the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2021) **108,** 1174–1192 doi: 10.1111/tpj.15482

RESOURCE

Chromosome-scale genome assembly and population genomics provide insights into the adaptation, domestication, and flavonoid metabolism of Chinese plum

Zhenyu Huang[1,†], Fei Shen[2,†], Yuling Chen[1], Ke Cao[1,*] and Lirong Wang[1,*]

[1]*Zhengzhou Fruit Research Institute, Chinese Academy of Agricultural Science, Zhengzhou, Henan 450009, China, and*

[2]*Beijing Agro-Biotechnology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China*

Received 8 October 2020; revised 27 August 2021; accepted 30 August 2021; published online 25 September 2021.

*For correspondence (e-mail wanglirong@caas.cn [Lirong Wang]; wyandck@126.com [Ke Cao]).

†These authors contributed equally to this work.

SUMMARY

Globally, commercialized plum cultivars are mostly diploid Chinese plums (*Prunus salicina* Lindl.), also known as Japanese plums, and are one of the most abundant and variable fruit tree species. To advance *Prunus* genomic research, we present a chromosome-scale *P. salicina* genome assembly, constructed using an integrated strategy that combines Illumina, Oxford Nanopore, and high-throughput chromosome conformation capture (Hi-C) sequencing. The high-quality genome assembly consists of a 318.6-Mb sequence (contig N50 length of 2.3 Mb) with eight pseudo-chromosomes. The expansion of the *P. salicina* genome is led by recent segmental duplications and a long terminal repeat burst of approximately 0.2 Mya. This resulted in a significant expansion of gene families associated with flavonoid metabolism and plant resistance, which impacted fruit flavor and increased species adaptability. Population structure and domestication history suggest that Chinese plum may have originated from South China and provides a domestication route with accompanying genomic variations. Selection sweep and genetic diversity analysis enabled the identification of several critical genes associated with flowering time, stress tolerance, and flavonoid metabolism, demonstrating the essential roles of related pathways during domestication. Furthermore, we reconstructed and exploited flavonoid–anthocyanin metabolism using multi-omics analysis in Chinese plum and proposed a complete metabolic pathway. Collectively, our results will facilitate further candidate gene discovery for important agronomic traits in Chinese plum and provide insights into future functional genomic studies and DNA-informed breeding.

Keywords: *Prunus salicina* L., chromosome-level reference genome, genome evolution, genetic diversity, gene identification.

INTRODUCTION

Plum is one of the most important horticultural crops worldwide, with an extensive genetic diversity and high economic value, and is an important source of vitamins, minerals, fiber, and antioxidants. Currently, the most commercially produced plums are the Chinese plum (*Prunus salicina* L., $2n = 2x = 16$) and European plum (*Prunus domestica* L., $2n = 6x = 48$) (Topp et al., 2012). China is currently the largest plum producer, with an annual production of 6 801 187 metric tons in 2018, accounting for 53.9% of the world's total (FAOSTAT, 2019).

Belonging to the *Prunus* genus, the Chinese plum accounts for most of the fresh market plums supplied globally and is one of the most abundant and variable groups of fruit tree species (Farcuh et al., 2018). The discovery of fossil plum stones dated to the Neolithic or Warring States period implies that plum fruits have been consumed for over 5000 years in China (Zhang and Zhou, 1998). The Chinese plum has a long growing history and extensive geographical distribution, with more than 1000 indigenous plum cultivars derived from *P. salicina* (Zhang, 1990). It was speculated from historical records that the Chinese plum originated in the middle reaches of the Yangtze River Basin in South China (Hartmann and Neumulle, 2009). A

palynological study showed that the *P. salicina* in South China is more primitive than that in North China (Guo, 2006). The genetic diversity of Chinese plums cultivated in Southwest China was higher than that of Chinese plums in Northern China using a large-scale phenotypic variation analysis (Yu et al., 2011) and molecular tests with multiple types of markers (Chen and Yang, 2014; Wei et al., 2019, 2021; Zuo et al., 2015), which was in accordance with the fact that there were wild plum populations discovered in Yunnan, Sichuan, and Guizhou Provinces (Wei et al., 2020; Zhang and Zhou, 1998). It is speculated that plums were introduced to Japan by a Chinese monk of high standing who brought the trees as a gift to the Emperor; alternatively, a naturalized Japanese, Wani, might also have brought plums to Japan from the Korean peninsula (Faust and Suranyi, 1999). *Prunus salicina*, also called Japanese plum, was first imported to the United States in 1870 from Japan by Luther Burbank, who went on to hybridize the species with American native species, resulting in the development of more than 100 varieties, which laid the foundation for fresh market plum production worldwide (Hartmann and Neumulle, 2009).

Generally, the Chinese plum type includes both pure Chinese plums (*P. salicina*) and its hybrids with other diploid plum species, such as *Prunus simonii*, *Prunus cerasifera*, and *Prunus americana*, and over 400 accessions are preserved in the Chinese National Germplasm Repository for Plums and Apricots (Liu et al., 2007). To boost the efficiency of conservation and utilize the biodiversity in Chinese plum germplasm collections, several studies related to genetic diversity evaluation, population structure analysis, and genetic characterization of Chinese plums have been conducted using random amplified polymorphic DNA and simple sequence repeat (SSR) markers (Liu et al., 2006, 2007). Furthermore, reduced-representation genome sequencing techniques have also been widely adopted as promising methods for crops without available reference genomic information (Carrasco et al., 2018; Marti et al., 2018; Salazar et al., 2017, 2019).

With markedly improved sequencing technologies, several important stone fruit species closely related to Chinese plum, including Japanese apricot (*Prunus mume*, $2n = 2x = 16$) (Zhang et al., 2012), peach (*Prunus persica*, $2n = 2x = 16$) (Verde et al., 2017), sweet cherry (*Prunus avium*, $2n = 2x = 16$) (Pinosio et al., 2020; Shirasawa et al., 2017), almond (*Prunus dulcis*, $2n = 2x = 16$) (Alioto et al., 2019; Sánchez-Pérez et al., 2019), and apricot (*Prunus armeniaca*, $2n = 2\times = 16$) (Jiang et al., 2019), have already been sequenced, which greatly contributes to our understanding of the structure of *Prunus* genomes in general and the rapid evolution of their genera. More recently, for the subgenus *Prunophora*, the genome architectures of a hexaploid European plum ('Improved French') and a diploid Chinese plum ('Sanyueli') have been dissected (Callahan et al., 2021; Fang et al., 2020; Liu et al., 2020b). To some extent, these newly released plum genome assemblies will advance comparative and evolutionary genomic studies; however, more continuous and complete genome sequences are required to accelerate scientific research.

The present study uses an integrated strategy combining Illumina, Oxford Nanopore, and high-throughput chromosome conformation capture (Hi-C) sequencing to obtain a chromosome-scale genome assembly of another diploid Chinese plum cultivar, 'Zhongli No.6', with an improved quality compared to those mentioned above. Furthermore, we analyze the genomic characteristics of 78 *Prunus* accessions using high-depth whole-genome re-sequencing data. Taken together, these results will facilitate downstream gene discovery, trait mapping for breeding, and other functional genomic applications for Chinese plum.

RESULTS

High-quality reference genome of *P. salicina*

The *P. salicina* cultivar sequenced was 'Zhongli No.6', selected by the Zhengzhou Fruit Research Institute, Chinese Academy of Agricultural Sciences (ZFRI-CAAS), and its features include early ripening, large fruit size, and high soluble solids content and productivity (Figure 1a). A set of approximately 79-fold coverage Illumina paired-end short reads (24.0 Gb), approximately 204-fold coverage Nanopore long reads (62.2 Gb), and approximately 150-fold coverage Hi-C data (45.7 Gb) was generated (Tables S2–S4). By conducting k-mer analysis, the *P. salicina* genome size and the genome heterozygosity rate were estimated to be 305.6 Mb and 0.92%, respectively (Figures S1 and S2). We initially assembled the reads into 335.6 Mb of contig sequences (Table S6), and the redundant sequences from the heterozygous genomic regions were then identified and filtered. This yielded a genome assembly of 318.6 Mb, with a contig N50 size of 2.3 Mb and a maximum contig size of 21.5 Mb (Figure S3, Table S7). Subsequently, using the Hi-C data, the contigs were corrected and scaffolded into eight pseudo-chromosomes that anchored 90.98% of the assembled sequence (Figure 1b,c; Table S5). Benchmarking Universal Single-Copy Orthologs (BUSCO) provides a quantitative assessment of genome assembly completeness in terms of expected gene content, while a higher percentage of complete terms suggests high integrity. The BUSCO analysis against the plant-specific database revealed 1348 (98.04%) complete gene models (Figure 1d; Table S8). The extent of comprehensive gene coverage was assessed by screening for the 248 most highly conserved core eukaryotic genes (CEGs), which revealed complete and partial matches for 233 (93.95%) and 243 (97.98%) genes, respectively (Figure S4, Table S9). Moreover, an average of 96.22% of the RNA sequencing

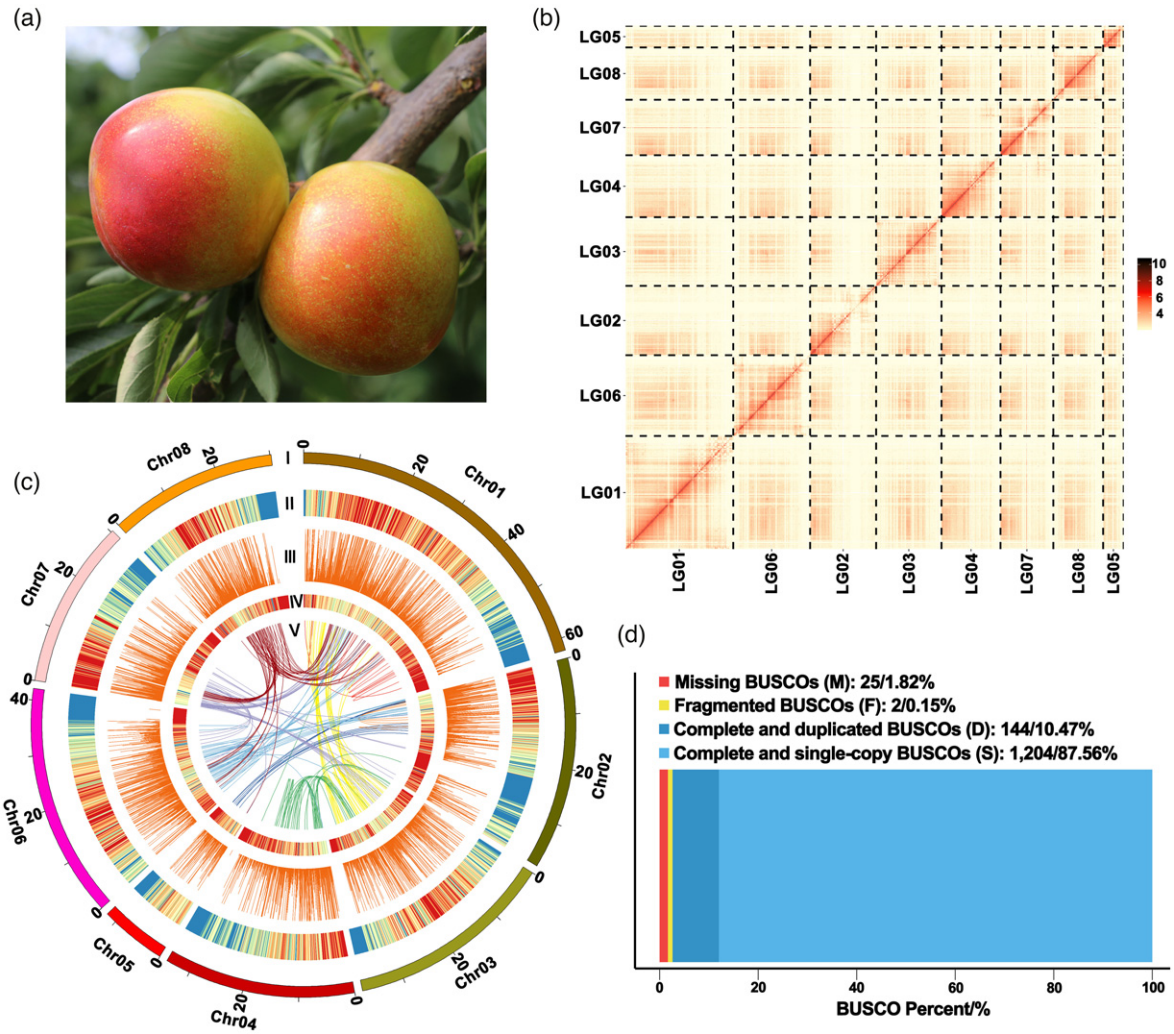

Figure 1. The genome features of *Prunus salicina*. (a) A fruit image of the sequenced Chinese plum cultivar 'Zhongli No.6'. (b) Genome-wide high-throughput chromosome conformation capture (Hi-C) interaction matrices of the ordered scaffolds along the eight pseudo-chromosomes. (c) Circular representation of the eight pseudo-chromosomes. Tracks from outside to the inner correspond to I, pseudo-chromosomes; II, gene density; III, repeat density; IV, GC content; and V, relationship between syntenic blocks. (d) Completeness evaluation of the Chinese plum genome using Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment. The numbers/percentages of different types of BUSCOs searched in the Chinese plum genome are indicated with different colors; a higher percentage of complete terms reveals a better assembly integrity.

