

Article

NB-TeaBase: A Multi-Omics Database and Genomic Selection Platform for Robust-Bud Tea Breeding

Yang Chen ^{1,2}, Lizhong Wang ³, Longjie Zhang ⁴, Keming Chen ³, Minghua Lou ^{1,2}, Dengfeng Shen ^{1,2}, Bin Wei ^{1,2} and Jianhong Zhang ^{1,2,*}

¹ Ningbo Academy of Agricultural Sciences, Ningbo 315040, China; cyevol@zju.edu.cn (Y.C.)

² Ningbo Key Laboratory of Characteristic Horticultural Crop Quality Regulation and Resistance Breeding, Ningbo 315040, China

³ Ningbo Fenghua District Agricultural Technology Service Station, Ningbo 315040, China

⁴ Ningbo Huangjinyun Tea Technology Co., Ltd., Ningbo 315040, China

* Correspondence: nbjianhong@163.com

Abstract

Tea breeding increasingly requires a continuous evidence chain from germplasm identity to molecular variation and progeny performance, but cultivar passports, genome resources, transcriptome profiles, metabolite data and breeding records are often managed separately. Here, we present Ningbo TeaBase (NB-TeaBase), an open, cultivar-centered database for robust-bud tea (*Camellia sinensis*) breeding. The platform uses named cultivars as the organizing unit and integrates germplasm passport information for eight elite cultivars with 23 chromosome-level genomes, whole-genome resequencing variants, RNA-seq expression profiles, untargeted LC-MS/MS metabolite features, annotations and downloadable records. Shared cultivar, gene, variant, pathway and metabolite identifiers allow users to characterize a cultivar across genomic, transcriptomic and metabolic layers and to connect these profiles with breeding records. A parental-prediction module incorporates genome-wide markers and phenotypes from 122 progeny assigned to seven observed parental combinations; it reports genomic estimated breeding values (GEBVs) for candidate crosses and observed-family general and specific combining-ability summaries for seven traits. By combining germplasm documentation, multi-omics evidence and progeny-informed cross ranking, NB-TeaBase supports cultivar evaluation, parent selection, cross prioritization and breeding strategy formulation. Current prediction outputs remain exploratory because the phenotypic training data cover only seven parental combinations and require broader validation.

Keywords: *Camellia sinensis*; robust-bud tea; multi-omics database; cultivar passport; metabolomics; transcriptomics; variome; genomic selection; combining ability; precision breeding



Academic Editor: Jing Zhuang

Received: 1 June 2026

Revised: 26 June 2026

Accepted: 29 June 2026

Published: 30 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Tea plant (*Camellia sinensis*) is one of the most important perennial beverage crops, and cultivar quality strongly determines the economic value of tea production. Cultivated tea is predominantly diploid ($2n = 2x = 30$), with 15 chromosome pairs and a haploid genome of approximately 3.0–3.2 Gb [1,2]. In premium green- and white-tea systems, robust-bud cultivars are valued for bud vigor, trichome abundance, early spring growth, appearance and quality-related chemical traits. Their improvement remains slow because tea has a

long juvenile period and because parental combinations must be evaluated across multiple years and environments.

High-throughput omics technologies have changed the scale at which tea germplasm can be characterized. Public resources now provide chromosome-scale *Camellia* genomes, population resequencing datasets, transcriptome resources and metabolite information. TPIA and TPIA2 established broad tea genomics and bioinformatics gateways [3,4]; TeaGVD and TeaPVs emphasize genomic variation and genotype-to-phenotype information [5,6]; TeaCoN and TeaAS provide transcript-derived functional views [7,8]; and TeaPGDB supports genome browsing and sequence-oriented analysis [9]. In parallel, smart-breeding platforms and genomic-enviromic prediction frameworks demonstrate how germplasm, phenotypic, genomic and analytical modules can be connected for breeding management and prediction [10,11].

Despite these advances, broad or gene-centered resources do not automatically answer cultivar-level breeding questions. A breeder planning crosses among elite robust-bud materials needs to know which parent carries useful variation, whether candidate genes are expressed in relevant cultivars, whether quality-related metabolites differ among parents and which crosses merit priority for field evaluation. These questions require a database architecture organized around breeding materials and decisions, rather than only around species, genomes, variants or gene identifiers.

We developed Ningbo TeaBase (NB-TeaBase) to address this need in robust-bud tea breeding. The current release integrates passports and multi-omics profiles for eight elite cultivars with a parental-prediction dataset comprising 122 progeny from seven parental combinations. It provides genome and variant search, expression and metabolite comparison, sequence tools, data download and cross-ranking functions through a single web interface.

NB-TeaBase is organized around three linked operations. Cultivar passports establish material identity and breeding context; linked genomic, transcriptomic and metabolomic records support molecular interpretation; and the parental-prediction module converts marker and phenotype data into exploratory cross rankings. Together, these operations create a traceable route from cultivar evaluation to parent selection, cross prioritization and breeding-strategy formulation. We describe the data architecture, implementation and representative workflows, and define the evidence boundaries of the current release.

2. Materials and Methods

2.1. Method Overview

NB-TeaBase uses cultivar, gene, variant, pathway and compound identifiers to connect heterogeneous records. Images and structured cultivar information are stored with searchable metadata, whereas sequence, variant, expression and metabolite records retain standard biological formats and identifiers. These links allow a query to move between a named cultivar, its molecular records and the corresponding breeding outputs.

The platform contains three functional layers: data curation and storage, cross-module retrieval and visualization, and parental prediction. The first layer standardizes incoming records; the second exposes linked searches and downloads; and the third produces GEBV, GCA and SCA summaries from marker and phenotype inputs.

2.2. Plant Materials

Eight elite robust-bud cultivars of *Camellia sinensis* were selected as the core multi-omics panel: FDDH, FDDDB, YS, CF, ZN117, ZN113, FY6 and BHZ. Young leaf and bud tissues from these cultivars were used for resequencing, RNA-seq and metabolomic profil-

ing. Cultivar descriptions and images were compiled from breeding records and project-generated materials.

