblies using RNA-Seq from two Australian faba bean cultivars (Doza and Farah) that differ in their breeding habit, adaptation characteristics, and disease resistance. Doza is resistant to rust infection and tolerant to frost events, and is best adapted to regions of New South Wales and Southern Queensland that experience warmer spring temperatures [47]. Since Doza is susceptible to ascochyta blight and, therefore, has a significantly lower yield, cultivar Farah is favoured in the region of South Australia, to which it is well-adapted [48]. Farah is an older cultivar (registered in 2003) [48] compared to Doza (registered in 2008) [49]. Comparisons

of the respective transcriptome assemblies have been made to the gene complements of several related species. The annotation of unigenes has been performed, and patterns of tissue-specific expression have been characterised. The faba bean transcriptome dataset generated in this study will provide an important resource for future genomics-assisted breeding activities in this species.

2. Results

2.1. RNA-Seq and De Novo Transcriptome Assembly

A total of seven RNA-Seq libraries based on various plant tissues were generated from both of the faba bean cultivars and sequenced aiming at similar depths. However, sequencing runs from Farah generated ca. 25% of more data as compared to Doza. The raw sequence data reads were trimmed to remove the adaptor sequences and subsequently filtered to exclude short read lengths and low quality reads. This resulted in a total of 776,387,394 and 1,037,821,214 paired-end reads from Doza and Farah, respectively. The details of the sequence reads from different source tissues are summarised in Table 1.

Table 1. Summary details of the reads (paired-end) used for *de novo* transcriptome assembly.

Tissue	Doza	Farah
Flower	139,572,392	180,879,976
Immature pod	90,270,053	192,279,013
Pod	42,557,479	159,411,032
Immature seed	65,603,441	163,953,461
Leaf	152,851,023	105,042,750
Stem	136,193,439	129,564,811
Root	149,339,568	106,690,173
Total	776,387,394	1,037,821,214

The raw sequence reads were optimised by comparing different word sizes (k-mer), and a k-mer size of 101 was found to be optimal, based on the outcomes prior to the compilation of the assembly. The statistics of the sequencing data filtering and outputs are summarised in Table 2. The clean reads were assembled using SOAPdenovo-TRANS and were further compiled by CAP3 assembler. This resulted in 60,012 and 59,391 transcripts with N50 (length of the longest contig, such that all contigs of at least that length compose at least 50% of the bases of the assembly) scaffold of 1588 and 1629 bp for Doza and Farah, respectively.

Table 2. Statistics of sequencing outputs and assembly.

Primary Assembly	Statistics-Doza	Statistics-Farah
SOAPdenovo-Trans		
Total number of contigs and scaffolds	79,782	73,989
Total assembled (without N *)	88.7 Mbp	78 Mbp
N50 scaffold	1751 bp	1726 bp
CAP3		
Total number of contigs and scaffolds	60,012	59,391
Total assembled	67.4 Mbp	65.5 Mbp
N50 scaffold	1588 bp	1629 bp

* without the bases that are N.

The distribution of the assembled contigs and scaffolds was determined (Figure 1). The majority of the transcripts were in the range of 301–400 bp (20.4% for Doza and 23.8% for Farah), followed by those that were above 2000 bp in length (13.9% on average). The longest transcript for

Doza (Vf-D-scaffold38544) and Farah (Vf-F-scaffold41695) had the length of 23,428 bp and 20,384 bp, respectively.
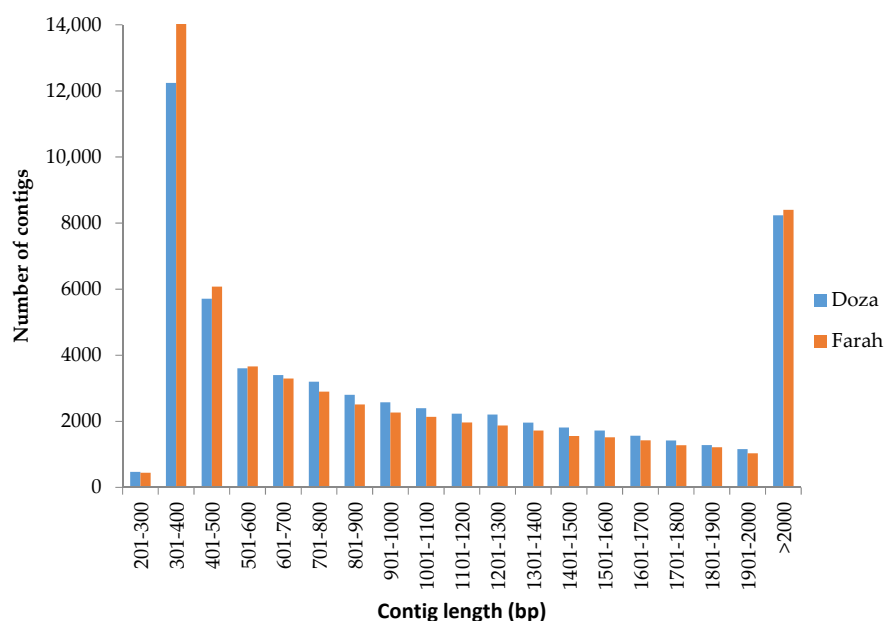

Figure 1. Distribution of the assembled transcript length from the Doza-specific and Farah-specific assemblies.

2.2. Classification and Functional Annotation of the Faba Bean Transcriptome

For the Doza-derived assembly, the BLASTN search to Nt database and the BLASTX search to Nr and UniRef100 databases identified 43,065, 44,501, and 44,523 transcripts, respectively (Table S1). For the Farah-derived assembly, 33,768 transcripts exhibited matches to the Nt database, 38,177 transcripts to Nr database as maintained by NCBI and 38,245 transcripts to UniRef 100 database (Table S2). A total of 47,477 Doza-specific transcripts had matches to either one or more Nt, Nr, and UniRef 100 databases with 39,654 transcripts having matches to all the three databases. For Farah, 36,922 transcripts had matches to either one or more Nt, Nr, and UniRef 100 databases with 28,065 Farah-specific transcripts having matches to each of the three databases.