(RNA-seq) reads of the four *P. salicina* tissues (flower, leaf, fruit, and stem bark) could be mapped to the genome (Table S12). In addition, when compared with two recently released genome assemblies for the Chinese plum cultivar 'Sanyueli' originating from Southern China (Fang et al., 2020; Liu et al., 2020b), the quality of the genome assembly for the Chinese plum cultivar 'Zhongli No.6', from Northern China, was improved in terms of sequence continuity and integrity, using the abovementioned evaluation approaches (Table S13). Collectively, this illustrates the high quality of the *P. salicina* genome assembly, indicating that it is adequate for subsequent genomic analysis.

A total of 27 481 genes were predicted, 95.46% of which could be functionally annotated based on the currently available databases, with an average gene length of 3500 bp, an average coding sequence length of 1208 bp, and an average of 5.36 exons (Tables S16 and S17). The annotation of the non-coding RNA genes yielded 422 rRNA, 3504 snRNA, 78 miRNA, 1502 tRNA, 4 regulatory RNA, 103 spliceosomal RNA, and 35 other types of genes in the *P. salicina* genome (Table S18). In addition, we annotated 166 495 SSRs (Table S14). These results provide a valuable genetic resource for future functional genomics and molecular breeding research.

Genome expansion of *P. salicina* caused by a long terminal repeat burst and large-scale segmental duplication events

The size of the *P. salicina* genome is 318.6 Mb, larger than that of its neighboring taxa (approximately 200 Mb)

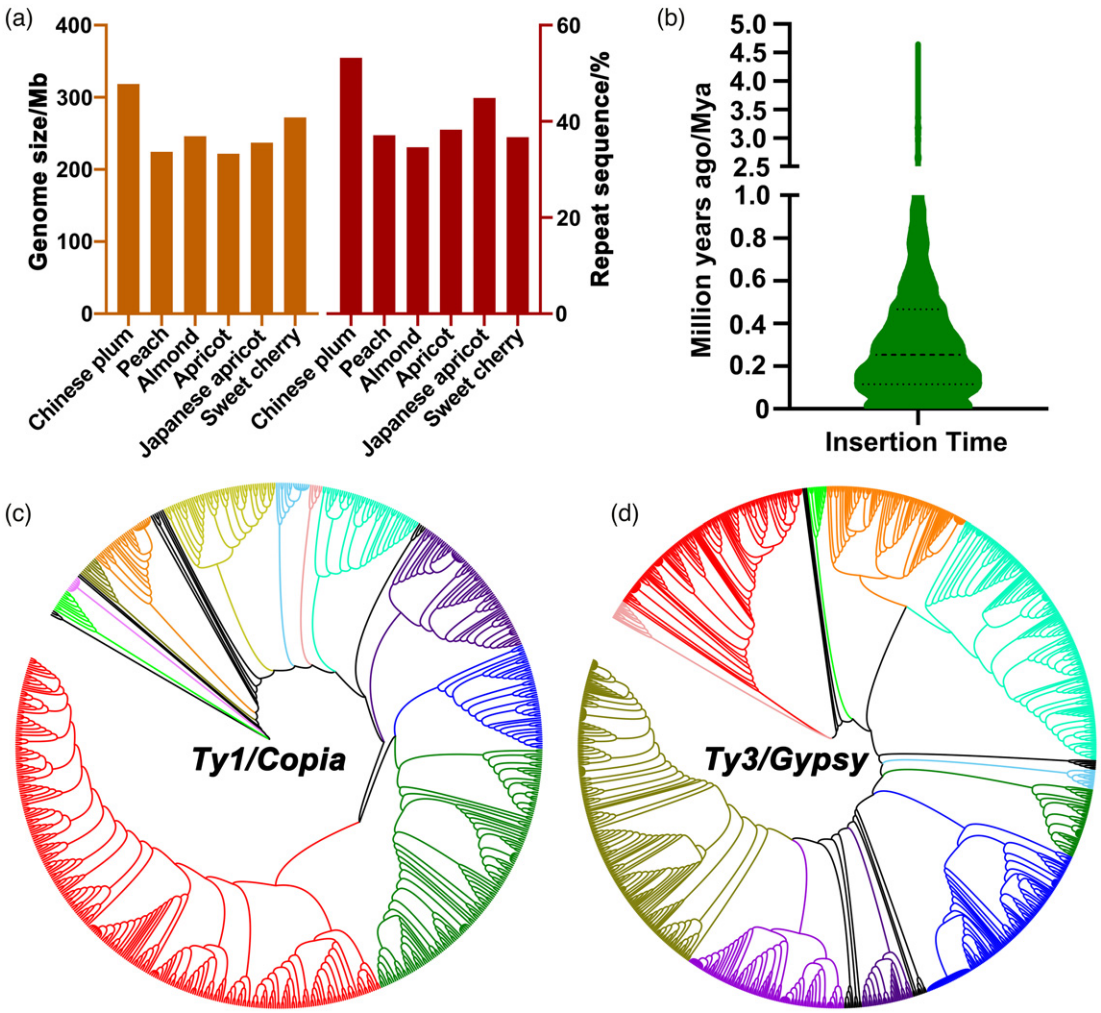

Figure 2. Repeat sequence analysis of the *Prunus salicina* genome. (a) Genome size (in orange) and repeat sequence content (in red) comparisons of Chinese plum and five other closely related stone fruit species. (b) Insertion time of the long terminal repeat (LTR) retrotransposons in the Chinese plum genome. (c, d) The phylogenetic tree of (c) *Ty1/Copia* type and (d) *Ty3/Gypsy* type LTR retrotransposons.

(Figure 2a). To explore the composition of the *P. salicina* genome, we conducted a detailed genome annotation. Altogether, 53.39% of the genome was repetitive (Figure 2a; Table S15). Among the predicted repeats, the long terminal repeats (LTRs) comprised the largest proportion of the genome (39.80%), including 56.06% LTR/*Gypsy* and 9.78% LTR/*Copia* retroelements; the DNA class repeat elements ranked second, accounting for 8.54% of the genome (Table S15). The phylogenetic trees of LTR/*Gypsy* and LTR/*Copia* showed that the repeat elements were grouped into different clades and expanded by clusters (Figure 2c,d). LTRs are the most abundant DNA components in all investigated plant species and are largely responsible for variations in plant genome size variations (Liu et al., 2020a). When compared with the neighboring taxa (almond, 34.60%; Japanese apricot, 44.92%; sweet cherry, 36.70%; peach, 37.14%; apricot, 38.28%), the repeat sequence content observed in Chinese plum was higher, especially for the LTRs (Figure 2a). Moreover, we detected a recent LTR burst of approximately 0.2 Mya, which involved several LTR families (Figure 2b).

We further investigated whole-genome duplication (WGD) events during the genome evolution of *P. salicina*, and identified 4905 reciprocal best-hit (RBH) paralogous gene pairs in the *P. salicina* genome assembly. Furthermore, we used synonymous substitutions per synonymous site (*Ks*) of paralogous gene pairs to measure the age distribution of duplication events. Interestingly, we identified two prominent peaks (PS_P1 and PS_P2) in the *Ks* profiles of the *P. salicina* genome, which were similar to those of apricot (Figure 3a; Figure S5). PS_P1 occurred at a *Ks* value of approximately 1.8, suggesting a common WGD event in the genus *Prunus*. As for PS_P2, with a lower *Ks* value of approximately 0.2 (just near the Chinese plum–apricot

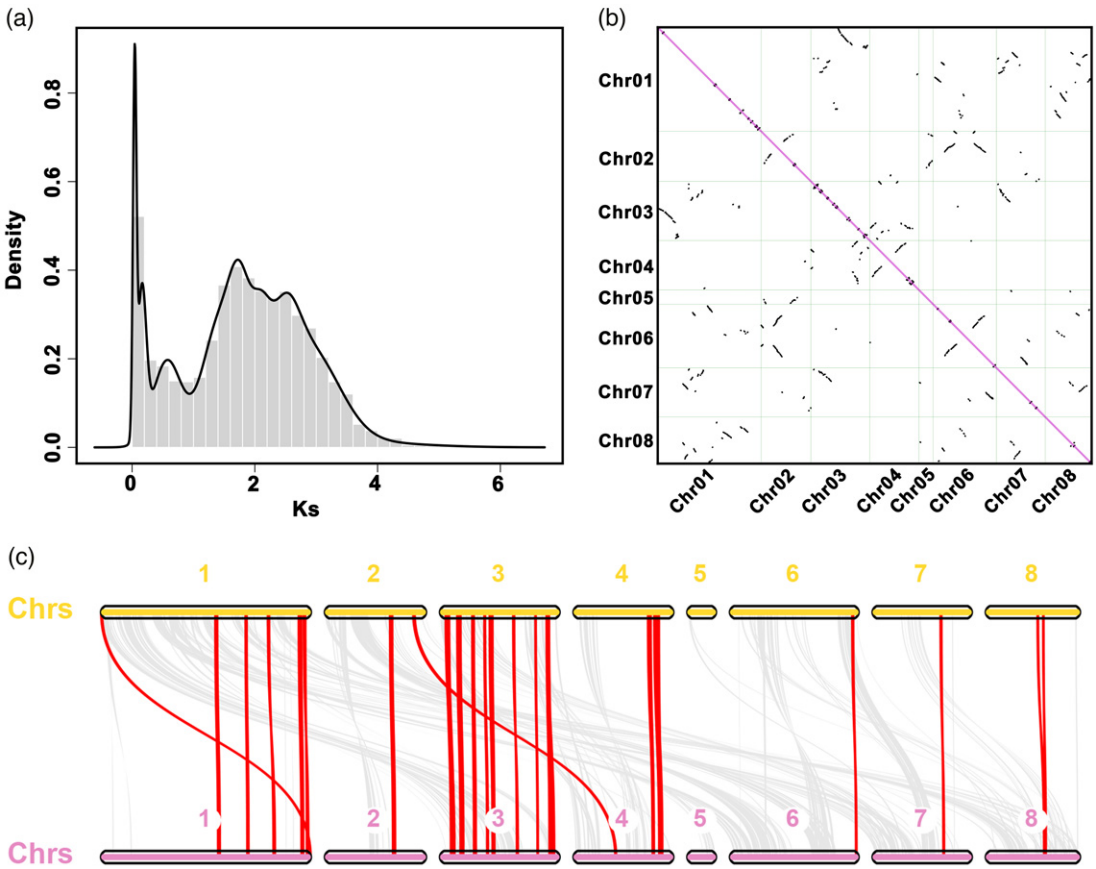

Figure 3. Genome duplication events and paralogous gene identification. (a) Synonymous *Ks* distribution of paralog genes in the *Prunus salicina* genome. (b) Dot plot matrix displaying the paralogs in *P. salicina*. (c) Macrosynteny pattern of the *P. salicina* karyotype. Gray lines highlight the syntenic blocks spanning the genome. Red lines highlight the major segmental duplications within and between chromosomes.

speciation curve), it could be a species-specific segmental duplication (SD) or a shared SD with apricot (Figure S6). We analyzed the *Ks* value distribution of the paralogous gene pairs in the collinearity region and identified two distinct *Ks* profiles within different segments, suggesting the existence of another large-scale SD event in Chinese plum besides the shared WGD across the *Prunus* genus (Figure 3b,c; Table S19). Large-scale genome duplications produce abundant duplicated genes related to the function of secondary metabolism and stress tolerance, which are important for the diversity of gene functions and adaptations to changing environments (Table S19).

In summary, the expansion of the *P. salicina* genome was led by the recent LTR burst and large-scale SDs; simultaneously, the LTR burst and SD events are fundamental causes of species differentiation.

Adaptability and secondary metabolism were reinforced during genome evolution

We elucidated the evolutionary location of Chinese plum by constructing a phylogenetic tree of *P. salicina* and related species based on 224 high-quality single-copy orthologous genes (Figure 4d) and estimating the divergence time among branches of the tree (Figure 4c). The phylogenetic tree indicated that, among the *Prunus* species, *P. salicina* was most closely related to *P. mume* and *P. armeniaca*, with a divergence time of approximately 16.73 Mya. Gene family analysis revealed that during the evolution of *P. salicina*, a total of 2237 gene families expanded, while 3612 families contracted, and we identified a total of 241 gene families comprising 550 genes specific to the *P. salicina* lineage (Figure 4c; Figures S8 and S9, Tables S20–S22). The expanded gene families were primarily enriched in sesquiterpenoid and triterpenoid biosynthesis, stilbenoid, diarylheptanoid, and gingerol biosynthesis, isoquinoline alkaloid biosynthesis, linoleic acid metabolism, and flavonoid biosynthesis (Figure 4a; Tables S23 and S24).