Parental prediction used a separate breeding dataset of 122 progeny from Huangjinya (HJY), Yujinxiang (YJX), Zijuan (ZJ), Simingxueya (SMXY), Huangjinjia (HJJ), Wuniuzao (WNZ) and Pingyangtezao (PYTZ). Reciprocal crosses and synonymous names were harmonized into seven unordered parental combinations containing 2–55 progeny. Non-missing counts for Traits 1–7 were 122, 122, 122, 122, 118, 104 and 100, respectively.

2.3. Resequencing and Variant Processing

Genomic DNA libraries were prepared from 0.2 µg of DNA per sample after fragmentation to approximately 350 bp and sequenced on the DNBSEQ-T7 platform. Raw paired reads were filtered with fastp v0.23.1 [12]. Read pairs were discarded when either read contained adapter sequence, more than 10% undetermined bases or more than 50% bases with Phred quality ≤ 5 . Clean reads were aligned to the FDDH reference genome with BWA-MEM [13], and SNPs and InDels were called with GATK [14]. The filtered variant set was imported into the variome module.

2.4. RNA-Seq Processing

mRNA libraries from the eight cultivars were sequenced on the DNBSEQ-T7 platform. Adapter-containing, poly-N and low-quality reads were removed with fastp v0.19.7. Paired-end clean reads were aligned to the reference genome with HISAT2 v2.2.1 [15]. Reads with mapping quality < 10 , non-paired alignments and multimapping alignments were excluded before gene-level counting with featureCounts v2.0.6 [16]. Reference-guided transcript assembly used StringTie v2.2.3 [17]. Gene expression was normalized as FPKM and imported into the transcriptome module.

2.5. Metabolome Processing

For each sample, 100 mg of tissue was ground in liquid nitrogen, extracted with prechilled 80% methanol, incubated on ice for 5 min and centrifuged at $15,000 \times g$ and 4 °C for 20 min. The supernatant was diluted to 53% methanol, centrifuged again under the same conditions and analyzed with a Vanquish UHPLC system (Thermo Fisher Scientific, Germering, Germany) coupled to a Q Exactive HF-X mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). Raw files were converted to mzXML with ProteoWizard [18] and processed with XCMS for peak detection, retention-time alignment and quantification [19]. Features missing from more than 50% of samples in any group or with a coefficient of variation $> 30\%$ in quality-control samples were removed. Metabolites were matched at 10 ppm mass tolerance against an in-house MS/MS library and annotated with KEGG [20], HMDB [21] and LIPID MAPS [22]. Differential features were defined by $VIP > 1$, fold change > 1.5 or < 0.667 , and $p < 0.05$.

2.6. Genomic Selection and Combining Ability

The parental-prediction analysis used phenotypes from 122 progeny. Reciprocal crosses and synonymous parent names were harmonized into seven unordered parental combinations, and progeny records were averaged within combination for each trait. Genome-wide SNPs were partitioned into 10 kb, 100 kb and 1 Mb windows, and parent-pair similarity was calculated from informative genotypes within each window. The seven combination means and genomic-similarity features were analyzed with Bayesian A, Bayesian B, Bayesian C, Bayesian LASSO, Bayesian ridge regression, genomic BLUP, LASSO, rrBLUP and support vector regression [23–27]. These models generated GEBVs for candidate parent pairs. Because the training set contained only seven observed combina-

tions, model outputs were retained as exploratory predictions rather than used to select a preferred algorithm.

Each trait was summarized by record count, missingness, mean, standard deviation, coefficient of variation and normality. Diagnostics were reported without automatic transformation or model substitution. Because only seven combination means were available, model comparisons and cross rankings were treated as exploratory.

Combining ability was calculated separately from the seven observed combination means. For parent i , GCA was the mean deviation of combinations involving that parent from the grand mean. For parents i and j , SCA was calculated as $SCA(i,j) = Y(i,j) - [Y(\text{grand}) + GCA(i) + GCA(j)]$, where $Y(i,j)$ is the observed combination mean. These observed-family summaries were calculated independently of the model-derived GEBV predictions for candidate parent pairs.

2.7. Database Construction

NB-TeaBase was deployed as a browser/server web application with modules for cultivar passport, genome search, variome query, expression visualization, metabolite comparison, sequence tools and parental prediction. JBrowse 2 v1.4.4 was used for genome visualization [28], BLAST+ v2.12.0 for sequence similarity search [29] and Primer3 v2.6.1 for primer design [30]. Interactive visualizations include heatmaps, volcano plots, PCA displays and prediction tables. The database is publicly accessible at <http://112.126.28.109:8082> (accessed on 26 June 2026).

3. Results

3.1. Database Overview and Cultivar-Centered Design

NB-TeaBase uses the cultivar passport as its primary navigation unit. Each passport links cultivar identity, origin, breeding background, morphology, images and available omics records. This structure keeps molecular results attached to the named materials used in breeding.

The database implements a three-stage breeding workflow (Figure 1). Passport records support parent identification; linked omics records support the interpretation of candidate traits and pathways; and genomic prediction produces trait-specific cross rankings. Field observations can subsequently be returned to the database as additional phenotypic evidence, creating a data feedback loop for future model updates.

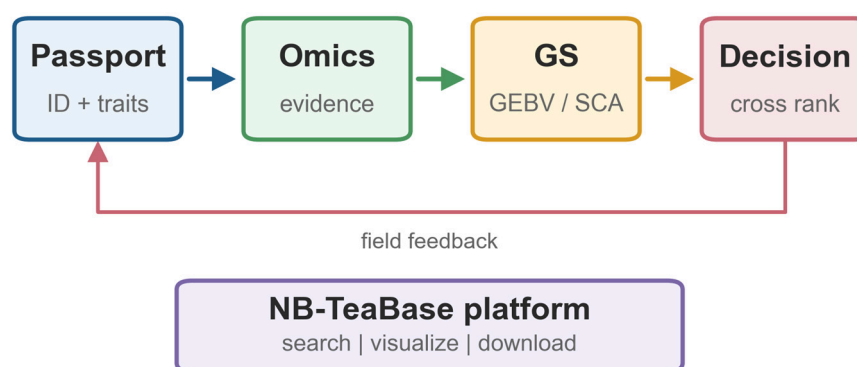


Figure 1. Conceptual architecture of NB-TeaBase. NB denotes Ningbo. The platform connects cultivar passports, genome and variome resources, transcriptome profiles, metabolome comparisons and pilot parental prediction into a Passport–Omics–GS Prediction workflow. The technical stack includes nginx v1.20.1, Docker Engine v20.10.24, Django REST Framework v3.14.0, MongoDB v6.0.8 with mongoengine v0.27.0, React v17.0.2, TypeScript v4.9.5 and Ant Design Pro v5.2.0.