A total of 376 transcripts from the Doza assembly and 5851 transcripts from the Farah assembly exhibited high-value matches of moderate similarity to non-plant-derived sources. A small proportion of these anomalies was resolved by comparison with BLASTN results to the closely related legume species. Totals of 270 (from Doza) and 5753 (from Farah) non-plant-derived sequences were removed from further analysis. Both the assemblies were also assessed for the presence of repeat elements and resulted in the identification of only circa 0.5% transcripts with annotations of repeat elements including retrotransposons components such as long terminal repeats (LTRs).

The BLASTN and BLASTX analysis to the Nt and Nr databases of NCBI revealed that the faba bean sequence annotations showed highest level matches to sequences from chickpea and *M. truncatula*. The *E*-value distribution of the significant matches (with *E*-value $<10^{-50}$) from the BLAST results exhibited high levels of similarity (89.8% for Doza and 83.3% for Farah) in the Nt database and approximately 71.8% (for Doza) and 69.8% (for Farah) in both the Nr and UniRef100 databases.

Transcripts from the reference Doza and Farah transcriptome assemblies were BLASTN analysed against the CDSs and genome of *M. truncatula*, CDSs of soybean and the genome of chickpea (Table S3 for Doza and Table S4 for Farah). The distribution of *E*-values based on BLASTN results of the comparator reference legume species to the generated faba bean assemblies is summarised in Table 3. In total, 43,284 (72.5%) Doza-specific transcripts had matches to either one or all the three comparator

legume species, with 23,299 transcripts displaying hits to all. In the Farah assembly, 33,075 (61.7%) transcripts displayed matches to either one or all of the *M. truncatula*, chickpea, and soybean datasets, with 16,727 transcripts exhibiting matches to all.

Table 3. *E*-value distribution of the faba bean transcripts BLAST results to sequences from other closely-related species.

Match to Comparator Legume	Doza Assembly		Farah Assembly	
	Transcripts with *E*-Value <10^{-50}	Transcripts with *E*-Value <10^{-10}	Transcripts with *E*-Value <10^{-50}	Transcripts with *E*-Value <10^{-10}
Chickpea genome	28,038	35,232	20,513	26,280
Soybean CDS	24,419	27,404	16,671	19,832
Medicago CDS	34,815	38,267	23,412	27,902
Medicago genome	32,916	39,968	24,289	30,265

The results of the comparison of Doza (Table S5) and Farah (Table S6) transcriptome assemblies to the previously published faba bean transcriptome sets (from Webb et al. [29] and Kaur et al. [22]) revealed that the current assemblies captured approximately 96% of transcripts from the Webb et al. [29] datasets and approximately 98% of contigs and 78% of singletons from the Kaur et al. [22] study. A total of 25.7% and 35.4% transcripts were found to be specific to the Doza and Farah transcriptome assemblies, respectively, based on comparison to the previously-published faba bean datasets. Reciprocal reference read mapping (Table S7) of Doza to Farah revealed that 92.2% of Doza-derived transcripts matched Farah-derived sequences. The reciprocal read mapping of Farah to Doza showed that 81.2% of Farah-derived transcripts exhibited matches to Doza-derived sequences. Furthermore, an overall comparison results of Farah versus Doza in addition to the datasets from Kaur et al. [22] and Webb et al. [29] studies are summarised in Figure 2. This identified a total of 8428 Doza-specific transcripts, 9437 Farah-specific transcripts and 9572 transcripts common to both the assemblies that was not previously characterised in *Vicia faba* L. datasets.
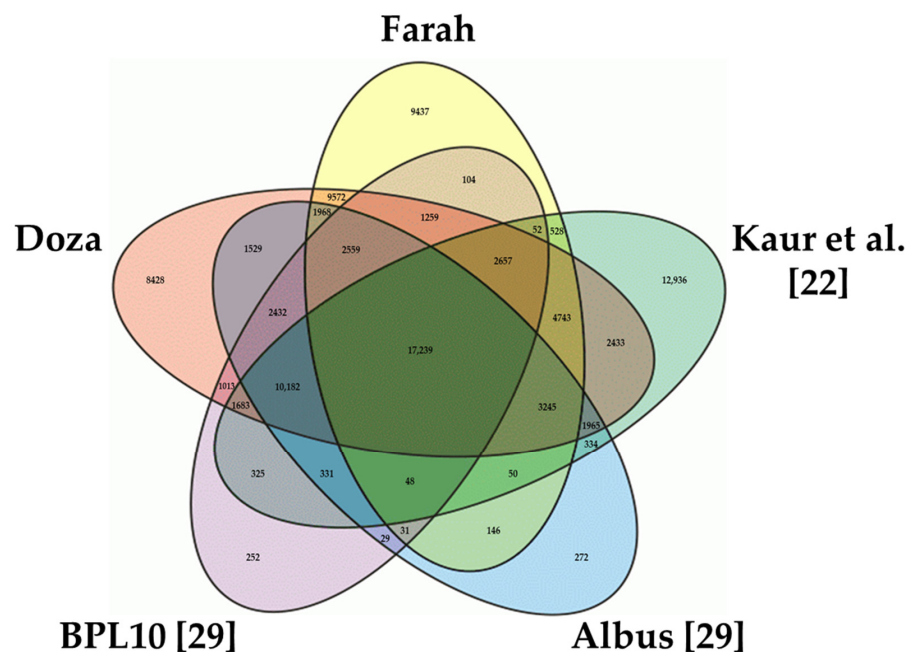

Figure 2. Venn diagram depicting the distribution of BLAST matches between Farah-specific and Doza-specific reference transcripts with sequences from Albus [29], BPL10 [29], and Kaur et al. [22] transcriptome studies.

To obtain gene function categories of the transcripts generated from Doza and Farah assemblies, Gene Ontology (GO) terms were assigned based on the sequence similarity to Nr databases. The analysis revealed a total of 30,581 transcripts from Doza and 27,285 transcripts from Farah were assigned at least one GO term. BLAST searches showed the highest similarity to *M. truncatula*, followed by chickpea (Figure S1). GO assignment to the biological process category was highest (41.9%), followed by cellular function (38.3%) and molecular function (19.9%; Figure S1). Among the biological process sub-categories, metabolic process (28.7%) and cellular process (24.5%) were prominently represented (Figure S1), indicating that tissues used in this study were undergoing extensive metabolic activity. A moderate number of transcripts were also involved in the single-organism process (15.9%), biological regulation (7.7%), regulation of biological process (6.6%), and response to stimulus (5.9%) categories. Under the molecular function category, catalytic activity (51.9%) and binding (48.1%) were the most common (Figure S1). For the cellular component category, the majority of the transcripts were assigned to the membrane (18.5%), cell (20.4%), cell part (20.1%), and organelle (13.8%) categories, while much smaller proportions were assigned to the organelle part and macromolecular complex categories (Figure S1).