The genes in the PS_P2 peak of the *Ks* profiles were investigated, and the genes involved in secondary metabolism (e.g., flavonoid biosynthesis and triterpenoid biosynthesis) and disease resistance were also identified (Table S19). Interestingly, the overexpressed genes were always distributed in the later large-segment genomic

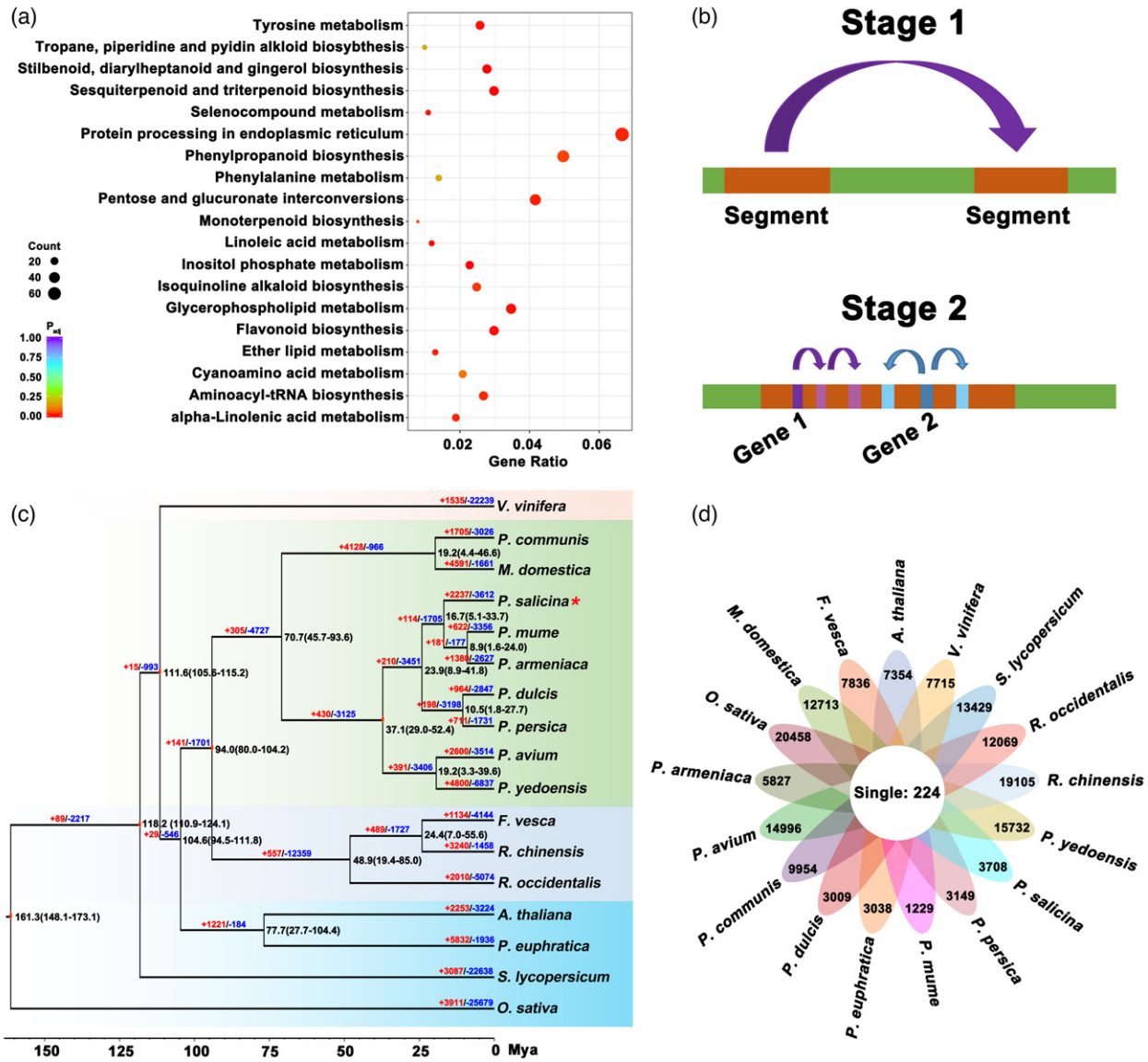

Figure 4. Gene family and divergence analysis of *P. salicina* and 16 related species. (a) Statistics of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment of the expanded gene families. (b) A model of the two-stage expansion of gene families. Stage 1, large-scale segment duplications; Stage 2, tandem duplications. (c) Phylogenetic tree of the 17 species, and gene family expansion and contraction compared with the most recent common ancestor. Gene family expansions are indicated in red. Gene family contractions are indicated in blue. Inferred divergence times (million years ago) are denoted at each node in black. (d) Venn diagram showing the shared and unique gene families among the Chinese plum genome and 16 other plant species. The number in the center represents the single-copy orthologous gene number. The numbers outside represent the unique gene numbers of corresponding species.

region and expanded by tandem duplications, such as genes encoding vinorine synthase, stemmadenine *O*-acetyltransferase (SAT), and shikimate *O*-hydroxycinnamoyltransferase (HCT). Vinorine synthase catalyzes the formation of vinorine, a precursor of the antiarrhythmic monoterpenoid indole alkaloid (MIA) ajmaline (Bayer et al., 2004). SAT is a component of the iboga and aspidosperma MIAs (e.g., tabersonine and catharanthine) biosynthesis pathways from 19*E*-geissoschizine and catalyzes the formation of *O*-acetylstemmadenine from stemmadenine (Qu et al., 2018). *HCT* genes participate in phenylpropanoid biosynthesis and affect the content of various materials, including flavonoids, monolignols, phenolic acids, stilbenes, and coumarins (Hoffmann et al., 2004). We also identified several stress-related and development-associated genes expanded significantly during genome evolution, such as disease resistance protein (Martin et al., 2003), (−)-alpha-pinene synthase (Phillips et al., 2003), and cytochrome P450 (Bak and Feyereisen, 2001).

Based on these results, we propose that several gene families/gene copies associated with stress tolerance and secondary metabolism have been expanded in two stages:

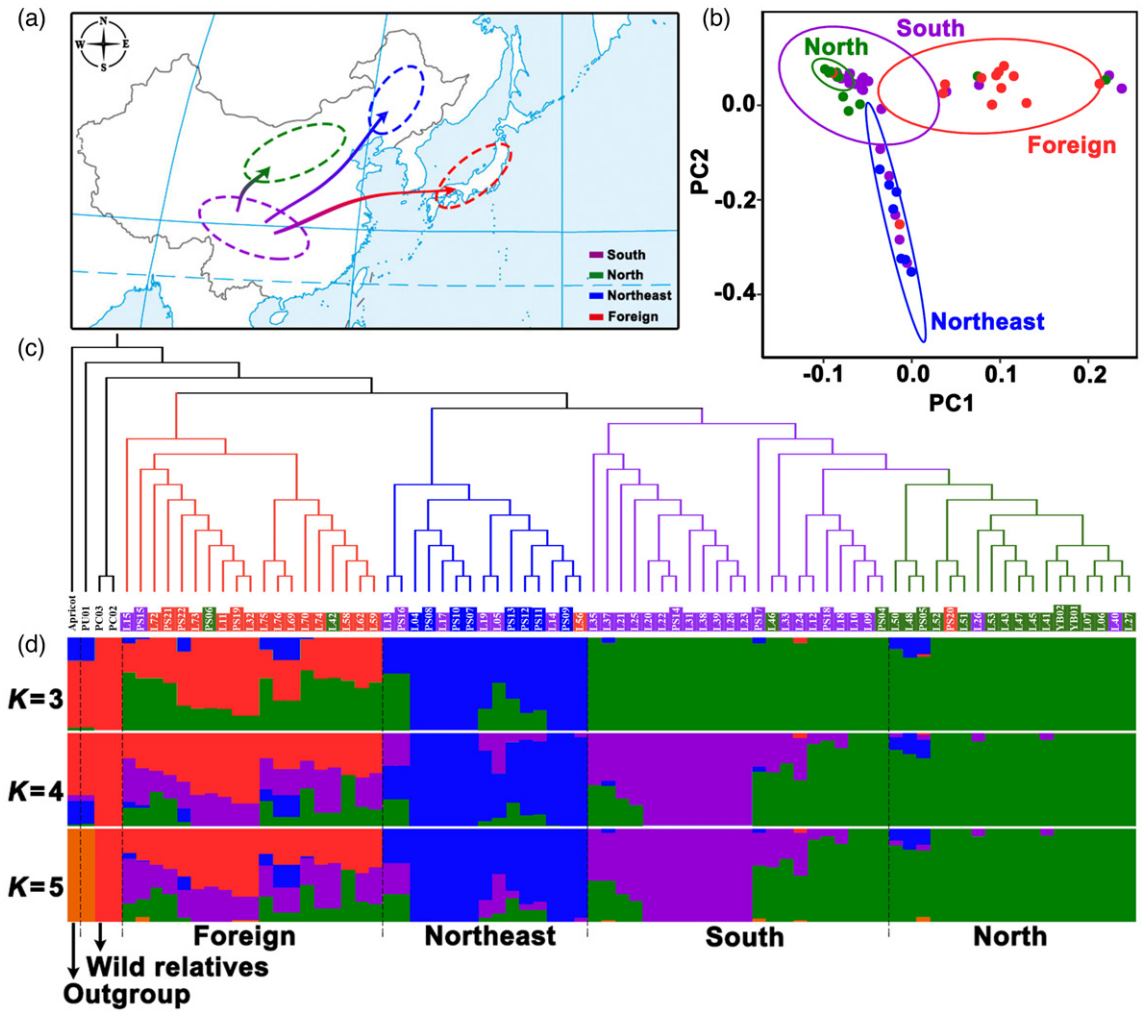

Figure 5. Population structure analysis of the *Prunus* accessions. (a) The deduced domestication route of Chinese plums from South China to other regions. The purple, green, blue, and red ellipses represent the South China group, the North China group, the Northeast China group, and the Foreign group, respectively. (b) Principal component analysis, (c) phylogenetic tree, (d) and population structure of the 78 *Prunus* accessions. The numbers of Chinese plum accessions within the South China group, the North China group, the Northeast China group, and the Foreign group are 22, 18, 15, and 19, respectively.

(i) large-scale SDs and (ii) tandem duplications, which reinforced the adaptability of *P. salicina* (Figure 4b).

Population structure of *Prunus* accessions

We examined the relationships and divergence among the different ecological populations of the 78 *Prunus* accessions (Table S1). Using the high-depth (approximately 20×) whole-genome re-sequencing data of these accessions generated in our study (22 of the 78) and reported by others (56 of the 78) (Jiang et al., 2019; Wei et al., 2021), we identified 13 548 616 single nucleotide polymorphisms (SNPs) and 1 090 237 small insertion/deletions (InDels). After imputation and filtering, we obtained 3 380 659 high-quality SNPs for further analysis; the average SNP frequency was nearly one per 100 bp (Figure S13).

Using apricot as an outgroup, we explored the phylogenetic relationships among the 78 accessions by examining whole-genome genetic variations. The neighbor-joining phylogenetic tree showed that the 74 cultivated Chinese plum varieties were classified into four geographical groups: the South China group (S), comprising 22 varieties mainly from the Yangtze River Basin; the North China group (N), including 18 varieties mainly from the Yellow River Basin, the Northeast China group (NE), comprising 15 varieties mainly from Heilongjiang, Jilin, and Liaoning Provinces, and the Foreign group (F), consisting of 19 varieties mostly from Japan and the USA (Figure 5c). The same population affinities were recovered by principal component analysis (PCA), with samples from each geographical region clustered together (Figure 5b). We used a Bayesian clustering algorithm with admixed models to estimate ancestry proportions for each accession, the results of which recapitulated the findings mentioned above: at $K = 3$, the N group exhibited a consistent genetic

constitution with the S group (Figure 5d), and individuals from the N group clustered together with the S group (Figure 5b). Individuals from the N group exhibited a relatively clear boundary with the S group in the population structure analysis. We suggest that the N group was derived from the S group, with similar genotypes but different ecotypes, possibly due to natural selection or human selection. As a case in point, we identified several genes associated with the phenological period in the selection sweeps (Figure 6e; Table S28).

For the NE group, as shown in the phylogenetic and PCA results, we observed a more complex genetic constitution as it contained a mixture of individuals from the F and S groups. The N group was independent of the NE group (Figure 5b,c), suggesting that the NE and N groups originated independently from the S group. Individuals in the F group showed a close yet relatively independent genetic composition compared to the S group. We deduced that inter-species hybrids kept them apart, thus increasing the genetic diversity of the F group (Figure 5b). In addition, we identified a distinct population structure between the NE and S groups, indicating the presence of hybrids between *P. salicina* and other *Prunus* species in the NE group (Figure 5d).

Genetic diversity of the different populations

The linkage disequilibrium (LD) decayed to its half-maximum within 5 kb for nearly all four populations, but with different decay rates (S > F > N > NE) (Figure 6a). We observed a decrease in nucleotide diversity in the N and S groups compared with the F and NE groups, suggesting a potential population bottleneck, artificial selection for the N and S groups, or the possible existence of inter-species hybrids in the F and NE groups (Figure 6b). We examined the genetic differentiation between different subgroups and found that the S group was much more closely related to the N group than the F and NE groups (Figure 6d). We also estimated Tajima's *D* values for different subpopulations to examine possible selection patterns. Figure S14 shows a consecutive increase in negative Tajima's *D* values (F < NE < S < N), suggesting a gradually enhanced positive selection during hybridization, domestication, and varietal improvement. Consistent with the phylogenetic results, gene flow occurred from the F group to the S group landraces (Figure 6c). In conclusion, analysis of the splits and migrations among the *P. salicina* accessions from different ecoregions illustrated that *P. salicina* mainly spread from the Southern regions of China to other parts of China and abroad (Figure 5a).

Selection sweep during migration, hybridization, and domestication

We exploited a combined strategy integrating nucleotide diversity (π-ratio), F_{ST}, and the cross-population composite likelihood ratio (XP-CLR) to identify reliable regions containing selective sweeps during domestication or improvement. Pairwise comparisons between the different subgroups identified 35 putative selection sweeps, encompassing 4.6 Mb of the *P. salicina* genome and involving 346 genes (Table S28).