Five evidence layers support this workflow: cultivar passports, genome and variome resources, transcriptomes, metabolomes and parental-prediction outputs. Common identifiers connect these layers and allow users to move from cultivar comparison to molecular evidence or from a candidate gene to relevant cultivars and breeding combinations.

3.2. Relationship to Existing Databases

NB-TeaBase complements rather than duplicates existing tea resources. TPIA2 provides the broadest Camellia-wide coverage [3,4], TeaGVD and TeaPVs emphasize genomic variation [5,6], TeaCoN and TeaAS provide transcript-derived functional information [7,8], and TeaPGDB supports genome-oriented search [9]. NB-TeaBase has a narrower germplasm scope, but links cultivar passports, multi-omics records and cross rankings for a defined robust-bud breeding program (Table 1).

Table 1. Positioning of NB-TeaBase relative to representative plant database papers and breeding platforms.

Resource	Primary Contribution	Main Limitation for Robust-Bud Breeding	NB-TeaBase Positioning
TPIA/TPIA2 [3,4]	Comprehensive Camellia gateway integrating genomes, variation, transcriptomes, metabolomes, methylomes and analysis tools.	Broad species-level and community-scale resource; not organized around robust-bud parent selection or cross ranking.	NB-TeaBase borrows the multi-omics database logic but re-centers it on elite robust-bud cultivars and breeding decisions.
TeaPGDB [9]	Published tea genomes with JBrowse, gene search, SSR search, BLAST and download services.	Genome-centered; does not provide cultivar-specific multi-omics linkage or parent prediction.	NB-TeaBase retains genome browsing but adds robust-bud cultivar passports, metabolome linkage and GS outputs.
TeaPVs [6]	Tea variation database integrating SNP, Indel, SV, SSR, expression search, BLAST and sequence extraction.	Variation-centered; less focused on breeding decision loops among elite robust-bud parents.	NB-TeaBase uses variation as one layer in a Passport-Omics-GS Prediction workflow.
TeaGVD [5]	Large-scale genomic variation and genotype-to-phenotype associations for tea metabolic and agronomic traits.	Excellent for variation-G2P mining but not designed as a cultivar passport and cross-prioritization platform.	NB-TeaBase links cultivar-level variants with expression, metabolites and parental prediction in a focused breeding panel.
TeaCoN [7]	Genome-wide tea co-expression network with expression heatmaps, network visualization, enrichment analysis and BLAST.	Co-expression-centered; candidate prioritization is not directly connected to parent-combination prediction.	NB-TeaBase uses expression as one evidence layer and connects it to metabolites and breeding outputs.
TeaAS [8]	Alternative-splicing database based on diverse tea RNA-seq datasets, with search, browser and download functions.	Transcript-isoform resource; not intended for robust-bud cultivar comparison or GS-based cross design.	NB-TeaBase complements transcript resources by emphasizing cultivar-specific trait and breeding interpretation.
Smart Breeding Platform [10]	General web platform for germplasm, testing, genomic data, GWAS and genomic selection.	Powerful generic breeding framework but not tea-specific and not linked to tea omics interpretation.	NB-TeaBase adapts the smart-breeding concept to tea cultivar passports and tea-specific omics evidence.
NB-TeaBase	Cultivar-centered robust-bud tea database with multi-omics linkage and parental prediction.	Current release still needs expanded validation populations and long-term maintenance documentation.	Provides a tea-specific bridge from resource browsing to breeding decision support.

3.3. Data Contents

3.3.1. Cultivar Passports

The passport module contains eight elite robust-bud cultivars: FDDH, FDDB, YS, CF, ZN117, ZN113, FY6 and BHZ. Each record includes the cultivar name and abbreviation, origin, breeding background, morphology, agronomic notes and representative images. Links from the passport page retrieve the corresponding genomic, transcriptomic and metabolomic records (Figure 2).

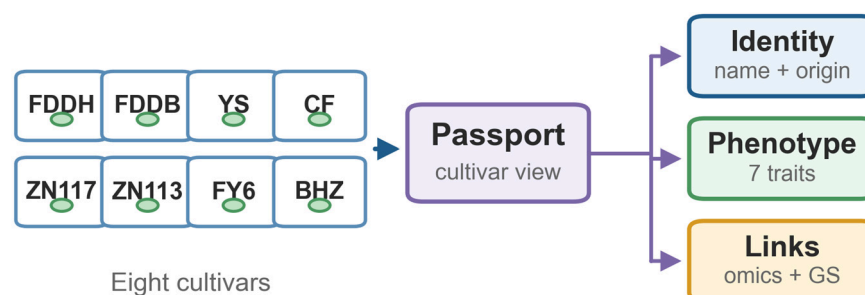


Figure 2. Cultivar passport module for eight elite robust-bud tea cultivars, including cultivar information, images and links to omics records.

3.3.2. Genome and Annotation Resources

The genome module contains 23 independently assembled chromosome-level *C. sinensis* genomes, including public assemblies and an in-house FDDH reference. The number 23 denotes genome assemblies, not chromosomes; a haploid tea assembly comprises 15 chromosome-scale pseudomolecules [1,2]. Gene models were annotated with Arabidopsis homologs, Swiss-Prot, NR, KEGG, Gene Ontology and Pfam. Users can search by gene identifier, functional term or protein domain and inspect local structure in the embedded genome browser.