A total of 50,506 (84.5%) Doza-specific transcripts and 40,462 (75.4%) Farah-specific transcripts exhibited matches to either one or all of the closely-related legume species. Following this process, totals of 2311 Doza-specific and 2511 Farah-specific uncharacterised transcripts were found to be annotated based on the BLAST results against Nt, Nr, and UniRef100 databases. However, 6925 transcripts (11.6%) from Doza and 10,665 transcripts (19.9%) from Farah were found to be uncharacterised based on these databases. The unannotated subsets from each assembly were searched for the presence of open reading frames (ORFs), which identified 5508 and 8592 transcripts from Doza and Farah, respectively. The reciprocal searches identified additional sets of 637 Doza-specific transcripts and 2436 Farah-specific transcripts. For Doza, the sequence length of the uncharacterised transcripts varied from 276 bp to 1355 bp with an average of 380 bp. For Farah, the sequence length of the uncharacterised transcripts ranged from 248 bp to 1142 bp with an average length of 398 bp. Hence, a total of 780 sequences from Doza and 363 sequences from Farah were removed from the final assemblies. The statistics of the final reference set for Doza and Farah is summarised in Table 4.

Table 4. Statistics of the filtered Doza and Farah datasets.

Statistics	Doza	Farah
Total number of sequences	58,962	53,275
Total length of transcripts	66,959,534 bp	60,943,125 bp
Number of transcripts with length <500 bp	17,518	17,606
Number of transcripts with length 500–1000 bp	15,482	12,860
Number of transcripts with length >1000 bp	25,962	22,809

2.3. Tissue-Specific Expression Analysis

Expression patterns within the faba bean transcript assemblies were analysed by aligning the sequence reads obtained from individual libraries to the cultivar-specific assembled transcriptome followed by normalization to the 75th percentile. It was found that 48,432 Doza-specific and 48,596 Farah-specific transcripts were present in all of the types of source tissue used in this study (Table S8). However, the expression level of these common transcripts varied significantly from one source tissue to another. For example, flower tissues from both the cultivars exhibited higher levels of transcripts corresponding to pollen-specific proteins, and leaf and stem tissue comprised of high levels of transcripts that had annotations to proteins involved in chlorophyll synthesis and photosynthesis. Similarly, in immature seeds and pods, the expression of seed-storage proteins, such as vicilin, convicilin, legumin, and embryonic abundant protein, was enriched. In contrast, only a small proportion of transcripts (352 in Doza and 192 in Farah) were expressed exclusively in one tissue-type group. For the Doza-specific assembly, a large number of transcripts were expressed in stem tissue,

while immature seeds exhibited a lower number of expressed transcripts (Figure 3). Conversely, for the Farah-specific assembly, the level of the expressed transcripts was higher in the flower tissue and at the lowest value for leaf tissue.
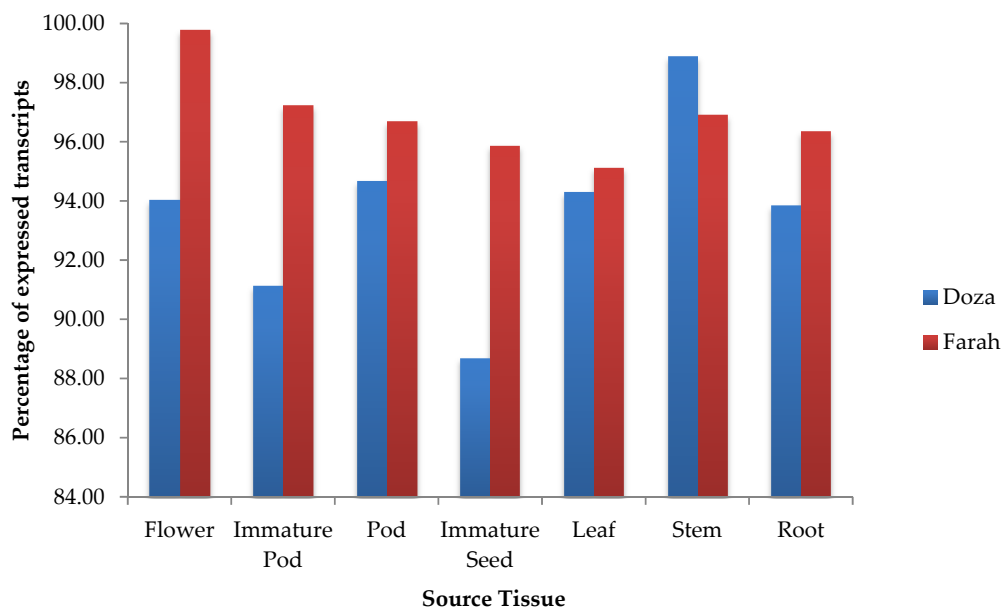

Figure 3. Differences in the level of gene expression observed from various tissue samples for Doza and Farah reference transcriptome assemblies.

Levels of expression were evaluated for 12 randomly-selected transcripts to analyse the level of expression among tissues from the transcriptome assembly by qRT-PCR. The transcripts associated with a range of functions in roots (peroxidase, lipoxygenase), pods and seeds (legumin, convicilin), flowers (pectinesterase, pollen-specific pectin methylesterase and leucine-rich repeat extension) and leaves and stems (photosynthesis-related proteins) were evaluated for the level of expression. The majority of the selected transcripts (10 out of 12) displayed good correlations between expression levels assessed by qRT-PCR and RNA-Seq (average Pearson's correlation coefficient of 0.999; Table S9). The remaining two transcripts showed only slight deviation from perfect concordance, with correlation coefficient values of 0.935 (discordant outcome for expression in pods) and 0.916 (discordant outcome for expression in flowers) (Table S9).