Several important genes related to flowering, stress tolerance, and secondary metabolism are highlighted in Figure 6(e) and Figures S15–S17. Flowering at appropriate and necessary time is an inevitable adaptation to each environment. We identified three critical genes associated with flowering located in the selection sweep region: *FPA*, encoding flowering time-related protein (NE versus S); *ELF3*, encoding early flowering protein (F versus S); and *PHL*, encoding phytochrome-dependent late flowering protein (NE versus S) (Table S28). *AtFPA1* plays a role in the regulation of flowering time in the autonomous flowering pathway by decreasing *FLOWERING LOCUS C* (*FLC*) mRNA levels and preventing the expression of distally polyadenylated antisense RNAs at the *FLC* locus (Schomburg et al., 2001). Mutations in the *ELF3* locus result in the loss of both photoperiod sensitivity and circadian regulation, making *ELF3* a candidate for linking circadian clock function with the photoperiodic induction of flowering in Arabidopsis; *elf3* mutant plants flower early and at the same developmental time under both long-day and short-day light conditions (Hicks et al., 2001). *AtPHL* triggers photoperiod-monitored flowering by repressing phytochrome B-dependent negative regulation of flowering (Endo et al., 2013).

The evolution of stress-related genes is a crucial part of the adaptation process. Notably, we observed that selection sweeps (harboring stress-related genes) overlapped with each other in the different comparisons, suggesting that a subset of genomic regions were selected during domestication and may have undergone successive selections, such as genes encoding hypersensitivity-related protein (*HSR*) (Lacomme and Roby, 1999), cellulose synthase (*CeS*) (Burn et al., 2002), plant disease resistance proteins (Martin et al., 2003), ABC transporters (*ABCTs*) (Martinoia et al., 2002), cytochrome P450 (*CYP450*) (Field and Osbourn, 2008), callose synthase (*CaS*) (Shi et al., 2016), CBL-interacting protein kinase (*CIPK*) (Singh et al., 2020), and α-farnesene synthase (*AFS*) (Jin et al., 2020; Wang et al., 2019) (Table S28). In addition, for the transcription factors involved in the development processes, disease resistance, and hormone-related functions (e.g., indole-3-butyric acid, abscisic acid, and ethylene), the nucleotide diversities of the F and NE groups were significantly higher (Figure S18a; Table S28).

Moreover, we identified several key genes associated with flavonoid metabolism, such as *FLS*, encoding flavonol synthase (FLS) (N versus S); *F3H*, encoding flavanone 3-dioxygenase (F versus N); *F3′H*, encoding flavonoid

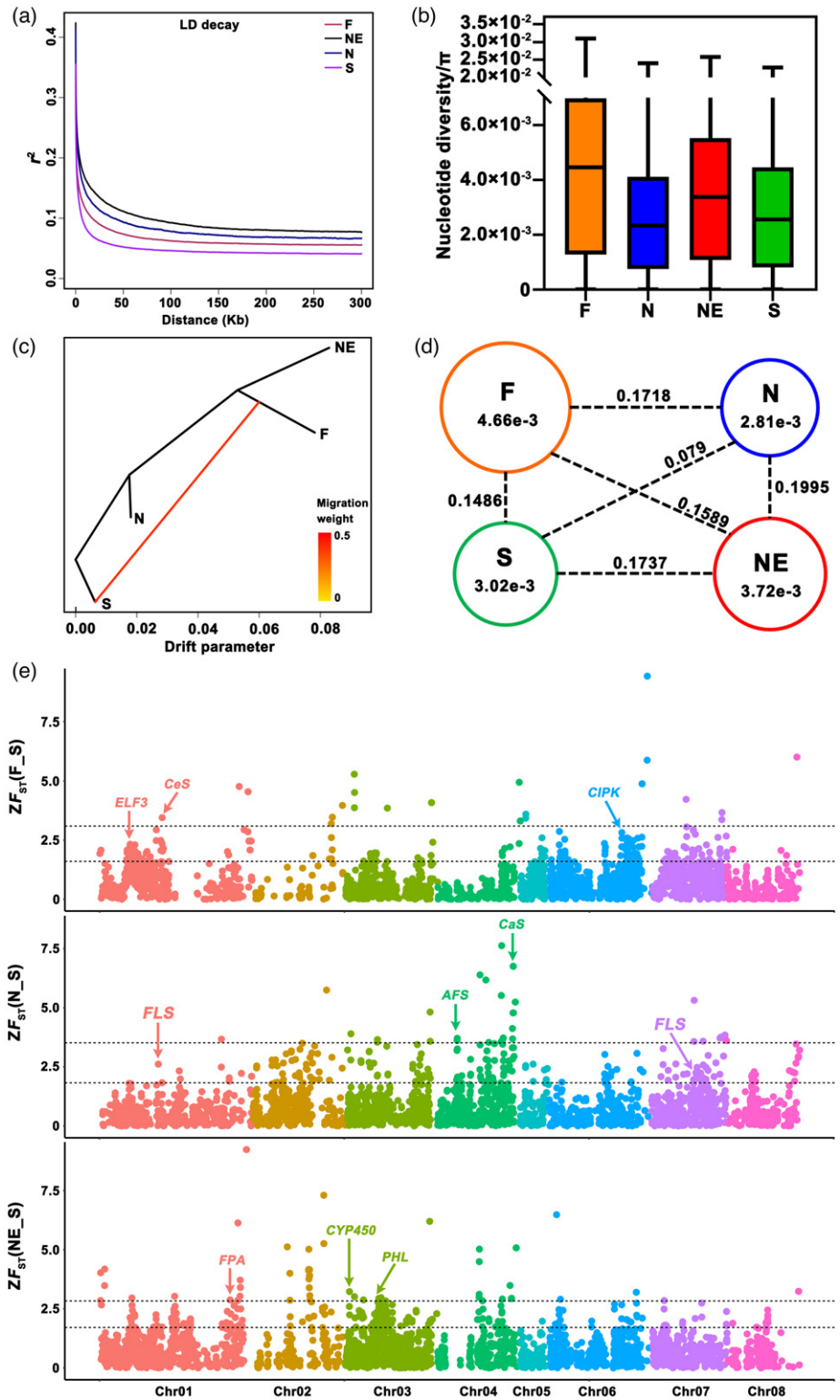

Figure 6. Population genetic diversity analysis and gene identification in the selective sweeps. (a) Linkage disequilibrium (LD) decay, (b) nucleotide diversity, (c) gene flow, and (d) F_{ST} value of different ecological groups. (e) Several key genes related to flowering, flavonoid metabolism, and stress resistance identified in the selective sweeps. The upper and lower dash lines represent the threshold of the top 1% and 5% ZF_{ST} scores, respectively. F, NE, N, and S represent the Foreign group, the Northeast China group, the North China group, and the South China group, respectively. *FPA*, flowering time control protein; *PHL*, phytochrome-dependent late flowering protein; *ELF*, early flowering protein; *FLS*, flavonol synthase; *AFS*, α-farnesene synthase; *CeS*, cellulose synthase; *CaS*, callose synthase; *CIPK*, CBL-interacting protein kinase; *CYP450*, cytochrome P450.

3′-monooxygenase (N versus S); *HCT*, encoding shikimate HCT (F versus NE); and *DFR*, encoding dihydroflavonol 4-reductase (F versus N). We also detected the nucleotide diversity values that differed greatly among subpopulations; that is, a highly domesticated group (N and S) had a lower nucleotide diversity, suggesting an underlying domestication process for flavonoid metabolism (Figure S18b, Table S28). FLS catalyzes the formation of flavonols from dihydroflavonols (Owens et al., 2008a). F3H is a 2-oxoglutarate-dependent dioxygenase that catalyzes the synthesis of dihydrokaempferol, a common precursor of three major classes of 3-hydroxy flavonoids, flavonols, anthocyanins, and proanthocyanidins (Owens et al., 2008b). F3′H catalyzes the 3-hydroxylation of the flavonoid B-ring to the 3,4-hydroxylated state (Wisman et al., 1998). HCT is involved in the biosynthesis of lignin, accepts caffeoyl-CoA and *p*-coumaroyl-CoA as substrates, and transfers the acyl group on both shikimate and quinate acceptors (Hoffmann et al., 2004). DFR catalyzes the reduction of dihydroflavonols to leucoanthocyanins and is a key enzyme in the biosynthesis of anthocyanidins, proanthocyanidins, and other flavonoids important for plant development and human nutrition (Martens et al., 2002).

We further identified overlapping regions with low nucleotide diversity for all cultivar groups, which may be related to domestication traits before their divergence. Among the top 1% lowest nucleotide diversity regions, 81 putative genes were included, and a considerable proportion were directly or indirectly involved in the disease resistance process, including genes encoding WRKY52/RRS1 (Rushton et al., 2010), disease resistance protein (Martin et al., 2003), and leaf rust 10 disease-resistance locus receptor-like protein kinase (Feuillet et al., 1997) (Table S29). In addition, vital gene members related to cell division and development were covered, for example, *SWA1*, encoding SLOW WALKER 1 (Shi et al., 2005); and *TSK*, encoding TONSOKU (Suzuki et al., 2005) (Table S29). However, we did not find the genes present in the identified selective sweep regions, which suggests that natural selection towards the genes was nearly completed before the divergence of the cultivar groups.

Reconstruction of flavonoid–anthocyanin metabolism in *P. salicina*

Using two cultivars with distinct fruit colors and flavors ('Friar' and 'Qingcuili'), we aimed to reconstruct the flavonoid–anthocyanin metabolism pathway and identify essential candidate genes associated with flavonoid content using integrated methods (time-serial RNA-seq and metabolomic analysis) (Figure 7). Using high-throughput transcriptome sequencing, we identified 23 516 expressed genes across all the samples, extracted the 5313 variably expressed genes, constructed 16 co-expression modules, and placed the top 100 hub genes in each module (Figure 7b,c). We annotated 94 genes involved in flavonoid–anthocyanin metabolism, of which 73 were expressed and 36 were differentially expressed genes, including those with dramatic expression changes, such as genes encoding FLS, anthocyanin synthase (*ANS*), chalcone synthase (*CHS*), chalcone isomerase (*CHI*), flavonoid 3-monooxygenase (*F3H*), leucoanthocyanidin reductase (*LAR*), and shikimate HCT (Table S32).

Eight important metabolites were quantified in the pathway, including peonidin *O*-hexoside (Peohex), cyanidin 3-*O*-glucoside (Cy3glu), procyanidin A1 (ProA1), procyanidin A2 (ProA2), procyanidin B2 (ProB2), cyanidin 3-*O*-rutinoside (Cy3rut), cyanidin 3-*O*-galactoside (Cy3gal), and peonidin 3-*O*-glucoside chloride (Peo3glu) (Figure 7a). We observed significant differences in anthocyanin composition between the two cultivars at different developmental stages: for 'Qingcuili', only ProA1 accumulated at the early stages and decreased with time, while the remaining seven kinds of metabolites were rarely detected across the three stages. For 'Friar', three types of procyanidins, especially A2 and B2, showed the same accumulation pattern to that of 'Qingcuili', but the other five metabolites were significantly enriched along with time, corresponding with the peel color formation. We concluded that the distinct fruit color of the two cultivars was caused by the difference in anthocyanin content, and the metabolic pathways for the flavonoids and procyanidins varied greatly between the two cultivars.

Furthermore, we related the eight important metabolites with the 16 co-expressed modules, and identified five modules (turquoise, pink, blue, brown, and yellow) associated with different metabolites ($P < 0.01$) (Figure 7b,c). We assigned 18 hub genes (five for turquoise, seven for pink, three for blue, and three for brown) in this pathway based on gene connectivity in the gene expression network (Figures S19 and S20, Table S33). The motifs of the promoter regions were analyzed, and several candidate regulators for each module were obtained, including *ERF*, *bHLH*, and *WRKY* (Table S34). Consistently, all genes were significantly differentially expressed between the two cultivars, suggesting that they play critical roles for the putative genes in flavonoid and procyanidin synthesis. Except for the lowly expressed genes, we did not find the hub genes located in the selection sweep, suggesting (i) that hub genes are functionally conserved and (ii) their diversity of regulation during selection and domestication.