The module supports gene-first and region-first queries. A gene query returns annotation, position and sequence information, whereas a region or marker query retrieves nearby genes and variants for subsequent interpretation (Figure 3).

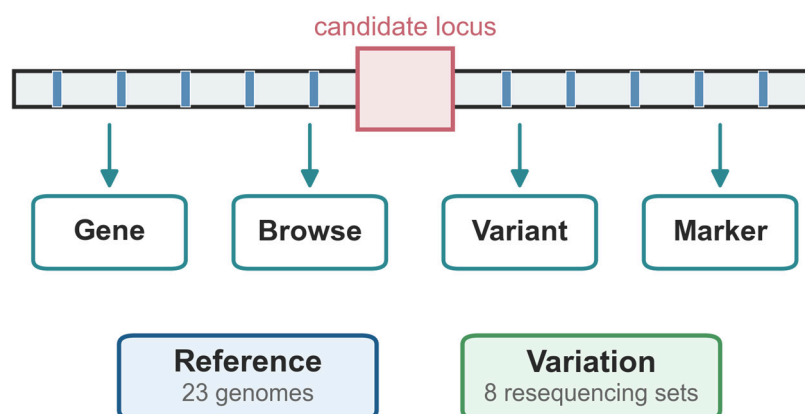


Figure 3. Genome and variome modules. Users can move from a candidate locus to gene search, genome browsing, variant comparison and marker-support functions.

3.3.3. Variome Resources

Resequencing of the eight core cultivars generated 395.04 Gb of raw data and 390.99 Gb of clean data. Clean-read retention ranged from 98.60% to 99.16%, and Q30 ranged from 95.58% to 97.24%. The resulting variant records can be searched by cultivar and genomic position and exported for marker comparison and parental prediction.

Connecting variant records to cultivar passports allows users to compare alleles among named breeding materials rather than across anonymous samples. Candidate variants can then be examined together with gene annotation, expression and metabolite information.

3.3.4. Transcriptome Resources and Transcript-Metabolite Linkage

The transcriptome module contains RNA-seq profiles for the eight cultivars. Sequencing generated 58.27 Gb of clean data, with Q30 values of 95.21–95.97%, alignment rates of 92.14–93.99% and uniquely mapped-read proportions of 86.52–89.40%. Gene-level FPKM values support search, cultivar comparison, correlation display and PCA.

Across cultivars, genes with FPKM > 60 represented 3.49–4.82% of annotated genes, whereas most expressed genes occurred in lower abundance classes. The interface displays these distributions and retrieves the expression profile of any indexed gene across the cultivar panel.

Transcript and metabolite records are linked through cultivar and pathway context. From a candidate gene, users can inspect expression across cultivars and retrieve metabolite contrasts for the same materials. Conversely, a differential metabolite can be traced to its pathway and associated gene-expression profiles (Figure 4).

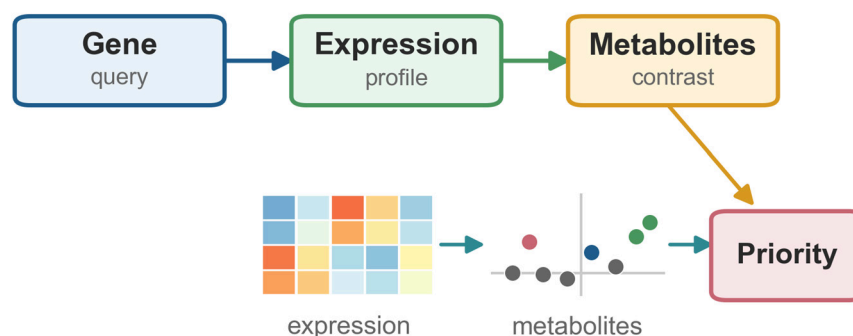


Figure 4. Transcriptome–metabolome linkage workflow. A candidate-gene query can be connected to cultivar expression profiles, pairwise metabolite contrasts and candidate prioritization.

3.3.5. Metabolome Resources

The metabolome module contains untargeted LC-MS/MS profiles from 24 samples. After peak extraction, alignment, normalization and QC filtering, 2456 positive-ion and 1746 negative-ion features were retained. Annotations combine accurate mass, MS/MS spectra, retention behavior and database matches. Users can compare cultivars, visualize differential accumulation and download feature-level statistics.

Metabolite records remain linked to cultivar identity and pathway annotation. This design supports direct comparison of chemical phenotypes relevant to taste, aroma, color and processing quality and provides molecular context for subsequent candidate-gene analysis.

Concordant variation, expression and metabolite patterns are treated as evidence for candidate prioritization, not as proof of causal regulation.

3.3.6. Parental Prediction Resources and Diagnostics

The parental-prediction module covers seven traits: tree posture, one-bud-and-one-leaf initiation stage, color of one bud and two leaves, trichome density, leaf morphology, calyx color and style-splitting position. The analysis used 100–122 non-missing progeny observations per trait, summarized as seven parental-combination means. Outputs distinguish model-derived GEBVs for candidate parent pairs from GCA and SCA calculated from the observed combination means.

The output archive preserves results by trait, genomic-window size and algorithm and records unavailable or non-converged runs as missing. Because the dataset contains

seven independent parental combinations, apparent differences among models were not interpreted as generalizable prediction accuracy.

Trait-level QC reports sample size, missingness, mean, standard deviation, coefficient of variation and normality. The normality test was applied to the seven combination means. No automatic transformation was applied; trichome density, which failed the diagnostic, was retained with an explicit caution flag. Collectively, these modules integrate cultivar passport information, genomic resources, variation data, transcriptomic profiles, metabolomic features and parental-prediction outputs. The content, primary functions and breeding relevance of these data layers are summarized in Table 2. The processing and reporting workflow of the parental-prediction module, including phenotype quality control, genomic-window comparison, model evaluation, output diagnostics and combining-ability calculation, is summarized in Table 3.

Table 2. Core data layers in the current NB-TeaBase release.