3. Discussion

3.1. De Novo *Transcriptome Assembly*

The present study describes the use of RNA-Seq to obtain comprehensive transcriptome assemblies from two Australian cultivars of faba bean (Doza and Farah) that differ in terms of breeding habit, adaptation characteristics, and disease resistance. Well-structured reference transcriptome assemblies for these cultivars were generated and characterised, with the aim of improving the application of marker-assisted selection strategies in breeding programs, in the absence of a complete faba bean whole-genome assembly.

RNA-Seq libraries from both Doza and Farah were prepared using multiple tissues obtained from three biological replicates and pooled before sequencing to ensure that genes expressed even at low levels were represented in the current assemblies. Sequencing from Farah generated 25% more data as compared to Doza, which may be due to the inconsistency in the quantification of RNA-Seq libraries. The final Doza assembly resulted in 58,962 transcripts with total assembly length of 66,959,534 bp (average length 1135.6 bp and N50 of 1595), whereas the final Farah assembly comprised

of 53,275 transcripts with total assembly length of 60,943,125 bp (average length 1143.9 bp and N50 of 1683). Both the Doza and Farah assemblies contained a substantial number of large transcripts with sequence length >500 bp (average 67.3%), which was comparable to the results obtained by previous studies [40,50] with a deep sequencing method for transcriptome generation. These results are highly comparable to those for the transcriptomes of other legume species, such as *M. truncatula,* which has a total of 66,028,174 bp (*M. truncatula* Genome Project v. 4.0), and soybean, which has 68,278,578 bp [51], and much higher than that of chickpea, which has 32,973,966 bp [52]. The average length statistics of both the assemblies were highly comparable to those for *M. truncatula* and chickpea, with average lengths of 1060 bp and 1166 bp, respectively.

3.2. Annotation of the Transcriptome Assemblies

A total of 83.5% and 71.7% of the transcripts from Doza and Farah, respectively (corresponding to 27,653 and 29,151 unigenes), were annotated based on the results of BLAST against the Nr database of NCBI. Based on the number of unigenes, it can be concluded that the Farah assembly is less fragmented (23.6%) than the Doza assembly (37.9%). The number of unigenes obtained from this study is highly comparable to those from other legume genomes including chickpea (28,269 gene models [52]) and lentil (27,396 unigenes based on the Nr database [50]). Moreover, BLASTX analysis revealed that the largest number of matches were to *M. truncatula* (around 43% and 34% for Doza and Farah, respectively), followed by chickpea (approximately 22% for Doza and 16% for Farah), and then any other plant species. This result is consistent with known phylogenetic relationships, as faba bean is more closely related to *M. truncatula* than to chickpea [12]. A small fraction of sequences in Doza (270), and a comparatively higher fraction in Farah (5753), displayed similarity matches to non-plant sequences based on BLAST results against NCBI and UniRef100 databases, probably due to the presence of contaminating seed-borne microbes and microbial communities in the rhizosphere, soil, and in planta.

Several transcriptome datasets have been previously generated for faba bean, from the cultivars Windsor [53], Icarus and Ascot [22], CDC Fatima, SSNS-1, and A01155 [39], Fiord [40,54], INRA-29H and Vf136 [42], BPL10 and Albus [29], and Wizard [43]. These datasets were generated from a selection of tissues including root, shoots, seedling, embryo, seed coat, leaves, or mixed tissues. The reference transcriptome assemblies from the current study were compared to the datasets obtained from the transcriptome studies of Webb et al. [29] and Kaur et al. [22]. The former was chosen as it has been used to generate a genetic linkage map with the densest SNP coverage that is currently available, while the latter used Australian cultivars, and a mixture of tissue types providing a broad survey of gene expression diversity. The transcriptome assembly from the present study was generated using the Illumina HiSeq 2000 second-generation sequencing system, while both of the previously-published faba bean datasets were obtained from GS FLX/454 reads. A high proportion of transcripts (96.5%) from Webb et al. [29] was captured in the current study, and 98% of contigs and 78% singletons from [22] were represented in the current reference transcriptome assembly. The remaining singletons (approximately 22%) that were not captured in the current assemblies could be specific to the Icarus/Ascot cultivars, or sequencing assembly artefacts. The two faba bean cultivars used in this study differ in terms of breeding habit and disease resistance characteristics. Doza differs from Farah in the response to biotic and abiotic stresses with Doza being resistant to rust infection and tolerant to frost events, whereas Farah is resistant to ascochyta blight. In addition to this, Farah is considered as the better yielding variety, compared to Doza. Reciprocal sequence analysis revealed an average of 14.2% transcripts that displayed no significant match to any transcript in the other cultivar which may account for some of the observed characteristics between the two faba bean varieties.

In total, 84.5% of Doza-specific transcripts and 75.4% of Farah-specific transcripts were annotated based on the comparisons to NCBI's Nt and Nr databases, UniRef100, comparator legume species genomes, and previously-published faba bean datasets. GO analysis revealed a total of 51.2% transcripts from Doza and 50.1% from Farah were assigned at least one GO term which is

comparable with other similar studies published in the literature [23,50]. The unannotated sequences (approximately 20.1%) were evaluated for the presence of ORFs and further for the similarity searches in the alternate cultivar that had annotations and the presence of ORF. The genes that were not annotated were more likely to be cultivar-specific and were included in the final assemblies for the completeness of the study as a previous study by Sudheesh et al. [50] also reported a similar percentage (21%) of unannotated genome components in their transcriptome study.

3.3. Analysis and Validation of Tissue-Specific Gene Expression Level

RNA-Seq libraries from multiple tissue sources were sequenced to obtain differential tissue gene expression levels based on the approach previously used in field pea and lentil [45,50]. Significant proportions of transcripts (96.3% in Doza and 96.7% in Farah) were expressed in more than one source tissue, but showed varying expression levels. The differential expression level from a selected set of transcripts was used to examine differential tissue expression obtained from the Doza and Farah transcriptome assemblies, and found to show good consistency in terms of the Pearson correlation coefficient. However, only two of twelve selected sequences failed to show concordance, possibly due to primer-design issues in the qRT-PCR test, or the effect of a complementary expression profile from a paralogous gene sequence.