DISCUSSION

Plum is an essential component of the phylogenetic architecture of the genus *Prunus*. To date, the genomic information for several other important drupe fruit species closely related to Chinese plum has been reported, including Japanese apricot, peach, sweet cherry, almond, and apricot, which supports further functional genomic studies and

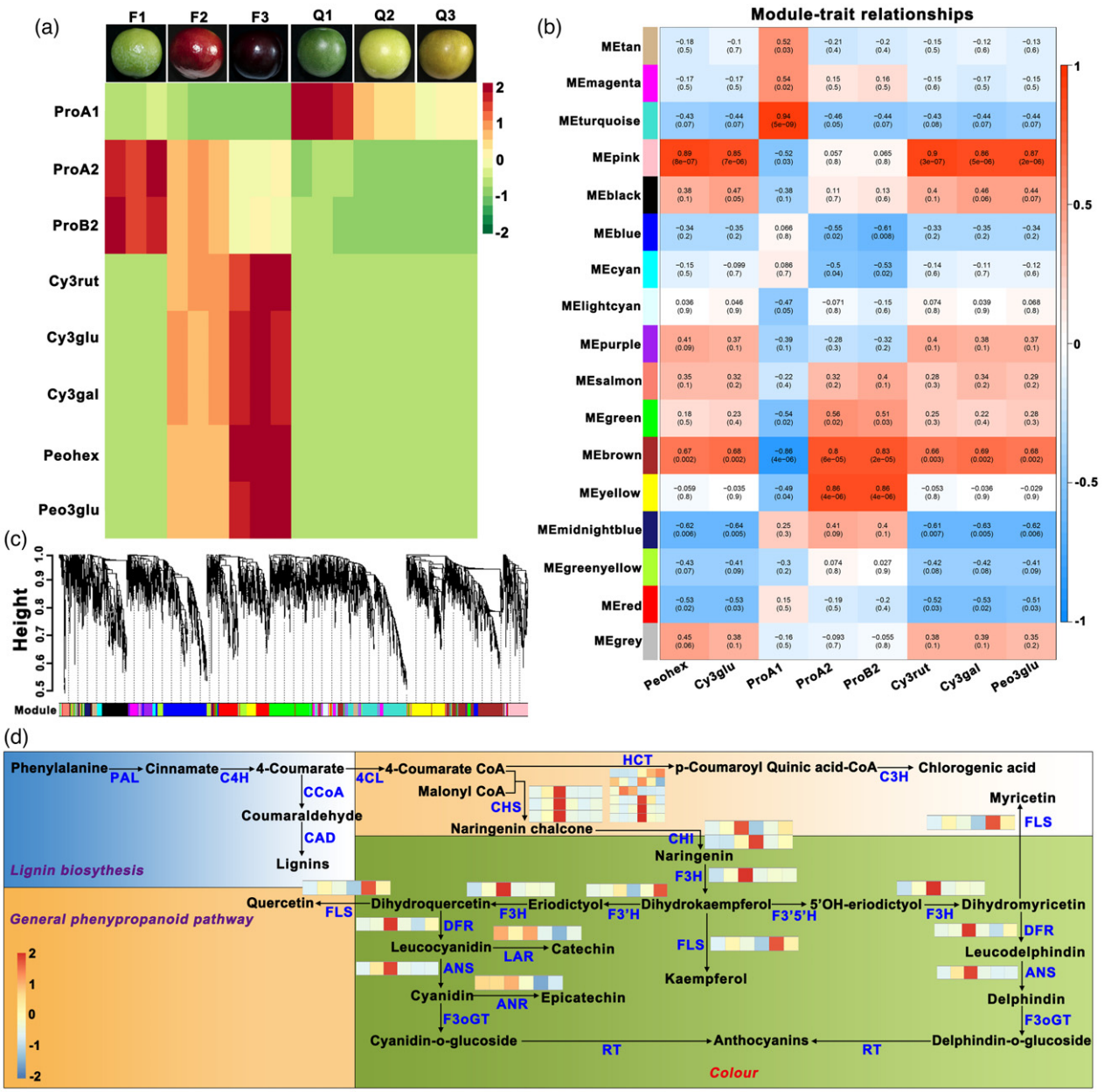

Figure 7. Multi-omics analysis of flavonoid metabolism in Chinese plum. (a) Metabolite identification of 'Friar' (F) and 'Qingcuili' (Q). (b) Heatmap of the correlation between module eigengenes (MEs) and different metabolites. For each module, the ME value was calculated, which represents the expression profile of the module. The ME values were correlated with binary variables (Spearman's correlation). (c) The cluster dendrogram of genes in the transcriptome dataset of F and Q. Each branch represents one gene, and every color below represents one co-expression module. (d) Reconstruction of the flavonoid–anthocyanin metabolic pathway. ProA1, procyanidin A1; ProA2, procyanidin A2; ProB2, procyanidin B2; Cy3rut, cyanidin 3-O-rutinoside; Cy3glu, cyanidin 3-*O*-glucoside; Cy3gal, cyanidin 3-*O*-galactoside; Peohex, peonidin *O*-hexoside; Peo3glu, peonidin 3-*O*-glucoside chloride.

DNA-informed breeding (Alioto et al., 2019; Callahan et al., 2021; Jiang et al., 2019; Pinosio et al., 2020; Sánchez-Pérez et al., 2019; Shirasawa et al., 2017; Verde et al., 2017; Zhang et al., 2012). Here, we present a chromosome-scale Chinese plum reference genome, comprising a high-quality sequence of 318.6 Mb with eight pseudo-chromosomes. The contig N50 size was 2.3 Mb, which was larger than that for the abovementioned stone fruit species genomes and the two recently released genomes for the diploid Chinese plum cultivar 'Sanyueli' (Fang et al., 2020; Liu et al., 2020b). We identified 5101 structural variations (SVs) and 2 705 789 single nucleotide variant (SNVs) (covering approximately 4% of the genome) between 'Zhongli No.6' and 'Sanyueli' genomes (Table S36). The significantly higher mapped reads ratio of the 78 accessions using 'Zhongli No.6' as reference suggested that our

assembly with high integrity and continuity could be preferentially selected as reference genome in population genetic and genomic studies (Table S35). We annotated a total of 26 234 highly confident protein-coding genes and 5648 non-coding RNA genes. These results were further confirmed by the favorable results of multiple evaluation approaches (Table S13), which suggests that our *P. salicina* genome assembly had high continuity, completeness, and accuracy. This is significant for future genetic diversity analysis, comparative genome evolution studies, and the development of genome-based breeding tools.

Gene duplications provide evolutionary potential to generate diversified functions, while polyploidization or WGD events double the chromosomes, initially resulting in hundreds to thousands of retained duplicates (Ren et al., 2018). WGDs and environmental factors, including animals, have contributed to the evolution of many fruits in Rosaceae (Xiang et al., 2017), as several WGD events have been identified, such as those in apples and pears (Velasco et al., 2010; Wu et al., 2013). In *Prunus*, a common WGD event was found in an ancestor and may contribute to the enhancement of its adaptation and diversity (Jiang et al., 2019; Verde et al., 2013). Recent SDs have also been recognized as important mediators of gene and genome evolution, and have led to various types of genome rearrangements and other genome structural changes, both between and within species (Zhao et al., 2017). Frequent SD events have also been detected in many other species, such as *Arabidopsis thaliana*, *Vitis vinifera*, *Bombyx mori*, and *Mikania micrantha* (Baumgarten et al., 2003; Giannuzzi et al., 2011; Liu et al., 2020a; Zhao et al., 2013). We identified large-scale SDs in the *P. salicina* genome (Figure 3; Table S19). By investigating the genes in the duplicated regions, we demonstrated their involvement in secondary metabolism and stress resistance, especially flavonoid metabolism and disease resistance, which could enhance the associated metabolism efficiency. A similar situation was detected in the apricot genome. We compared the genes in the SDs of the two genomes and found distinct genes, and deduced that the SD events in each genome were independent, based on their *K*s profiles and increased gene copy numbers. We propose that the SDs in the *P. salicina* genome could contribute to the enhancement of flavonoid metabolism and adaptations to the environment, and even cause species differentiation.

Southwest China was previously assumed to be a primitive domestication center of Chinese plum, which was evidenced by the discovery of wild plum communities in the middle reaches of the Yangtze River Basin (Hartmann and Neumulle, 2009; Wei et al., 2020; Zhang and Zhou, 1998), a palynology investigation (Guo, 2006), a large-scale phenotypic variation analysis (Yu et al., 2011), and multiple genetic diversity analyses (Chen and Yang, 2014; Wei et al., 2019, 2021; Zuo et al., 2015). The results of the present study were consistent with those findings. Based on the population genetic study, we found several cultivars from South China distributed in three ecological groups (Figure 5d), along with gene flow between the F and S groups (Figure 6c). Previous studies have reported that most Chinese plum cultivars from Japan and the improved Chinese plum hybrids from the USA were distributed across the Chinese indigenous plum groups in the dendrogram analysis (Liu et al., 2006, 2007; Wei et al., 2019), which is in accordance with the predominant *P. salicina* genetic components found in the improved Chinese plum hybrids (Faust and Suranyi, 1999). This is in agreement with the hypothesis that plum cultivars that originated from China were initially introduced to Japan, and then exported to the USA in 1870 from Japan (Hartmann and Neumulle, 2009; Topp et al., 2012). In summary, we propose that Chinese plums in South China possess a broader genetic background and more frequent gene exchange with those from other geographical groups, and their possible center of origin is in South China.

Plum contains high levels of flavonoids, such as anthocyanins and chlorogenic acid, which have been shown to have a broad range of health-promoting activities, including protection against cardiovascular disease, diabetes, digestive disorders, and osteoporosis (Stacewicz-Sapuntzakis, 2013). In this study, we determined the underlying reasons for the high flavonoid content at the genome level: Gene copies of key genes involved in flavonoid metabolism (e.g., *HCT* and *FLS*) were increased in two stages: SD and tandem duplication (Figures 4b and 7d; Tables S28 and S29). Genes involved in flavonoid metabolism may also play an important role in human selection and domestication. We identified several genes (e.g., *HCT* and *FLS*) located in the selective sweep regions, and for the F and NE groups, the Tajima's *D* values were higher, suggesting the underlying roles of those genes in flavonoid metabolism (Figure S18). To reconstruct the gene network and underlying regulators of flavonoid metabolism, we performed weighted gene co-expression network analysis and identified 18 hub genes (e.g., *HCT*, *LAR*, and *FLS*) and regulators (e.g., *ERF*, *bHLH*, and *WRKY*) using multi-omics analysis (Figures S19 and S20). These studies have shed light on the flavonoid metabolism features of Chinese plum.

In conclusion, the high-quality reference genome and population genomic analysis of *P. salicina* performed here will expand our understanding of the genetic basis of agronomically important traits in Chinese plum cultivars. As expected, we identified hundreds of genes selected during domestication/improvement, many of which are functionally involved in Chinese plum quality and stress resistance. However, before the establishment of highly effective molecular breeding programs, extensive sampling and systemic tests at the gene level are warranted to determine whether these candidate genes were truly selected during

domestication/improvement or are simply genetic hitchhikers in other regions targeted by artificial selection.

EXPERIMENTAL PROCEDURES

Sample preparation

Plant materials were sourced from the experimental orchard of ZFRI-CAAS, Zhengzhou, Henan Province, China (34°45′23.7962″N, 113°38′58.72″E). The latest selected cultivar *Prunus salicina* L. cv. Zhongli No.6 from ZFRI-CAAS was used for *de novo* genome assembly and genome annotation. Four different tissues from the same tree (flower, leaf, fruit, and stem bark) were collected for transcriptome sequencing to assist in genome assembly and annotation. The whole-genome re-sequencing data of a panel of 78 *Prunus* accessions were used to investigate the genetic diversity and adaptive evolution of Chinese plums (Table S1). The full-colored Chinese plum cultivar 'Friar' and non-colored Chinese plum cultivar 'Qingcuili' were used to dissect the mechanisms of peel color formation by integrating their transcriptome and metabolome analyses.

Nucleic acid extraction, library construction, and sequencing

Genomic DNA was extracted from young leaves bursting after blooming using the QIAGEN® Genomic DNA Kit, and total RNA was extracted from four different tissues (flower, leaf, fruit, and stem bark) using the QIAGEN RNeasy® Plus Mini Kit following the manufacturer's instructions. The quality and quantity of the isolated DNA and RNA were separately assessed using electrophoresis on a 0.75% agarose gel, a NanoDrop™ D-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA), and an Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA).

For Nanopore sequencing, the purified DNA was fragmented and size-fractionated (1050 kb) using a g-tube and BluePippin. The purified DNA was prepared for Nanopore sequencing following the protocol provided with the genomic sequencing kit SQK-LSK109 (Oxford Nanopore Technologies, Oxford, UK), and single-molecule real-time sequencing of long reads was conducted on a PromethION platform (ONT, Oxford, UK). A total of 59.91 clean ONT data with an average pass read length of 23.5 kb were generated after quality filtering, the longest of which was 322.2 kb (Table S3). Illumina paired-end sequencing libraries were generated following the manufacturer's standard protocols and sequenced using the Illumina HiSeq X-ten platform (Illumina, San Diego, CA, USA). Paired-end 150-bp (PE150) reads were generated from libraries with an insert size of 350 bp.

For RNA sequencing, paired-end libraries were constructed using the TruSeq Sample Preparation kit following the manufacturer's instructions and then sequenced using the Illumina HiSeq X-ten platform (Illumina) with a read length of 150 bp. Full-length cDNA libraries were then constructed using the IsoSeq Library Construction Kit and sequenced using the PacBio Sequel system (Pacific Biosciences, Menlo Park, CA, USA). Using the RNA-seq reads from different libraries, we conducted *de novo* assembly using the commonly used Trinity software (Grabherr et al., 2013) and SOAPdenovo-Trans (Xie et al., 2014). All transcripts were merged and fed into the downstream genome.

For Hi-C sequencing, approximately 2.0 g of young leaves was cut into small sections and immersed in a mixture of 35 ml of pre-cooled NIB buffer, 35 μl of mercaptoethanol, 35 μl of phenylmethylsulfonyl fluoride, and 2 ml of formaldehyde for 15 min for 90 min in a rotary hybridizer. Glycine (2.5 ml) was added to terminate the crosslinking reaction, and the leaf sections were removed from the mixture, rinsed with ddH_2O several times, and pounded to a fine powder in liquid nitrogen for the isolation of the cross-linked DNA. To further generate a chromosome-scale assembly of the *P. salicina* genome, we used the Hi-C sequencing data to increase contiguity (Belton et al., 2012). The isolated cross-linked DNAs were purified, digested with *Dpn*II enzymes, blunt-end-repaired, and tagged with biotin. Then, the biotin-containing DNA fragments were captured and PCR-enriched for the construction of the Hi-C library. The final libraries were sequenced using the Illumina HiSeq X-ten platform (Illumina) with the PE150 sequencing strategy.