Layer	Current Content	Primary Function	Breeding Value
Cultivar passport	Eight robust-bud cultivars with descriptions and images	Cultivar browsing and comparison	Defines the parental materials used in breeding
Genome	23 chromosome-level tea genomes and harmonized annotations	Gene search, genome browser, sequence retrieval	Candidate gene and locus interpretation
Variome	390.99 Gb clean resequencing data from eight cultivars	Variant discovery and cultivar comparison	Marker development and genetic diversity assessment
Transcriptome	58.27 Gb clean RNA-seq data from eight cultivars	Expression search and visualization	Functional support for candidate genes
Metabolome	24 samples; 2456 positive-ion and 1746 negative-ion features	Compound search and pairwise differential analysis	Quality-related biomarker discovery
Parental prediction	Seven traits with model-derived GEBVs and observed-family GCA/SCA summaries	Genomic prediction and descriptive combining-ability reporting	Generates hypotheses for candidate parent pairs and subsequent field evaluation

Table 3. Data-processing and reporting steps in the parental-prediction module.

Step	Implementation	Purpose
Phenotype QC	N, valid N, missing rate, mean, SD, CV and a normality diagnostic are reported. The diagnostic flags non-normality but does not automatically transform the data.	Identifies traits that require cautious interpretation or distribution-appropriate modeling in expanded datasets.
Genomic-window comparison	Parent-pair genomic similarity is evaluated with 10 kb, 100 kb and 1 Mb SNP windows.	Assesses sensitivity to the genomic resolution used to summarize parental similarity.
Model ensemble	BA, BB, BC, BL, BRR, gBLUP, LASSO, rrBLUP and SVR are evaluated.	Reduces dependence on a single statistical assumption.
Output diagnostics	Model-specific prediction and status files are retained; unavailable or non-converged outputs are recorded as missing.	Makes incomplete or unstable model runs visible rather than silently omitting them.
Combining-ability summaries	Observed combination means are used to calculate parental GCA and cross-specific SCA; model-derived GEBVs are reported separately.	Separates descriptive combining ability from genomic prediction output.

3.4. FAIR Data Access and Download

NB-TeaBase implements the FAIR principles through searchable identifiers, open browsing, standard file formats and downloadable records [31]. Cultivar names, gene identifiers, genomic coordinates, compound identifiers and trait labels support findability. Public web access and file downloads support accessibility; FASTA, VCF and common annotation systems support interoperability; and documented outputs and repository accessions support reuse.

3.5. Database Implementation and User Tools

NB-TeaBase is deployed on an Alibaba Cloud server running Linux CentOS. Docker containers manage deployment, nginx handles static files and reverse proxying, and uWSGI serves the Django REST Framework back end. Python 3.9.7 provides the application logic, and MongoDB with mongoengine stores heterogeneous document records.

The front end uses React v17.0.2, TypeScript v4.9.5, UmiJS v3.5.43 and Ant Design Pro v5.2.0. API, task, database-query and file-stream services are separated in the back end. JWT authentication is available for restricted functions, whereas routine browsing and public downloads do not require registration.

Integrated tools include JBrowse for genome exploration [28], BLAST+ for sequence search [29], Primer3 for primer design [30], batch sequence extraction, heatmaps, PCA, volcano plots and downloadable result tables. The parental-prediction interface displays GEBV, GCA and SCA outputs by trait.

3.6. Recommended User Workflows

Three entry routes illustrate how the modules are used. Cultivar comparison links passport, expression and metabolite records for selected parents. Candidate-gene search links annotation and genomic context to expression, variants and metabolites. Cross design links trait-level phenotype summaries to model outputs and ranked parental combinations.

Together, these routes connect material identification, molecular interpretation and cross prioritization. The same identifiers are retained across modules, allowing each output to be traced to its source cultivar, gene, feature or phenotype record.

3.7. Utility Demonstrations

3.7.1. Overview: From Data Storage to Breeding Decisions

These demonstrations show how NB-TeaBase converts separated data layers into a traceable breeding workflow. They assess retrieval, cross-module linkage and output traceability rather than causal gene function or field-validated prediction accuracy. In practice, a breeder can begin with passport records to verify cultivar identity, origin and trait descriptions; inspect genome and variant records to identify candidate loci or allelic differences; compare transcriptome and metabolome profiles to judge whether the same cultivar contrast has molecular support; and then consult the parental-prediction module for a progeny-informed ranking of candidate crosses. This multi-layered framework provides a broader evidence base for cultivar evaluation, parent selection, cross prioritization and breeding strategy formulation. Its immediate practical value is not to replace multi-environment field testing, but to reduce manual reconciliation across independent files and make preliminary breeding decisions more transparent and reproducible.

3.7.2. Use Case 1: FDDH Versus FDDDB Metabolomic Differentiation

FDDH and FDDDB were compared to demonstrate cultivar-level metabolite retrieval. The query returned 854 differentially accumulated positive-ion features, of which 446 were

higher and 408 lower in FDDH. In negative-ion mode, 620 features differed, with 457 higher and 163 lower in FDDH (Figure 5).