3.4. Applications to Genomics-Assisted Breeding

A comparison of the Doza-derived and Farah-derived transcriptomes will permit identification of SNPs that discriminate between these cultivars, and which may be shared with other, particularly closely-related, germplasm. A sufficient number of genome-wide-distributed SNPs will be suitable for use in genome-wide association studies to identify genomic regions controlling key traits, and in genomic selection. SNPs can be formatted for detection at a number of different levels of the multiplex ratio, but the most obvious approach to implementation in a species such as faba bean is genotyping-by-sequencing (GBS) in the absence of a reference faba bean genome [55]. A GBS methodology based on sampling of the expressed component of the genome by RNA-Seq has been previously reported [56]. The unigene sets described in the present study will be a valuable resource for implementation of such a method, which ideally uses alignment to a high-quality reference assembly.

4. Materials and Methods

4.1. Plant Material

Three plants from each of the Doza and Farah cultivars were used in this study. The plants were germinated and maintained in standard potting mix in 200 mm plastic pots at 22 ± 2 °C with a photoperiod of 16/8-h (light/dark) within the glasshouse at AgriBio, Bundoora, Victoria, Australia. To prevent any problems with cross-pollination, the plants were isolated through the use of net enclosures during periods of flowering. Multiple tissue sources of plant material from both cultivars were harvested at various time points, in three replicates. Stem and leaf tissues from multiple nodes, along with roots, were sampled from four-week-old plants. Immature pods and fully-open flowers were collected within 8–12 days after flowering. Pods and immature seeds were sampled within 18–23 days post-flowering. The harvested tissue was snap-frozen in liquid nitrogen and stored at -80 °C until RNA extraction was performed.

4.2. RNA Extraction

The three replicates from each of the different source tissues were combined in equimolar quantities prior to grinding of tissue for the RNA isolation step to minimise variability across biological replicates. Total RNA was extracted and treated with DNase I (Qiagen, Hilden, Germany) using the RNeasy® Plant Mini Kit (Qiagen) following the manufacturer's protocol. The isolated total RNA samples were quantified using a spectrophotometer (Thermo-Scientific, Wilmington, DE, USA) at the

wavelength ratios of A260/230 and A260/230. The extracted samples were resolved on 1.2% (*w/v*) denaturing agarose gel to assess the integrity of RNA.

4.3. Library Preparation and Sequencing

RNA-Seq libraries were prepared with the SureSelect Strand-Specific RNA Library Kit according to the protocol described by the manufacturer, with the exception of the poly(A) RNA fragmentation time. The purified poly(A) RNA was fragmented to an approximate insert size of 350 bp at 94 °C for a minute, instead of 8 min as recommended in the protocol. The libraries were assessed on an Agilent TapeStation 2200 platform with D1000 ScreenTape (Agilent Technologies, Santa Clara, CA, USA) following the manufacturer's protocol. Each library was prepared with a unique indexing primer, and all the libraries were multiplexed in an equimolar concentration to generate a single pool. The multiplexed pooled sample was quantified using a KAPA library quantification kit (KAPA Biosystems, Boston, MA, USA) according to the protocol described by the manufacturer. The quantified sample was subjected to pair-end sequencing using the HiSeq 2000 system (Illumina Inc., San Diego, CA, USA).

4.4. Sequence Data Processing/Data Filtering and De Novo *Assembly*

The raw reads of sequences were filtered by employing a custom perl script and Cutadapt v. 1.9 [57]. Adaptor sequences and low quality reads (reads with >10% bases with Q $\leq$ 20) were removed from the resulting data. Trimming of the data involved removal of the reads that had three or more consecutive unassigned Ns with a phred score of $\leq$20. Sequence reads that were less than 50 bp were discarded prior to the *de novo* transcriptome assembly step. The filtered data was assembled using the transcriptome assembler, SOAPdenovo-TRANS [58] with k-mer size of 101. To generate more complete sequences with longer length, fork, bubble and complex loci from SOAPdenovo-TRANS assembly were further combined using the CAP3 assembler [59] with 95% identity and minimum overlap of 50 bp. Furthermore, the contigs and scaffolds having a total length of less than 240 bp were omitted, as these were considered shorter than the length of a single pair of the sequence.

4.5. Transcriptome Annotation

The Doza- and Farah-derived assemblies were analysed using BLASTN [60] and BLASTX [61] against the nucleotide (Nt) and protein (Nr) database maintained by NCBI with the threshold *E*-value of <10^{-10}. Both the assemblies were also searched against UniRef100 [62] using the same threshold parameter. The assemblies were further compared by performing a nucleotide search against the genomes of related legume species against the coding DNA sequences (CDSs) and the genome of *Medicago truncatula* Gaertn. (*M. truncatula* Genome Project v. 4.0 [63]), the chickpea (*Cicer arietinum* L.) genome [52], and soybean (*Glycine max* L.) CDSs [51].

For further analysis, those transcripts that displayed a significant match to non-plant databases based on their annotation were removed from both the assemblies. The transcripts were also BLASTN analysed against the previously-generated faba bean transcriptome databases of Kaur et al. [22] and Webb et al. [29]. The unannotated transcripts from both assemblies were searched for the presence of open reading frames (ORFs) using the 'getorf' command in the EMBOSS package [64]. The transcripts that returned no match as part of the ORF search were analysed for the presence of annotated contigs in the alternate cultivar. The assembled faba bean transcripts were characterised on the basis of Gene Ontology (GO) using the Blast2GO PRO software program [65] with the *E*-value threshold of <10^{-10}.

4.6. Tissue-Specific Expression Analysis

The BWA-MEM software package [66] was employed to generate tissue-specific expression profiles by aligning the reads obtained from each of the individual RNA-Seq libraries of Doza and Farah to their respective assembled transcriptome using the default parameters. The read counts were normalised as originally described by Sudheesh et al. [45] to generate the source tissue-specific expression profile. The normalised data from the specific source tissue was classified into the three

major groups, namely, reproductive (flower, pod, immature pod, and seed), vegetative (leaf and stem), and subterranean (root).

4.7. Validation of Tissue Expression Analysis

A set of twelve unigenes with differences in the level of expression were randomly selected based on their annotation and putative biological function from the three major groups as described above. RNA extractions from different tissues (leaf, stem, root, pod, and flower) of the 'Nura' cultivar of faba bean were performed as detailed above. The primer sequences for the selected transcripts based on their annotation NCBI's Nr database (Table S10) were designed using BatchPrimer3 [67] with default parameters for the product size of 100 to 120 bp, GC content ranging from 40% to 60% and an optimum annealing temperature between 55 and 60 °C. The GADPH gene was used as an internal reference gene. The qRT-PCR, melting curve analysis and normalisation of the obtained data against the internal control was performed as described by Sudheesh et al. [50]. The correlation between the RNA-Seq and qRT-PCR data was assessed by calculating the Pearson's correlation coefficient in Microsoft Excel.