Genome assembly and quality assessment

To estimate the *P. salicina* genome size and heterozygosity rate, we performed k-mer analysis using the 22.9-Gb Illumina clean pair-end reads generated from the Chinese plum cultivar 'Zhongli No.6' (Kajitani et al., 2014). For genome assembly, the filtered ONT long reads (mean_qscore > 7) were first corrected using Canu (version 1.8) with default parameters (Koren et al., 2017). The corrected long reads were subjected to *de novo* assembly using SMARTdenovo with default parameters (Ruan, 2016). The initial assembly was polished using NextPolish (version 1.0) with five rounds of iteration (Hu et al., 2020). The redundant contigs led by heterozygosity were identified and removed using Redundans (version 0-13c) with the parameters '--identity 0.93--overlap 0.93' (Pryszcz and Gabaldón, 2016). Multiple methods were used to evaluate the quality of our final genome assembly, including a BUSCO evaluation (version 4.0.5) (Simão et al., 2015), CEGs assembled with CEGMA (version 2.0) (Parra et al., 2007), DNA read mapping with BWA (version 0.7.12-r1039) (Li and Durbin, 2010), and RNA read mapping using GMAP (version 2019-09-12) (Wu and Watanabe, 2005) (Tables S8–S12). The genome alignment of 'Zhongli No.6' and 'Sanyueli' was conducted using MUMmer with default parameters (Delcher et al., 2002) and SVs were analyzed using Assemblytics (Nattestad and Schatz, 2016).

Hi-C assisted scaffolding

We obtained a total of 45.7 Gb of Hi-C raw data, providing approximately 150-fold coverage of the *P. salicina* genome, which we used for chromosome-level assembly with the following steps: (i) filtering the raw data using fastp (version 0.12.3) with default parameters (Chen et al., 2018); (ii) mapping the clean paired-end reads to the *P. salicina* assembly using Bowtie 2 (version 2.3.2) with the mode end-to-end and the parameters --very-sensitive -L 30 (Langmead and Salzberg, 2012); (iii) eliminating invalid interaction pairs, including dangling end, self-circle, and dumped pairs, by recognizing the *Dpn*II enzyme cutting sites (Table S4); and (iv) chromosome-level scaffolding using an agglomerative hierarchical clustering strategy with LACHESIS (Burton et al., 2009). In total, 227 adjacent anchored scaffolds (representing 90.98% of the total length) were connected using 100 bp Ns to form eight pseudo-chromosomes with lengths ranging from 11.99 to 62.48 Mb (Table S5).

Genome annotation

We adopted three strategies for predicting the protein-coding genes in the *P. salicina* genome: (i) *ab initio* gene prediction using AUGUSTUS (version 3.3.3) (Stanke et al., 2008), FGenesh (Salamov and Solovyev, 2000), and GlimmerHMM (Majoros et al., 2004); (ii) homology-based gene prediction using GeneWise (version 2.4.1) (Birney et al., 2004); and (iii) RNA-seq assisted gene

prediction using the Program to Assemble Spliced Alignments (PASA) (Haas et al., 2003). All gene models were integrated using EVidenceModeler (EVM, version 1.1.1) to generate a consensus set (Haas et al., 2008) (Table S16).

In total, six types of SSRs, from mono- to hexa-nucleotides, were annotated by running the MIcroSAtellite identification (MISA) Perl script (https://webblast.ipk-gatersleben.de/misa/) (Table S14). To identify other types of repeat sequences in the *P. salicina* genome assembly, we first built an NR repeat sequence library by searching for repetitive sequences using LTR-FINDER (Xu and Wang, 2007), MITE-Hunter (Han and Wessler, 2010), and RepeatModeler (www.repeatmasker.org/RepeatModeler/), and repetitive sequences in the *P. salicina* genome assembly were identified using RepeatMasker (http://www.repeatmasker.org). The NR repeat sequence library was then integrated with the Repbase database (Bao et al., 2015) and subjected to further detection of the repeat sequences using RepeatMasker and the Tandem Repeat Finder package (version 4.04) (Benson, 1999), including transposable elements, tandem repeats, and non-interspersed repeat sequences (Table S14). Moreover, the non-coding RNA genes were predicted using the tRNAscan-SE package (version 1.3.1) (Lowe and Chan, 2016), RNAmmer algorithms (Lagesen et al., 2007), INFERNAL (version 1.1.3) (Nawrocki and Eddy, 2013), and the publicly available Rfam database (release 13.0) (Kalvari et al., 2018) (Table S18).

Functional annotation of protein-coding genes

Gene functions were assigned to the protein-coding gene models and compared to the National Center for Biotechnology Information (NCBI) non-redundant (NR) protein database and the Universal Protein Resource Knowledgebase (UniProtKB) Swiss-Prot protein database (release 2019_11) using BLASTP (Altschul et al., 1997) with an *E*-value threshold of 1e−5. The motifs and domains were identified using InterProScan (version 5.2-45.0, ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/) against multiple publicly available databases including ProDom, PRINTS, Pfam, SMRT, PANTHER, and PROSITE (Hunter et al., 2009). Gene Ontology (GO, http://www.geneontology.org) annotation was performed based on the corresponding InterPro entries using Blast2GO (Götz et al., 2008) with an *E*-value threshold of 1e−5. Metabolic pathway annotations were performed using sequence comparisons with the Kyoto Encyclopedia of Genes and Genomes database (release 92.0) with BLASTP and an E-value threshold of 1e−5. Functional comparisons were also carried out in the Cluster of Orthologous Groups of proteins database using BLASTP with an E-value threshold of 1e−5 to classify proteins from completely sequenced genomes based on the orthology concept (Tatusov et al., 2000) (Table S17).

Genome evolution analysis

The gene families of *P. salicina* and 16 other representative plant genomes, including *Arabidopsis thaliana*, *Fragaria vesca*, *Malus domestica*, *Oryza sativa*, *Prunus armeniaca*, *Prunus avium*, *Pyrus communis*, *Prunus dulcis*, *Populus euphratica*, *Prunus mume*, *Prunus persica*, *Prunus yedoensis*, *Rosa chinensis*, *Rubus occidentalis*, *Solanum lycopersicum*, and *Vitis vinifera*, were identified using the OrthoMCL package (version 2.0.9) (Li et al., 2003) and the Markov cluster algorithm (MCL) with default inflation parameters (Enright et al., 2002) (Table S20). The species-specific gene families were determined according to the presence or absence of genes for a given species, according to the results of OrthoMCL (Tables S20–S22). A phylogenetic tree for the 17 plant species was constructed using the RAxML package (version 8.2.11) with the PROTGAMMAAUTO model based on 224 high-quality single-copy orthologous genes (Stamatakis, 2014) (Figure S10). The divergence times were estimated using the MCMCTree program (Arvestad et al., 2003) which was incorporated in the Phylogenetic Analysis using the Maximum Likelihood (PAML) software package (version 4.9e) (Yang, 2007), and the time tree was calibrated using the estimated divergence times obtained from the TimeTree database (http://www.timetree.org/) (Figure S11). The expansion and contraction of the gene families were determined using Computational Analysis of gene Family Evolution (CAFE, version 3.1) (Bie et al., 2006). Positive selection sites were detected with *P. salicina* as a predetermined branch using Codeml implemented in the PAML package with a branch-site model (Yang, 2007) (Figure S12). The tandem gene duplications were identified using the synteny-based approach implemented in CoGE SynMap tools (https://genomevolution.org/coge/) (Lyons and Freeling, 2008).

To detect the polyploidization event, we performed all-versus-all paralog analysis in the *P. salicina* genome using RBHs from primary protein sequences by self-*BLASTp* in Chinese plum. RBHs are defined as reciprocal best *BLASTp matches* with an E-value threshold of 1e−5, a *c*-score (BLAST score/best BLAST score) threshold of 0.3 (Putnam et al., 2008), and an alignment length threshold of 100 amino acids. The synonymous substitution rate (*Ks*) of the RBH gene pairs was calculated based on the YN model using the *KaKs_Calculator* (version 2.0) (Wang et al., 2010a).

Population evolution and genetic diversity

Of the 78 *Prunus* accessions used in our study, the raw reads of *P. armeniaca* were obtained from the China National GeneBank Sequence Archive (CNSA) (https://db.cngb.org/cnsa/) with enquiry number CNP0000755 (Jiang et al., 2019). The raw reads of the other 55 *Prunus* accessions were obtained from the Sequence Read Archive of the NCBI (https://www.ncbi.nlm.nih.gov/sra/) with enquiry number PRJNA659814 (Wei et al., 2021). Whole-genome re-sequencing of the remaining 22 accessions was conducted in the current study at high coverage (approximately 20×) using the Illumina HiSeq X-ten platform (Illumina) with the PE150 sequencing strategy (Table S1). Reads were aligned to the reference genome using BWA (version 0.7.12-r1039) (Li and Durbin, 2010). Picard tools (version 2.23.6) were processed to remove duplicates (broadinstitute.github.io/picard/). To improve the accuracy, InDel realignment was conducted using RealignerTargetCreator and IndelRealigner packed in the Genome Analysis ToolKit (version 4.2.0.0) (McKenna et al., 2010). SNPs and InDels were identified using SAMtools (version 1.4) (Li et al., 2009). We inferred missing genotype data using the hidden Markov model (HMM) in Beagle software (version 4.0) (Browning and Browning, 2016) with the following default parameter settings: unphased and non-reference; iterations = 10, window = 50 000, nthreads = 10. To confirm the quality of the variations in the population analysis, we set strict filtering parameters for each sample (depth $\geq$ 5, phred-scaled quality $\geq$ 40). Variations were annotated using the ANNOVAR package (version 2018-04-16) (Wang et al., 2010b).

To analyze the phylogenetic relationships, we constructed a neighbor-joining tree with TreeBeST (version 1.9.2) (Vilella et al., 2009), inferred the population structure using ADMIXTURE (version 1.23) (Alexander et al., 2009), and conducted PCA using PLINK (version 1.07) (Purcell et al., 2007). TreeMix (version 1.13) was used to evaluate gene flow among different geographical groups (Pickrell and Pritchard, 2012).

The LD of each group was estimated by computing the squared correlation coefficient (r^2) values between any two SNPs within 500 kb using PopLDdecay (Zhang et al., 2019). The genetic nucleotide diversity (π), Tajima's *D* value, and the population-

differentiation statistics (fixation index, F_{ST}) were calculated using VCFtools (version 0.1.14) with a 20-kb sliding window and a step size of 10 kb across the *P. salicina* genome (Danecek et al., 2011).

Genome scanning for divergent regions

We calculated the genome-wide distribution of nucleotide diversity (π), Z-transformed F_{ST}, and the XP-CLR to detect selective signatures between different ecological populations with a sliding window size of 100 kb and a step size of 50 kb (windows with fewer than 10 SNPs were ignored). Putative selection targets with the top 5% and bottom 5% of π-ratios, the top 5% of F_{ST} values, and the top 5% of XP-CLR scores were extracted as high-confidence outliers.

Anthocyanin metabolism

The peel samples of the full-colored Chinese plum cultivar 'Friar' and the non-colored Chinese plum cultivar 'Qingcuili' were collected from three fruit developmental stages, including the green-mature stage, the color-turning stage, and the fully mature stage, with three replicates for each stage and 10 uniform fruits for each replicate. The preparation, extract analysis, metabolite identification, and quantification of the samples were performed at Wuhan METWARE Biotechnology Co., Ltd., following standard procedures, and metabolite data analysis was conducted using Analyst software (version 1.6.1; AB SCIEX, Ontario, Canada). Metabolites with $|\log_2(\text{fold change})| \geq 1$ were considered as differential metabolites for group discrimination (Chen et al., 2013; Yuan et al., 2018).

The same peel samples of 'Friar' and 'Qingcuili' used in the metabolome analysis were also subjected to transcriptome sequencing using the Illumina HiSeq X-ten platform (Illumina) with a read length of 150 bp. Approximately 10 Gb clean data for each sample were obtained, and the clean reads were mapped to the reference genome using HISAT2 (version 2.0.5) with default parameters (Kim et al., 2015) (Table S30). Then, transcript abundances were quantified using StringTie (version 2.1.3) (Pertea et al., 2015) with the expected number of fragments per kilobase of transcript sequence per million base pairs sequenced method. Differential gene expression between different samples was identified using DESeq2 (version 3.11) (Love et al., 2014).

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program (2019YFD1000203), the Agricultural Science and Technology Innovation Program (CAAS-ASTIP-2021-ZFRI), the National Horticulture Germplasm Resources Center, and the Key Scientific and Technological Project of Henan Province (No. 202102110049).

AUTHOR CONTRIBUTIONS

LW and KC conceived the project; YC and ZH prepared the plant materials; ZH and FS managed the genomic bioinformatics analyses; ZH and FS wrote the manuscript; LW and KC revised the manuscript. All authors read and approved the final manuscript.

CONFLICT OF INTEREST

The authors declare that they have no competing financial interests.