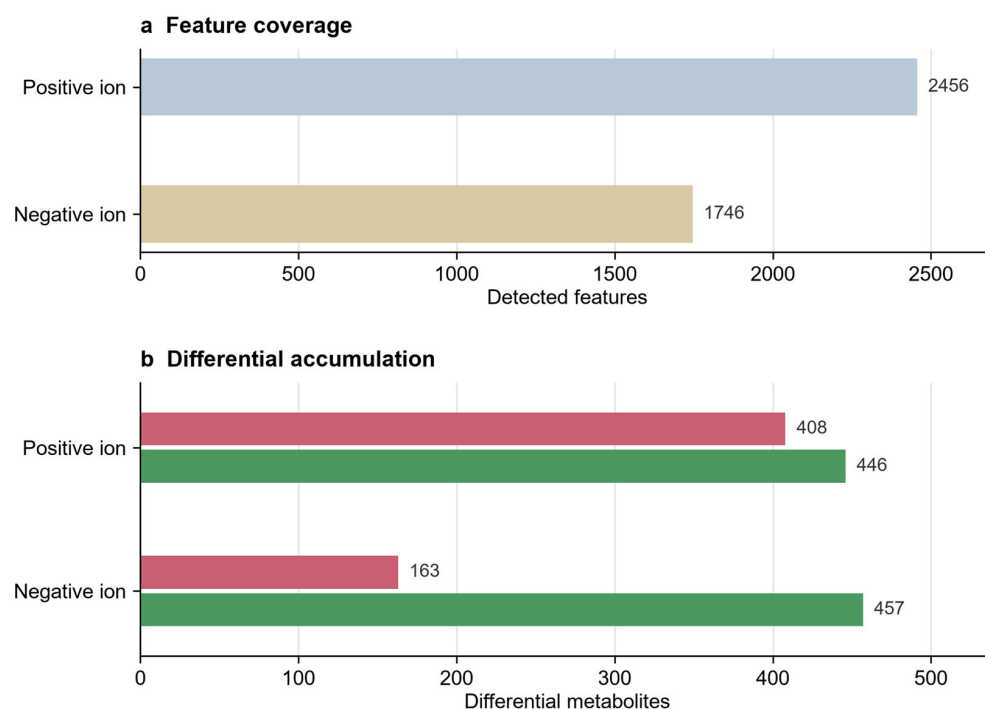


Figure 5. Metabolome feature coverage and differential accumulation in FDDH versus FDDH. (a) Total detected metabolic features in positive- and negative-ion modes. (b) Differential metabolites higher in FDDH (green) or lower in FDDH (red) for each ion mode.

For each comparison, users can inspect volcano plots and heatmaps, filter compound-level statistics, follow pathway annotations and download the result table. The FDDH-FDDH example therefore links a named cultivar comparison to reusable metabolite evidence.

3.7.3. Use Case 2: Linked Transcript-Metabolite Query

The linked query begins with either a gene or a metabolite. A gene query displays expression across the eight cultivars and retrieves metabolite contrasts from the same cultivar panel. A metabolite query returns pathway annotation and the expression profiles of associated genes. This bidirectional route supports pathway-level candidate prioritization while retaining the cultivar context of both evidence types.

3.7.4. Use Case 3: Trait-Specific Combining-Ability Summary

The parental-prediction use case evaluated seven breeding traits from 122 progeny assigned to seven parental combinations. After aggregation, each trait was represented by seven combination means. Six traits passed the normality diagnostic, whereas trichome density did not ($p = 0.0338$). Table 4 summarizes the combination-level phenotypes used by the module.

Within the seven observed combinations, the maximum SCA was associated with HJY $\times$ YJX for tree posture, HJY $\times$ SMXY for bud initiation and PYTZ $\times$ SMXY for bud-leaf color, trichome density and calyx color. HJJ $\times$ ZJ had the maximum SCA for leaf morphology, whereas YJX $\times$ ZJ had the maximum for style-splitting position (Figure 6; Table 5). These values demonstrate combining-ability reporting. A high SCA denotes positive deviation from the expectation on the encoded trait scale; it does not by itself identify a favorable or field-superior cross.

Table 4. Combination-level phenotype summary used in the pilot parental-prediction workflow. Seven combination means were derived from 122 progeny records; valid individual-level counts for Traits 1–7 were 122, 122, 122, 122, 118, 104 and 100, respectively.

Trait	Cross N	Mean	SD	CV	Normality
Tree posture	7	3.5	1.24	0.353	normal
One-bud-and-one-leaf initiation stage	7	3.15	1.16	0.37	normal
Color of one bud and two leaves	7	3.17	1.63	0.515	normal
Trichome density of young shoots and bud leaves	7	1.99	0.187	0.094	non_normal
Leaf morphology	7	1.76	0.463	0.262	normal
Calyx color	7	0.195	0.21	1.08	normal
Style-splitting position	7	2.92	0.701	0.24	normal

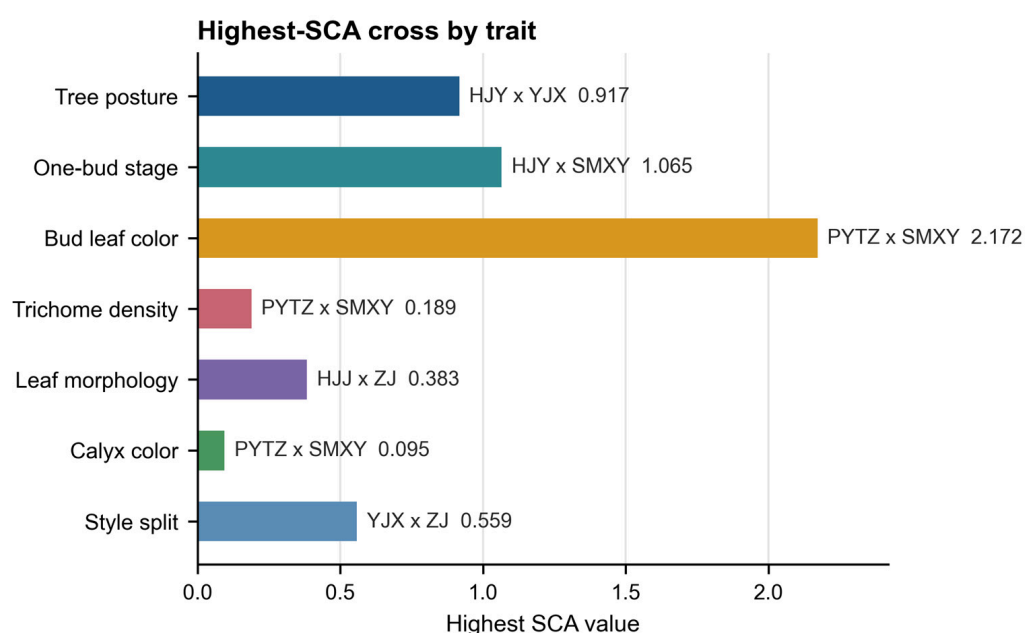


Figure 6. Observed parental combination with the highest SCA for each trait. Values were calculated from seven combination means derived from 122 progeny records. SCA indicates deviation from the expectation based on the two parental GCAs and the grand mean; it should not be interpreted as a validated recommendation without considering the desired direction of the encoded trait and independent field evaluation.