5. Conclusions

In conclusion, the present study contributes to a total of 26,295 transcripts, with 7648 Doza-specific and 9075 Farah-specific transcripts, which have not been characterized previously in *Vicia faba*. The validation results confirm that the transcriptome assemblies are the most comprehensive to be generated for faba bean, and will be of significant value in genomics-assisted breeding, as well as a support for functional genomics studies.

References

1. Bohra, A.; Pandey, M.K.; Jha, U.C.; Singh, B.; Singh, I.P.; Datta, D.; Chaturvedi, S.K.; Nadarajan, N.; Varshney, R.K. Genomics-assisted breeding in four major pulse crops of developing countries: Present status and prospects. *Theor. Appl. Genet.* **2014**, *127*, 1263–1291. [CrossRef] [PubMed]

2. Alghamdi, S.S.; Migdadi, H.M.; Ammar, M.H.; Paull, J.G.; Siddique, K.H.M. Faba bean genomics: Current status and future prospects. *Euphytica* **2012**, *186*, 609–624. [CrossRef]

3. Food and Agricultural Organisation of the United Nations-FAO Statistical Database. Available online: http://faostat.fao.org (accessed on 18 January 2017).

4.   Gnanasambandam, A.; Paull, J.; Torres, A.; Kaur, S.; Leonforte, T.; Li, H.; Zong, X.; Yang, T.; Materne, M. Impact of molecular technologies on faba bean (*Vicia faba* L.) breeding strategies. *Agronomy* **2012**, *2*, 132. [CrossRef]

5.   Sillero, J.C.; Villegas-Fernández, A.M.; Thomas, J.; Rojas-Molina, M.M.; Emeran, A.A.; Fernández-Aparicio, M.; Rubiales, D. Faba bean breeding for disease resistance. *Field Crops Res.* **2010**, *115*, 297–307. [CrossRef]

6.   Khan, H.R.; Paull, J.G.; Siddique, K.H.M.; Stoddard, F.L. Faba bean breeding for drought-affected environments: A physiological and agronomic perspective. *Field Crops Res.* **2010**, *115*, 279–286. [CrossRef]

7.   Slabu, C.; Zörb, C.; Steffens, D.; Schubert, S. Is salt stress of faba bean (*Vicia faba*) caused by Na+ or Cl− toxicity? *JPNSS* **2009**, *172*, 644–651. [CrossRef]

8.   Arbaoui, M.; Balko, C.; Link, W. Study of faba bean (*Vicia faba* L.) winter-hardiness and development of screening methods. *Field Crops Res.* **2008**, *106*, 60–67. [CrossRef]

9.   El-Sherbeeny, M.H.; Robertson, L.D. Protein content variation in a pure line faba bean (*Vicia faba*) collection. *JSFA* **1992**, *58*, 193–196. [CrossRef]

10.   Kumar, R.; Singh, M. Tannins: Their adverse role in ruminant nutrition. *J. Agric. Food Chem.* **1984**, *32*, 447–453. [CrossRef]

11.   Choi, H.-K.; Mun, J.-H.; Kim, D.-J.; Zhu, H.; Baek, J.-M.; Mudge, J.; Roe, B.; Ellis, N.; Doyle, J.; Kiss, G.B.; et al. Estimating genome conservation between crop and model legume species. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 15289–15294. [CrossRef] [PubMed]

12.   Chooi, W.Y. Variation in nuclear DNA content in the genus vicia. *Genetics* **1971**, *68*, 195–211. [PubMed]

13.   Bennett, M.D.; Leitch, I.J. Nuclear DNA amounts in angiosperms: Progress, problems and prospects. *Ann. Bot.* **2005**, *95*, 45–90. [CrossRef] [PubMed]

14.   Ellwood, S.R.; Phan, H.T.; Jordan, M.; Hane, J.; Torres, A.M.; Avila, C.M.; Cruz-Izquierdo, S.; Oliver, R.P. Construction of a comparative genetic map in faba bean (*Vicia faba* L.); conservation of genome structure with lens culinaris. *BMC Genom.* **2008**, *9*, 380. [CrossRef] [PubMed]

15.   Neumann, P.; Koblížková, A.; Navrátilová, A.; Macas, J. Significant expansion of *Vicia pannonica* genome size mediated by amplification of a single type of giant retroelement. *Genetics* **2006**, *173*, 1047–1056. [CrossRef] [PubMed]

16.   Bond, D.A. Recent developments in breeding field beans (*Vicia faba* L.). *Plant. Breed.* **1987**, *99*, 1–26. [CrossRef]

17.   Torres, A.M.; Weeden, N.F.; Martín, A. Linkage among isozyme, rflp and rapd markers in *Vicia faba*. *Theor. Appl. Genet.* **1993**, *85*, 937–945. [CrossRef] [PubMed]

18.   Avila, C.M.; Sillero, J.C.; Rubiales, D.; Moreno, M.T.; Torres, A.M. Identification of rapd markers linked to the *Uvf-1* gene conferring hypersensitive resistance against rust (*Uromyces viciae-fabae*) in *Vicia faba* L. *Theor. Appl. Genet.* **2003**, *107*, 353–358. [CrossRef] [PubMed]

19.   Gutierrez, N.; Avila, C.M.; Rodriguez-Suarez, C.; Moreno, M.T.; Torres, A.M. Development of scar markers linked to a gene controlling absence of tannins in faba bean. *Mol. Breed.* **2007**, *19*, 305–314. [CrossRef]

20.   Zeid, M.; Mitchell, S.; Link, W.; Carter, M.; Nawar, A.; Fulton, T.; Kresovich, S. Simple sequence repeats (SSRs) in faba bean: New loci from orobanche-resistant cultivar 'giza 402'. *Plant. Breed.* **2009**, *128*, 149–155. [CrossRef]