DATA AVAILABILITY STATEMENT

The sequencing data were deposited in the NCBI database under the BioProject accession PRJNA723734. Whole genome information of *Prunus salicina* cv. Zhongli No. 6 (v1.0) is also available from the Genome Database for Rosaceae (https://www.rosaceae.org/Analysis/9019655).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Frequency distribution of the 17-mer graph analysis used to estimate the size of the *Prunus salicina* genome.

Figure S2. Schematic diagram of simulation curve of the *Prunus salicina* heterozygosity rate.

Figure S3. Contig length distribution of the final assembly of the *Prunus salicina* genome.

Figure S4. Evaluation of the *Prunus salicina* genome assembly by mapping the core eukaryotic genes (CEGs).

Figure S5. Synonymous substitutions per synonymous site (*Ks*) distribution of the apricot genome.

Figure S6. Synonymous substitutions per synonymous site (*Ks*) distribution of the apricot genome versus the Chinese plum genome.

Figure S7. Evaluation of the *Prunus salicina* genome assembly by analysis of the GC depth distribution with ONT long reads.

Figure S8. Significantly enriched KEGG pathways of the genes specific to the *Prunus salicina* lineage.

Figure S9. Significantly enriched GO terms of the genes specific to the *Prunus salicina* lineage.

Figure S10. The phylogenetic tree of *Prunus salicina* and 16 other fully sequenced plant genomes.

Figure S11. The divergence times of *Prunus salicina* and 16 other fully sequenced plant genomes.

Figure S12. Expansion and contraction of gene families of the *Prunus salicina* genome and 16 other fully sequenced plant genomes.

Figure S13. SNP density across the *P. salicina* genome.

Figure S14. The Tajima's *D* values of different populations across the Chinese plum genome.

Figure S15. The Z-transformed F_{ST} values of different populations across the Chinese plum genome.

Figure S16. The nucleotide diversity values of different populations across the Chinese plum genome.

Figure S17. The XP-CLR scores of different populations across the Chinese plum genome.

Figure S18. Nucleotide diversity values of the selected genes related to (a) stress tolerance and (b) flavonoid metabolism in different populations.

Figure S19. Motif analysis of hub genes in the pink module.

Figure S20. Motif analysis of hub genes in the turquoise module.

Table S1. List of *Prunus* accessions used in the population genetic study.

Table S2. Summary of Illumina sequencing.

Table S3. Summary of Nanopore sequencing.

Table S4. Summary of Hi-C sequencing.

Table S5. Statistics of the pseudo-chromosomes of *Prunus salicina*.

Table S6. Statistics of the initial assembly results of *Prunus salicina*.

Table S7. Statistics of the final assembly results of *Prunus salicina*.

Table S8. BUSCO evaluation of the *Prunus salicina* genome.

Table S9. CEGMA evaluation of the *Prunus salicina* genome.

Table S10. Quality assessment of the *Prunus salicina* genome using sequence identity evaluation.

Table S11. Quality assessment of the *Prunus salicina* genome by calculating the genome single-nucleotide error rate.

Table S12. Quality assessment of the assembled genome of *Prunus salicina* by aligning the RNA-seq reads.

Table S13. Comparisons between the three versions of *Prunus salicina* genome assemblies.

Table S14. Specific statistics of the annotated SSRs in the *Prunus salicina* genome.

Table S15. Summary of repetitive sequences identified in the *Prunus salicina* genome.

Table S16. Summary statistic of annotated genes in the *Prunus salicina* genome.

Table S17. Annotation of the protein-coding genes in the *Prunus salicina* genome.

Table S18. Annotation of the non-coding RNA genes in the *Prunus salicina* genome.

Table S19. *Ks* value of the syntenic gene pairs in the *Prunus salicina* genome.

Table S20. Statistics of the gene families of *Prunus salicina* and 16 other representative plant species.

Table S21. KEGG functional enrichment analysis of the gene families specific to *Prunus salicina*.

Table S22. Gene Ontology enrichment analysis of the gene families specific to *Prunus salicina*.

Table S23. KEGG functional enrichment analysis of the expanded gene families of the *Prunus salicina* genome.

Table S24. Gene Ontology enrichment analysis of the expanded gene families of the *Prunus salicina* genome.

Table S25. KEGG functional enrichment analysis of the contracted gene families of the *Prunus salicina* genome.

Table S26. Gene Ontology enrichment analysis of the contracted gene families of the *Prunus salicina* genome.

Table S27. Functional annotation of the positively selected genes of the *Prunus salicina* genome.

Table S28. Genes annotated in the selection sweeps between different ecological groups.

Table S29. Putative selective regions and genes of all cultivated groups.

Table S30. Statistics of the transcriptome of peel samples of 'Friar' (F) and 'Qingcuili' (Q).

Table S31. Number of the differentially expressed genes between the samples of 'Friar' (F) and 'Qingcuili' (Q).

Table S32. Key genes annotated in the flavonoid–anthocyanin metabolism pathway by RNA-seq and metabolomic analysis of 'Friar' (F) and 'Qingcuili' (Q).

Table S33. Identified hub genes and gene expression in the flavonoid metabolism pathway.

Table S34. Candidate transcription factors in the flavonoid metabolism pathway.

Table S35. Statistics of the mapped reads using the two reference genomes of 'Zhongli No.6' and 'Sanyueli'.

Table S36. Statistics of genomic variations between 'Zhongli No.6' and 'Sanyueli' genomes.

Appendix S1. Supplementary methods.

REFERENCES

Alexander, D.H., Novembre, J. & Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664.

Alioto, T., Alexiou, K.G., Bardil, A., Barteri, F., Castanera, R., Cruz, F. *et al.* (2019) Transposons played a major role in the diversification between the closely related almond and peach genomes: results from the almond genome sequence. *The Plant Journal*, **101**, 455–472.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Arvestad, L., Berglund, A.C., Lagergren, J. & Sennblad, B. (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics*, **19**, 7–15.

Bak, S. & Feyereisen, R. (2001) The involvement of two P450 enzymes, CYP83B1 and CYP83A1, in auxin homeostasis and glucosinolate biosynthesis. *Plant Physiology*, **127**, 108–118.

Bao, W., Kojima, K.K. & Kohany, O. (2015) Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 4–9.

Baumgarten, A., Cannon, S., Spangler, R. & May, G. (2003) Genome-level evolution of resistance genes in *Arabidopsis thaliana*. *Genetics*, **165**, 309–319.

Bayer, A., Ma, X. & Stöckigt, J. (2004) Acetyltransfer in natural product biosynthesis-functional cloning and molecular analysis of vinorine synthase. *Bioorganic & Medicinal Chemistry*, **12**, 2787–2795.

Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y. & Dekker, J. (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*, **58**, 268–276.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, **27**, 573–580.

Birney, E., Clamp, M. & Durbin, R. (2004) GeneWise and genomewise. *Genome Research*, **14**, 988–995.

Browning, B.L. & Browning, S.R. (2016) Genotype imputation with millions of reference samples. *American Journal of Human Genetics*, **98**, 116–126.

Burn, J.E., Hocart, C.H., Birch, R.J., Cork, A.C. & Williamson, R.E. (2002) Functional analysis of the cellulose synthase genes *CesA1*, *CesA2*, and *CesA3* in *Arabidopsis*. *Plant Physiology*, **129**, 797–807.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. & Shendure, J. (2009) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, **31**, 1119–1125.

Callahan, A.M., Zhebentyayeva, T.N., Humann, J.L., Saski, C.A., Galimba, K.D., Georgi, L.L. *et al.* (2021) Defining the 'HoneySweet' insertion event utilizing NextGen sequencing and a de novo genome assembly of plum (*Prunus domestica*). *Horticulture Research*, **8**(1). https://doi.org/10.1038/s41438-020-00438-2

Carrasco, B., González, M., Gebauer, M., García-González, R., Maldonado, J. & Silva, H. (2018) Construction of a highly saturated linkage map in Japanese plum (*Prunus salicina* L.) using GBS for SNP marker calling. *PLoS One*, **13**(12), e0208032. https://doi.org/10.1371/journal.pone.0208032

Chen, H. & Yang, Y. (2014) Genetic diversity and relationship of plum resources in Guizhou analysed by ISSR markers. *Journal of Fruit Science*, **31**, 175–180.

Chen, W., Gong, L., Guo, Z., Wang, W., Zhang, H., Liu, X. *et al.* (2013) A novel integrated method for large-scale detection, identification, and quantification of widely targeted metabolites: application in the study of rice metabolomics. *Molecular Plant*, **6**, 1769–1780.

Chen, S., Zhou, Y., Chen, Y. & Gu, J. (2018) Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

De Bie, T., Cristianini, N., Demuth, J.P. & Hahn, M.W. (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, **22**, 1269–1271.

Delcher, A.L., Phillippy, A., Carlton, J. & Salzberg, S.L. (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research*, **30**, 2478–2483.

Endo, M., Tanigawa, Y., Murakami, T., Araki, T. & Nagatani, A. (2013) Phytochrome-dependent late-flowering accelerates flowering through physical interactions with phytochrome B and CONSTANS. *Proceedings*

of the National Academy of Sciences United States of America, **110**, 18017–18022.

Enright, A.J., Dongen, S.V. & Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, **30**, 1575–1584.

Fang, Z., Wang, K., Dai, H., Zhou, D., Jiang, C., Espley, R.V. *et al.* (2020) The genome of low-chill Chinese plum 'Sanyueli' (*Prunus salicina* Lindl.) provides insights into the regulation of the chilling requirement of flower buds. *bioRxiv*. Available from: http://biorxiv.org/content/early/2020/09/15/2020.07.31.193243[Accessed 20th September 2020].

FAOSTAT. (2019) *Food agriculture organization of the United Nations. Statistical Database*. Available from: http://www.fao.org/faostat/en/#data [Accessed 20th August 2020].

Farcuh, M., Saha, P. & Blumwald, E. (2018) Using the genetic diversity of plum to explore the complexity of fruit ripening. *Acta Horticulture*, **1194**, 1337–1343.

Faust, M. & Suranyi, D. (1999) Origin and dissemination of plum. *Horticultural Reviews*, **23**, 179–231.

Feuillet, C., Schachermayr, G. & Keller, B. (1997) Molecular cloning of a new receptor-like kinase gene encoded at the Lr10 disease resistance locus of wheat. *The Plant Journal*, **11**, 45–52.

Field, B. & Osbourn, A.E. (2008) Metabolic diversification – independent assembly of operon-like gene clusters in different plants. *Science*, **194**, 543–547.

Giannuzzi, G., D'Addabbo, P., Gasparro, M., Martinelli, M., Carelli, F.N., Antonacci, D. *et al.* (2011) Analysis of high-identity segmental duplications in the grapevine genome. *BMC Genomics*, **12**. https://doi.org/10.1186/1471-2164-12-436

Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J *et al.* (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, **36**, 3420–3435.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I. *et al.* (2013) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, **29**, 644–652.

Guo, Z. (2006) *Collection and utilization of Prunus salicina* germplasm resource in South China. Nanjing: Nanjing Agricultural University.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R, Smith Jr, R.K., Hannick, L.I. *et al.* (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, **31**, 5654–5666.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J. *et al.* (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology*, **9**, R7. https://doi.org/10.1186/gb-2008-9-1-r7

Han, Y. & Wessler, S.R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research*, **38**, e199. https://doi.org/10.1093/nar/gkq862

Hartmann, W. & Neumulle, M. (2009) Plum breeding. In: Jain, S.M. & Priyadarshan, P.M. (Eds.) *Breeding plantation tree crops: temperate species*. New York: Springer Science + Business Media, pp. 161–231.

Hicks, K.A., Albertson, T.M. & Wagner, D.R. (2001) *EARLY FLOWERING 3* encodes a novel protein that regulates circadian clock function and flowering in *Arabidopsis*. *The Plant Cell*, **13**, 1281–1292.

Hoffmann, L., Besseau, S., Geoffroy, P., Ritzenthaler, C., Meyer, D., Lapierre, C. *et al.* (2004) Silencing of hydroxycinnamoyl-coenzyme A shikimate/quinate hydroxycinnamoyltransferase affects phenylpropanoid biosynthesis. *The Plant Cell*, **16**, 1446–1465.

Hu, J., Fan, J., Sun, Z., Liu, S. & Berger, B. (2020) NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics*, **36**, 2253–2255.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Research*, **37**, 211–215.

Jiang, F., Zhang, J., Wang, S., Yang, L., Luo, Y., Gao, S. *et al.* (2019) The apricot (*Prunus armeniaca* L.) genome elucidates Rosaceae evolution and beta-carotenoid synthesis. *Horticulture Research*, **6**. https://doi.org/10.1038/s41438-019-0215-6

Jin, J., Zhang, S., Zhao, M., Jing, T., Zhang, N., Wang, J. *et al.* (2020) Scenarios of genes-to-terpenoids network led to the identification of a novel α/β-farnesene/β-ocimene synthase in *Camellia sinensis*. *International Journal of Molecular Sciences*, **21**, 1–14. https://doi.org/10.3390/ijms21020655

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M. *et al.* (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, **24**, 1384–1395.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R. *et al.* (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Research*, **46**, D335–D342.

Kim, D., Langmead, B. & Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, **12**, 357–360.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. & Phillippy, A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome sResearch*, **27**, 722–736.

Lacomme, C. & Roby, D. (1999) Identification of new early markers of the hypersensitive response in *Arabidopsis thaliana*. *FEBS Letters*, **459**, 149–153.

Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.H., Rognes, T. & Ussery, D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, **35**, 3100–3108.

Langmead, B. & Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Li, H. & Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, L., Stoeckert, C.J.J. & Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, **13**, 2178–2189.