Table 5. Observed parental combination with the highest SCA for each trait. The parental GCA sum is $GCA(i) + GCA(j)$; SCA is the deviation of the observed combination mean from the expectation based on the grand mean and the two parental GCAs.

Trait	Observed Cross	Cross Mean	Parental GCA Sum	SCA
Tree posture	HJY × YJX	5.000	0.583	0.917
One-bud-and-one-leaf initiation stage	HJY × SMXY	4.000	−0.215	1.065
Color of one bud and two leaves	PYTZ × SMXY	1.000	−4.344	2.172
Trichome density of young shoots and bud leaves	PYTZ × SMXY	2.000	−0.178	0.189
Leaf morphology	HJJ × ZJ	1.250	−0.896	0.383
Calyx color	PYTZ × SMXY	0.000	−0.290	0.095
Style-splitting position	YJX × ZJ	3.477	−0.001	0.559

The interface also reports the remaining observed combinations. Supplementary Table S1 lists the three highest-SCA combinations for each trait as descriptive outputs, not validated breeding recommendations.

4. Discussion

4.1. Cultivar-Centered Resource Design

NB-TeaBase contributes a cultivar-centered organization of breeding and multi-omics data. Unlike gene- or genome-first resources, it retains the named parent as the common unit linking identity, molecular records and prediction output. This organization is particularly relevant to perennial breeding, where each proposed cross carries substantial time and field costs.

4.2. Complementarity with TPIA2 and Other Tea Resources

The platform is complementary to broader tea resources. TPIA2 provides extensive *Camellia*-wide coverage [4], and TeaGVD, TeaPVs, TeaCoN, TeaAS and TeaPGDB provide specialized variation, transcript or genome functions [5–9]. NB-TeaBase contributes a narrower but operationally linked panel in which passports, omics evidence and cross rankings refer to the same breeding materials.

4.3. Strengths and Limitations

The current release has two principal limitations. First, the parental-prediction dataset contains 122 progeny but only seven independent parental combinations, with 2–55 progeny per combination. These data support estimation of observed combination means and descriptive GCA/SCA but not robust evaluation of GEBV performance beyond the represented families. Second, trichome density failed the normality diagnostic and was analyzed without transformation. The corresponding combining-ability and prediction outputs should therefore be treated as exploratory.

The multi-omics links also define an evidential boundary. Concordant variants, expression profiles and metabolites can strengthen candidate prioritization, but they do not establish gene function. Functional claims require independent genetic or experimental validation.

Prediction reliability should be assessed by the number and balance of independent parental combinations, family-blocked cross-validation and an external validation population, rather than by progeny count alone. No universal minimum sample size applies across traits and populations [23–25,27].

4.4. Future Development

Future releases will prioritize additional parental combinations, balanced family sizes and multi-year or multi-location phenotyping. Distribution-appropriate models will be evaluated for non-normal or ordinal traits, and prediction performance will be assessed with family-blocked and independent validation. The database will also expand curated omics links and user documentation.

5. Conclusions

NB-TeaBase links cultivar passports, genomic variation, gene expression, metabolite profiles and progeny-informed prediction and combining-ability summaries within one robust-bud tea breeding resource. By organizing germplasm, genome, transcriptome, metabolome and breeding records around named cultivars, the platform provides a traceable evidence base for cultivar evaluation, parent selection, cross prioritization and breeding strategy formulation. Open browsing and downloadable records further support reuse. Its current prediction outputs are limited by seven observed parental combinations and require independent field validation. NB-TeaBase therefore provides an extensible foundation for data-informed breeding decisions rather than a substitute for multi-environment breeding trials.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/horticulturae12070802/s1>, Table S1: Three observed parental combinations with the highest SCA for each trait.

Author Contributions: Conceptualization, Y.C. and J.Z.; methodology, Y.C., M.L. and D.S.; software, Y.C.; validation, L.W., L.Z., K.C., D.S. and B.W.; investigation, Y.C., L.W., L.Z., K.C. and B.W.; resources, L.W., L.Z., K.C., B.W. and J.Z.; data curation, M.L. and Y.C.; writing-original draft preparation, Y.C. and M.L.; writing-review and editing, D.S., B.W. and J.Z.; supervision, J.Z.; project administration, Y.C. and J.Z.; funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the 2024 Ningbo Science and Technology Development Special Fund, Social Welfare Research Program (No. 2024S112) and the President Fund of Ningbo Academy of Agricultural Sciences (No. 2024NKYP002, 2025NKYP007).

Data Availability Statement: NB-TeaBase is available at <http://112.126.28.109:8082> (accessed on 26 June 2026). Public browsing does not require registration. Processed data are available through the database modules and file-download services. The raw resequencing and transcriptome data have been deposited in the NCBI BioProject database under accession number PRJNA1466621 (<https://www.ncbi.nlm.nih.gov/bioproject/>) (accessed on 26 June 2026). The metabolome data are available in the Figshare repository at <https://doi.org/10.6084/m9.figshare.32252196> (accessed on 26 June 2026).

Acknowledgments: During the preparation of this manuscript, the authors used OpenAI Chat-GPT/Codex (GPT-based generative AI tool, accessed in 2026) for language polishing, structural editing, formatting refinement and assistance in preparing schematic figure layouts based on author-provided data and descriptions. The authors reviewed and edited all output and take full responsibility for the content of this publication.

Conflicts of Interest: Author Longjie Zhang was employed by Ningbo Huangjinyun Tea Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

NB, Ningbo; FDDH, Fudingdahaocha; FDDB, Fudingdabaicha; GEBV, genomic estimated breeding value; GCA, general combining ability; SCA, specific combining ability; RNA-seq, RNA sequencing; LC-MS/MS, liquid chromatography-tandem mass spectrometry; FPKM, fragments per kilobase of transcript per million mapped reads; PCA, principal component analysis; QC, quality control; VIP, variable importance in projection.