21.   Gong, Y.-M.; Xu, S.-C.; Mao, W.-H.; Hu, Q.-Z.; Zhang, G.-W.; Ding, J.; Li, Z.-Y. Generation and characterization of 11 novel EST-derived microsatellites from *Vicia faba* (fabaceae). *Am. J. Bot.* **2010**, *97*, e69–e71. [CrossRef] [PubMed]

22.   Kaur, S.; Pembleton, L.W.; Cogan, N.O.I.; Savin, K.W.; Leonforte, T.; Paull, J.; Materne, M.; Forster, J.W. Transcriptome sequencing of field pea and faba bean for discovery and validation of SSR genetic markers. *BMC Genom.* **2012**, *13*, 104. [CrossRef] [PubMed]

23.   Suresh, S.; Kim, T.-S.; Raveendar, S.; Cho, J.-H.; Yi, J.Y.; Lee, M.C.; Lee, S.-Y.; Baek, H.-J.; Cho, G.-T.; Chung, J.-W. Transcriptome characterization and large-scale identification of SSR/SNP markers in symbiotic nitrogen fixation crop faba bean (*Vicia faba* L.). *Turk. J. Agric. For.* **2015**, *39*, 459–469. [CrossRef]

24.   Yang, T.; Bao, S.-Y.; Ford, R.; Jia, T.-J.; Guan, J.-P.; He, Y.-H.; Sun, X.-L.; Jiang, J.-Y.; Hao, J.-J.; Zhang, X.-Y.; et al. High-throughput novel microsatellite marker of faba bean via next generation sequencing. *BMC Genom.* **2012**, *13*, 602. [CrossRef] [PubMed]

25. Cruz-Izquierdo, S.; Avila, C.M.; Satovic, Z.; Palomino, C.; Gutierrez, N.; Ellwood, S.R.; Phan, H.T.; Cubero, J.I.; Torres, A.M. Comparative genomics to bridge *Vicia faba* with model and closely-related legume species: Stability of QTLs for flowering and yield-related traits. *Theor. Appl. Genet.* **2012**, *125*, 1767–1782. [CrossRef] [PubMed]

26. Satovic, Z.; Avila, C.M.; Cruz-Izquierdo, S.; Díaz-Ruíz, R.; García-Ruíz, G.M.; Palomino, C.; Gutiérrez, N.; Vitale, S.; Ocaña-Moral, S.; Gutiérrez, M.V.; et al. A reference consensus genetic map for molecular markers and economically important traits in faba bean (*Vicia faba* L.). *BMC Genom.* **2013**, *14*, 932. [CrossRef] [PubMed]

27. El-Rodeny, W.; Kimura, M.; Hirakawa, H.; Sabah, A.; Shirasawa, K.; Sato, S.; Tabata, S.; Sasamoto, S.; Watanabe, A.; Kawashima, K.; et al. Development of EST-SSR markers and construction of a linkage map in faba bean (*Vicia faba*). *Breed. Sci.* **2014**, *64*, 252–263. [CrossRef] [PubMed]

28. Kaur, S.; Kimber, R.B.; Cogan, N.O.; Materne, M.; Forster, J.W.; Paull, J.G. SNP discovery and high-density genetic mapping in faba bean (*Vicia faba* L.) permits identification of QTLs for ascochyta blight resistance. *Plant. Sci.* **2014**, *217–218*, 47–55. [CrossRef] [PubMed]

29. Webb, A.; Cottage, A.; Wood, T.; Khamassi, K.; Hobbs, D.; Gostkiewicz, K.; White, M.; Khazaei, H.; Ali, M.; Street, D.; et al. A SNP-based consensus genetic map for synteny-based trait targeting in faba bean (*Vicia faba* L.). *Plant. Biotechnol. J.* **2016**, *14*, 177–185. [CrossRef] [PubMed]

30. Avila, C.M.; Satovic, Z.; Sillero, J.C.; Nadal, S.; Rubiales, D.; Moreno, M.T.; Torres, A.M. QTL detection for agronomic traits in faba bean (*Vicia faba* L.). *Agric. Conspec. Sci.* **2005**, *70*, 65–73.

31. Arbaoui, M.; Link, W.; Satovic, Z.; Torres, A.M. Quantitative trait loci of frost tolerance and physiologically related traits in faba bean (*Vicia faba* L.). *Euphytica* **2008**, *164*, 93–104. [CrossRef]

32. Avila, C.M.; Satovic, Z.; Sillero, J.C.; Rubiales, D.; Moreno, M.T.; Torres, A.M. Isolate and organ-specific QTLs for ascochyta blight resistance in faba bean (*Vicia faba* L.). *Theor. Appl. Genet.* **2004**, *108*, 1071–1708. [CrossRef] [PubMed]

33. Román, B.; Torres, A.M.; Rubiales, D.; Cubero, J.I.; Satovic, Z. Mapping of quantitative trait loci controlling broomrape (*Orobanche crenata* Forsk.) resistance in faba bean (*Vicia faba* L.). *Genome* **2002**, *45*, 1057–1063. [CrossRef] [PubMed]

34. Díaz, R.; Torres, A.M.; Satovic, Z.; Gutierrez, M.V.; Cubero, J.I.; Román, B. Validation of QTLs for *Orobanche crenata* resistance in faba bean (*Vicia faba* L.) across environments and generations. *Theor. Appl. Genet.* **2010**, *120*, 909–919. [CrossRef] [PubMed]

35. Atienza, S.G.; Palomino, C.; Gutiérrez, N.; Alfaro, C.M.; Rubiales, D.; Torres, A.M.; Ávila, C.M. QTLs for ascochyta blight resistance in faba bean (*Vicia faba* L.): Validation in field and controlled conditions. *Crop. Pasture Sci.* **2016**, *67*, 216–224. [CrossRef]

36. Kumar, J.; Choudhary, A.K.; Solanki, R.K.; Pratap, A. Towards marker-assisted selection in pulses: A review. *Plant. Breed.* **2011**, *130*, 297–313. [CrossRef]

37. 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]

38. Garg, R.; Jain, M. RNA-Seq for transcriptome analysis in non-model plants. *Methods Mol. Biol.* **2013**, *1069*, 43–58. [PubMed]

39. Ray, H.; Bock, C.; Georges, F. Faba bean: Transcriptome analysis from etiolated seedling and developing seed coat of key cultivars for synthesis of proanthocyanidins, phytate, raffinose family oligosaccharides, vicine, and convicine. *Plant. Genome* **2015**, *8*, 1–11. [CrossRef]

40. Arun-Chinnappa, K.S.; McCurdy, D.W. *De novo* assembly of a genome-wide transcriptome map of *Vicia faba* (L.) for transfer cell research. *Front. Plant. Sci.* **2015**, *6*, 217. [CrossRef] [PubMed]

41. Madrid, E.; Horres, R.; Krezdorn, N.; Palomino, C.; Plötner, A.; Rotter, B.; Torres, A.M.; Winter, P. DeepSuperSage analysis of the *Vicia faba* transcriptome in response to *Ascochyta fabae* infection. *Phytopathol. Mediterr.* **2013**, *52*, 166–182.