Liu, B., Yan, J., Li, W., Yin, L., Li, P., Yu, H. *et al.* (2020a) *Mikania micrantha* genome provides insights into the molecular mechanism of rapid growth. *Nature Communications*, **11**. https://doi.org/10.1038/s41467-019-13926-4

Liu, C., Feng, C., Peng, W., Hao, J., Wang, J., Pan, J. *et al.* (2020b) Chromosome-level draft genome of a diploid plum (*Prunus salicina*). *Gigascience*, **9**. https://doi.org/10.1093/gigascience/giaa130

Liu, W.S., Liu, D.C., Feng, C.J., Zhang, A.M. & Li, S.H. (2006) Genetic diversity and phylogenetic relationships in plum germplasm resources revealed by RAPD markers. *The Journal of Horticultural Science & Biotechnology*, **81**, 242–250.

Liu, W., Liu, D., Zhang, A., Feng, C., Yang, J., Yoon, J. *et al.* (2007) Genetic diversity and phylogenetic relationships among plum germplasm resources in China assessed with inter-simple sequence repeat markers. *Journal of the American Society for Horticultural Science*, **132**, 619–628.

Love, M.I., Huber, W. & Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**. https://doi.org/10.1186/s13059-014-0550-8

Lowe, T.M. & Chan, P.P. (2016) tRNAscan-SE on-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Research*, **44**, W54–W57.

Lyons, E. & Freeling, M. (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal*, **53**, 661–673.

Majoros, W.H., Pertea, M. & Salzberg, S.L. (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.

Martens, S., Teeri, T. & Forkmann, G. (2002) Heterologous expression of dihydroflavonol 4-reductases from various plants. *FEBS Letters*, **531**, 453–458.

Marti, A.F.I., Saski, C.A., Manganaris, G.A., Gasic, K. & Crisosto, C.H. (2018) Genomic sequencing of Japanese plum (*Prunus salicina* Lindl.) mutants provides a new model for rosaceae fruit ripening studies. *Frontiers in Plant Science*, **9**. https://doi.org/10.3389/fpls.2018.00021

Martin, G.B., Bogdanove, A.J. & Sessa, G. (2003) Understanding the functions of plant disease resistance proteins. *Annual Review of Plant Biology*, **54**, 23–61.

Martinoia, E., Klein, M., Geisler, M., Bovet, L., Forestier, C., Kolukisaoglu, Ü. *et al.* (2002) Multifunctionality of plant ABC transporters-more than just detoxifiers. *Planta*, **214**, 345–355.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Nattestad, M. & Schatz, M.C. (2016) Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics*, **32**, 3021–3023.

Nawrocki, E.P. & Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

Owens, D.K., Alerding, A.B., Crosby, K.C., Bandara, A.B., Westwood, J.H. & Winkel, B.S.J. (2008a) Functional analysis of a predicted flavonol synthase gene family in *Arabidopsis*. *Plant Physiology*, **147**, 1046–1061.

Owens, D.K., Crosby, K.C., Runac, J., Howard, B.A. & Winkel, B.S.J. (2008b) Biochemical and genetic characterization of *Arabidopsis* flavanone 3β-hydroxylase. *Plant Physiology and Biochemistry*, **46**, 833–843.

Parra, G., Bradnam, K. & Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. & Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**, 290–295.

Phillips, M.A., Wildung, M.R., Williams, D.C., Hyatt, D.C. & Croteau, R. (2003) cDNA isolation, functional expression, and characterization of (+)-α-pinene synthase and (-)-α-pinene synthase from loblolly pine (*Pinus taeda*): stereocontrol in pinene biosynthesis. *Archives of Biochemistry and Biophysics*, **411**, 267–276.

Pickrell, J.K. & Pritchard, J.K. (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, **8**, e1002967. https://doi.org/10.1371/journal.pgen.1002967

Pinosio, S., Marroni, F., Zuccolo, A., Vitulo, N., Mariette, S., Sonnante, G. *et al.* (2020) A draft genome of sweet cherry (*Prunus avium* L.) reveals genome-wide and local effects of domestication. *The Plant Journal*, **103**, 1420–1432.

Pryszcz, L.P. & Gabaldón, T. (2016) Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, **44**, e113. https://doi.org/10.1093/nar/gkw294

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575.

Putnam, N.H., Butts, T., Ferrier, D.E.K., Furlong, R.F., Hellsten, U., Kawashima, T. *et al.* (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature*, **453**, 1064–1071.

Qu, Y., Easson, M.E.A.M., Simionescu, R., Hajicek, J., Thamm, A.M.K., Salim, V. *et al.* (2018) Solution of the multistep pathway for assembly of corynanthean, strychnos, iboga, and aspidosperma monoterpenoid indole alkaloids from 19E-geissoschizine. *Proceedings of the National Academy of Sciences United States of America*, **115**, 3180–3185.

Ren, R., Wang, H., Guo, C., Zhang, N., Zeng, L., Chen, Y. *et al.* (2018) Widespread whole genome duplications contribute to genome complexity and species diversity in Angiosperms. *Molecular Plant*, **11**, 414–428.

Ruan, J. (2016) *Ultra-fast de novo* assembler using long noisy reads. Available at: https://github.com/ruanjue/smartdenovo [Accessed 29th December 2019].

Rushton, P.J., Somssich, I.E., Ringler, P. & Shen, Q.J. (2010) WRKY transcription factors. *Trends in Plant Science*, **15**, 247–258.

Salamov, A.A. & Solovyev, V.V. (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Research*, **10**, 516–522.

Salazar, J.A., Pacheco, I., Shinya, P., Zapata, P., Silva, C., Aradhya, M. *et al.* (2017) Genotyping by sequencing for SNP-based linkage analysis and identification of QTLs linked to fruit quality traits in Japanese plum (*Prunus salicina* Lindl.). *Frontiers in Plant Science*, **8**. https://doi.org/10.3389/fpls.2017.00476

Salazar, J.A., Pacheco, I., Silva, C., Zapata, P., Shinya, P., Ruiz, D. *et al.* (2019) Development and applicability of GBS approach for genomic studies in Japanese plum (*Prunus salicina* Lindl.). *The Journal of Horticultural Science & Biotechnology*, **94**, 284–294.

Sánchez-Pérez, R., Pavan, S., Mazzeo, R., Moldovan, C., Aiese Cigliano, R., Del Cueto, J. *et al.* (2019) Mutation of a bHLH transcription factor allowed almond domestication. *Science*, **364**, 1095–1098.

Schomburg, F.M., Patton, D.A., Meinke, D.W. & Amasino, R.M. (2001) *FPA*, a gene involved in floral induction in *Arabidopsis*, encodes a protein containing RNA-recognition motifs. *The Plant Cell*, **13**, 1427–1436.

Shi, D.Q., Liu, J., Xiang, Y.H., Ye, D., Sundaresan, V. & Yang, W.C. (2005) *Slow walker1*, essential for gametogenesis in *Arabidopsis*, encodes a WD40 protein involved in 18S ribosomal RNA biogenesis. *The Plant Cell*, **17**, 2340–2354.

Shi, X., Han, X. & Lu, T.-G. (2016) Callose synthesis during reproductive development in monocotyledonous and dicotyledonous plants. *Plant Signaling & Behavior*, **1**. e1062196. https://doi.org/10.1080/15592324.2015.1062196

Shirasawa, K., Isuzugawa, K., Ikenaga, M., Saito, Y., Yamamoto, T., Hirakawa, H. *et al.* (2017) The genome sequence of sweet cherry (*Prunus avium*) for use in genomics-assisted breeding. *DNA Research*, **24**, 499–508.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Singh, N.K., Shukla, P. & Kirti, P.B. (2020) A CBL-interacting protein kinase AdCIPK5 confers salt and osmotic stress tolerance in transgenic tobacco. *Scientific Reports*, **10**, 1–14.

Stacewicz-Sapuntzakis, M. (2013) Dried plums and their products: composition and health effects-an updated review. *Critical Reviews in Food Science and Nutrition*, **53**, 1277–1302.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. (2008) Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics*, **24**, 637–644.

Suzuki, T., Nakajima, S., Morikami, A. & Nakamura, K. (2005) An *Arabidopsis* protein with a novel calcium-binding repeat sequence interacts with TONSOKU/MGOUN3/BRUSHY1 involved in meristem maintenance. *Plant and Cell Physiology*, **46**, 1452–1461.

Tatusov, R.L., Galperin, M.Y., Natale, D.A. & Koonin, E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, **28**, 33–36.

Topp, B., Russell, D., Neumuller, M., Dalbo, M.A. & Liu, W. (2012) Plum. In: Badenes, M. & Byrne, D. *Fruit breeding*. Springer Science Business Media, pp. 571–621.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A. *et al.* (2010) The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nature Genetics*, **42**, 833–839.

Verde, I., Abbott, A.G., Scalabrin, S., Jung, S., Shu, S., Marroni, F. *et al.* (2013) The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics*, **45**, 487–494.

Verde, I., Jenkins, J., Dondini, L., Micali, S., Pagliarani, G., Vendramin, E. *et al.* (2017) The Peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC Genomics*, **18**. https://doi.org/10.1186/s12864-017-3606-9

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. & Birney, E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, **19**, 327–335.

Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. (2010) *KaKs_Calculator* 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, Proteomics & Bioinformatics*, **8**, 77–80.

Wang, K., Li, M. & Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, **38**, e164. https://doi.org/10.1093/nar/gkq603

Wang, X., Zeng, L., Liao, Y., Li, J., Tang, J. & Yang, Z. (2019) Formation of α-farnesene in tea (*Camellia sinensis*) leaves induced by herbivore-derived wounding and its effect on neighboring tea plants. *International Journal of Molecular Sciences*, **20**, 4151. https://doi.org/10.3390/ijms20174151

Wei, X., Shen, F., Zhang, Q., Liu, N., Zhang, Y., Xu, M. *et al.* (2021) Genetic diversity analysis of Chinese plums (*Prunus salicina* L.) based on whole-genome resequencing. *Tree Genetics & Genomes*, **17**. https://doi.org/10.1007/s11295-021-01506-x

Wei, X., Zhang, Q., Liu, N., Zhang, Y., Xu, M., Liu, S. *et al.* (2019) Genetic diversity of the *Prunus salicina* L. from different sources and their related species. *Scientia Agricultura Sinica*, **52**, 568–578.

Wei, X., Zhang, Q. & Liu, W. (2020) Research progress on plum germplasm resources in China. *Acta Horticulturae Sinica*, **47**, 1203–1212.

Wisman, E., Hartmann, U., Sagasser, M., Baumann, E., Palme, K., Hahlbrock, K. *et al.* (1998) Knock-out mutants from an *En-1* mutagenized *Arabidopsis thaliana* population generate phenylpropanoid biosynthesis phenotypes. *Proceedings of the National Academy of Sciences United States of America*, **95**, 12432–12437.

Wu, J., Wang, Z., Shi, Z., Zhang, S., Ming, R., Zhu, S. *et al.* (2013) The genome of the pear (*Pyrus bretschneideri* Rehd.). *Genome Research*, **23**, 396–408.

Wu, T.D. & Watanabe, C.K. (2005) GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

Xiang, Y., Huang, C.H., Hu, Y., Wen, J., Li, S., Yi, T. *et al.* (2017) Evolution of *Rosaceae* fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Molecular Biology and Evolution*, **34**, 262–281.

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S. *et al.* (2014) SOAPdenovo-trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, **30**, 1660–1666.

Xu, Z. & Wang, H. (2007) LTR-FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, **35**, 265–268.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

Yu, X., Zhang, Q., Liu, W., Sun, M., Liu, N., Zhang, Y. *et al.* (2011) Genetic diversity analysis of morphological and agronomic characters of Chinese plum (*Prunus salicina* Lindl.) germplasm. *Journal of Plant Genetic Resources*, **12**, 402–407.

Yuan, H., Zeng, X., Shi, J., Xu, Q., Wang, Y., Jabu, D. *et al.* (2018) Time-course comparative metabolite profiling under osmotic stress in tolerant and sensitive Tibetan hulless barley. *BioMed Research International*, **2018**. https://doi.org/10.1155/2018/9415409

Zhang, C., Dong, S.S., Xu, J.Y., He, W.M. & Yang, T.L. (2019) PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*, **35**, 1786–1788.

Zhang, J. (1990) Report on the investigation for national plum and apricot resources. *China Fruits*, **4**, 29–34.

Zhang, J. & Zhou, E. (1998) *China fruit-plant monographs, plum flora*. Beijing: China Forestry Press.

Zhang, Q., Chen, W., Sun, L., Zhao, F., Huang, B., Yang, W. *et al.* (2012) The genome of *Prunus mume*. *Nature Communications*, **3**. https://doi.org/10.1038/ncomms2290

Zhao, Q., Ma, D., Vasseur, L. & You, M. (2017) Segmental duplications: evolution and impact among the current Lepidoptera genomes. *BMC Evolutionary Biology*, **17**. https://doi.org/10.1186/s12862-017-1007-y

Zhao, Q., Zhu, Z., Kasahara, M., Morishita, S. & Zhang, Z. (2013) Segmental duplications in the silkworm genome. *BMC Genomics*, **14**. https://doi.org/10.1186/1471-2164-14-521

Zuo, L., Han, Z., Liang, H. & Yang, M. (2015) Analysis of genetic diversity of *Prunus salicina* from different producing area by SSR markers. *Acta Horticulturae Sinica*, **42**, 111–118.