References

1. Xia, E.H.; Zhang, H.B.; Sheng, J.; Li, K.; Zhang, Q.J.; Kim, C.; Zhang, Y.; Liu, Y.; Zhu, T.; Li, W.; et al. The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. *Mol. Plant* **2017**, *10*, 866–877. [[CrossRef](#)] [[PubMed](#)]
2. Zhang, X.; Chen, S.; Shi, L.; Gong, D.; Zhang, S.; Zhao, Q.; Zhan, D.; Vasseur, L.; Wang, Y.; Yu, J.; et al. Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. *Nat. Genet.* **2021**, *53*, 1250–1259. [[CrossRef](#)] [[PubMed](#)]
3. Xia, E.-H.; Li, F.-D.; Tong, W.; Li, P.-H.; Wu, Q.; Zhao, H.-J.; Ge, R.-H.; Li, R.-P.; Li, Y.-Y.; Zhang, Z.-Z.; et al. Tea Plant Information Archive: A comprehensive genomics and bioinformatics platform for tea plant. *Plant Biotechnol. J.* **2019**, *17*, 1938–1953. [[CrossRef](#)] [[PubMed](#)]
4. Gao, Q.; Tong, W.; Li, F.; Wang, Y.; Wu, Q.; Wan, X.; Xia, E. TPIA2: An updated tea plant information archive for *Camellia* genomics. *Nucleic Acids Res.* **2024**, *52*, D1661–D1667. [[CrossRef](#)] [[PubMed](#)]
5. Chen, J.-D.; He, W.-Z.; Chen, S.; Chen, Q.-Y.; Ma, J.-Q.; Jin, J.-Q.; Ma, C.-L.; Moon, D.-G.; Ercisli, S.; Yao, M.-Z.; et al. TeaGVD: A comprehensive database of genomic variations for uncovering the genetic architecture of metabolic traits in tea plants. *Front. Plant Sci.* **2022**, *13*, 1056891. [[CrossRef](#)] [[PubMed](#)]
6. An, Y.; Zhang, X.; Jiang, S.; Zhao, J.; Zhang, F. TeaPVs: A comprehensive genomic variation database for tea plant (*Camellia sinensis*). *BMC Plant Biol.* **2022**, *22*, 513. [[CrossRef](#)] [[PubMed](#)]

7. Zhang, R.; Ma, Y.; Hu, X.; Chen, Y.; He, X.; Wang, P.; Chen, Q.; Ho, C.-T.; Wan, X.; Zhang, Y.; et al. TeaCoN: A database of gene co-expression network for tea plant (*Camellia sinensis*). *BMC Genom.* **2020**, *21*, 461. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Mi, X.; Yue, Y.; Tang, M.; An, Y.; Xie, H.; Qiao, D.; Ma, Z.; Liu, S.; Wei, C. TeaAS: A comprehensive database for alternative splicing in tea plants (*Camellia sinensis*). *BMC Plant Biol.* **2021**, *21*, 280. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Lei, X.; Wang, Y.; Zhou, Y.; Chen, Y.; Chen, H.; Zou, Z.; Zhou, L.; Ma, Y.; Chen, F.; Fang, W. TeaPGDB: Tea Plant Genome Database. *Beverage Plant Res.* **2021**, *1*, 5. [\[CrossRef\]](#)
10. Li, H.; Li, X.; Zhang, P.; Feng, Y.; Mi, J.; Gao, S.; Sheng, L.; Ali, M.; Yang, Z.; Li, L.; et al. Smart Breeding Platform: A web-based tool for high-throughput population genetics, phenomics, and genomic selection. *Mol. Plant* **2024**, *17*, 677–681. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Xu, Y.; Zhang, X.; Li, H.; Zheng, H.; Zhang, J.; Olsen, M.S.; Varshney, R.K.; Prasanna, B.M.; Qian, Q. Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction. *Mol. Plant* **2022**, *15*, 1664–1695. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018**, *34*, i884–i890. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **2013**, arXiv:1303.3997.
14. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernysky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation sequencing data. *Genome Res.* **2010**, *20*, 1297–1303. [\[CrossRef\]](#) [\[PubMed\]](#)
15. 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. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Liao, Y.; Smyth, G.K.; Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **2014**, *30*, 923–930. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Pertea, M.; Pertea, G.M.; Antonescu, C.M.; Chang, T.-C.; Mendell, J.T.; Salzberg, S.L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **2015**, *33*, 290–295. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Chambers, M.C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D.L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012**, *30*, 918–920. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Smith, C.A.; Want, E.J.; O’Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **2006**, *78*, 779–787. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Wishart, D.S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B.L.; et al. HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Res.* **2022**, *50*, D622–D631. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E.A.; Glass, C.K.; Merrill, A.H.; Murphy, R.C.; Raetz, C.R.H.; Russell, D.W.; et al. LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* **2007**, *35*, D527–D532. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Meuwissen, T.H.E.; Hayes, B.J.; Goddard, M.E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **2001**, *157*, 1819–1829. [\[CrossRef\]](#) [\[PubMed\]](#)
24. VanRaden, P.M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **2008**, *91*, 4414–4423. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Endelman, J.B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* **2011**, *4*, 250–255. [\[CrossRef\]](#)
26. Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27. [\[CrossRef\]](#)
27. Crossa, J.; Perez-Rodriguez, P.; Cuevas, J.; Montesinos-Lopez, O.; Jarquin, D.; de los Campos, G.; Burgueno, J.; Gonzalez-Camacho, J.M.; Perez-Elizalde, S.; Beyene, Y.; et al. Genomic selection in plant breeding: Methods, models, and perspectives. *Trends Plant Sci.* **2017**, *22*, 961–975. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Buels, R.; Yao, E.; Diesh, C.M.; Hayes, R.D.; Munoz-Torres, M.; Helt, G.; Goodstein, D.M.; Elsik, C.G.; Lewis, S.E.; Stein, L.; et al. JBrowse: A dynamic web platform for genome visualization and analysis. *Genome Biol.* **2016**, *17*, 66. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [\[CrossRef\]](#) [\[PubMed\]](#)
30. 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*, e115. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; Bonino da Silva Santos, L.O.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.