42. Ocaña, S.; Seoane, P.; Bautista, R.; Palomino, C.; Claros, G.M.; Torres, A.M.; Madrid, E. Large-scale transcriptome analysis in faba bean (*Vicia faba* L.) under *Ascochyta fabae* infection. *PLoS ONE* **2015**, *10*, e0135143. [CrossRef] [PubMed]

43. Cooper, J.W.; Wilson, M.H.; Derks, M.F.L.; Smit, S.; Kunert, K.J.; Cullis, C.; Foyer, C.H. Enhancing faba bean (*Vicia faba* L.) genome resources. *J. Exp. Bot.* **2017**, *68*, 1941–1953. [CrossRef] [PubMed]

44. Libault, M.; Farmer, A.; Joshi, T.; Takahashi, K.; Langley, R.J.; Franklin, L.D.; He, J.; Xu, D.; May, G.; Stacey, G. An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. *Plant. J.* **2010**, *63*, 86–99. [CrossRef] [PubMed]

45. Sudheesh, S.; Sawbridge, T.I.; Cogan, N.O.I.; Kennedy, P.; Forster, J.W.; Kaur, S. *De novo* assembly and characterisation of the field pea transcriptome using RNA-Seq. *BMC Genom.* **2015**, *16*, 611. [CrossRef] [PubMed]

46. Moreton, J.; Izquierdo, A.; Emes, R.D. Assembly, assessment, and availability of *De novo* generated eukaryotic transcriptomes. *Front. Genet.* **2015**, *6*, 361. [CrossRef] [PubMed]

47. ABB Seeds-Doza (Faba Bean), NSW Department of Primary Industries. Available online: http://www.nvtonline.com.au/wp-content/uploads/2013/03/Fact-Sheet-Faba-Bean-Doza.pdf (accessed on 11 February 2017).

48. 2017 Sowing Guide, South Australia. Available online: http://pir.sa.gov.au/__data/assets/pdf_file/0005/268862/SA_Sowing_Guide_2017.pdf (accessed on 15 April 2017).

49. Plant Breeder's Rights, IP Australia. Available online: https://www.ipaustralia.gov.au/plant-breeders-rights (accessed on 30 January 2017).

50. Sudheesh, S.; Verma, P.; Forster, J.W.; Cogan, N.O.; Kaur, S. Generation and characterisation of a reference transcriptome for lentil (*Lens culinaris* Medik.). *Int. J. Mol. Sci.* **2016**, *17*, 1887. [CrossRef] [PubMed]

51. 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] [PubMed]

52. 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]

53. Ray, H.; Georges, F. A genomic approach to nutritional, pharmacological and genetic issues of faba bean (*Vicia faba*): Prospects for genetic modifications. *GM Crops* **2010**, *1*, 99–106. [CrossRef] [PubMed]

54. Zhang, H.-M.; Wheeler, S.; Xia, X.; Radchuk, R.; Weber, H.; Offler, C.E.; Patrick, J.W. Differential transcriptional networks associated with key phases of ingrowth wall construction in trans-differentiating epidermal transfer cells of *Vicia faba* cotyledons. *BMC Plant. Biol.* **2015**, *15*, 103. [CrossRef] [PubMed]

55. He, J.; Zhao, X.; Laroche, A.; Lu, Z.X.; Liu, H.; Li, Z. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front. Plant. Sci.* **2014**, *5*, 484. [CrossRef] [PubMed]

56. Harper, A.L.; Trick, M.; Higgins, J.; Fraser, F.; Clissold, L.; Wells, R.; Hattori, C.; Werner, P.; Bancroft, I. Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. *Nat. Biotechnol.* **2012**, *30*, 798–802. [CrossRef] [PubMed]

57. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet* **2011**, *17*, 10–12. [CrossRef]

58. 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. *Bioinformatics* **2014**, *30*, 1660–1666. [CrossRef] [PubMed]

59. Huang, X.; Madan, A. Cap3: A DNA sequence assembly program. *Genome Res.* **1999**, *9*, 868–877. [CrossRef] [PubMed]

60. Chen, Y.; Ye, W.; Zhang, Y.; Xu, Y. High speed BLASTN: An accelerated megablast search tool. *Nucleic Acids Res.* **2015**, *43*, 7762–7768. [CrossRef] [PubMed]

61. Altschul, S.F.; Madden, T.L.; Schaffer, 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]

62. Suzek, B.E.; Huang, H.; McGarvey, P.; Mazumder, R.; Wu, C.H. UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **2007**, *23*, 1282–1288. [CrossRef] [PubMed]

63. *Medicago Truncatula* Genome Database. Available online: http://www.medicagogenome.org/ (accessed on 15 November 2016).

64. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The european molecular biology open software suite. *Trends Genet.* **2000**, *16*, 276–277. [CrossRef]

65. Conesa, A.; Gotz, S.; Garcia-Gomez, J.M.; Terol, J.; Talon, M.; Robles, M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **2005**, *21*, 3674–3676. [CrossRef] [PubMed]

66. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.* **2013**. Available online: http://arxiv.org/abs/1303.3997 (accessed on 15 September 2016).

67. You, F.M.; Huo, N.; Gu, Y.Q.; Luo, M.C.; Ma, Y.; Hane, D.; Lazo, G.R.; Dvorak, J.; Anderson, O.D. BatchPrimer3: A high throughput web application for PCR and sequencing primer design. *BMC Bioinform.* **2008**, *9*, 253. [CrossRef] [PubMed]