GigaScience

The chromosome-level draft genome of a diploid plum (Prunus salicina)

--Manuscript Draft--

Manuscript Number:	GIGA-D-20-00195
Full Title:	The chromosome-level draft genome of a diploid plum (Prunus salicina)
Article Type:	Data Note
Funding Information:	The Industry University Research Collaborative Innovation Major Projects of Guangzhou Science Technology Innovation Commission (201704020021) Dr Yehua He Guangdong Key Laboratory of Innovation Method and Decision Management System (CN) (2016LM1128) Dr Yehua He
Abstract:	<p> <b> Background: </b> Plums are the economically important Rosaceae fruit crops and include dozens of species distributed across the world. They are the most taxonomically diverse within the <i> Prunus </i> genus and hold the center of the <i> Prunus </i> genetic stage <i> . </i> However, limited genomic information is available for the genetic studies and breeding programs of plums. <i> Prunus salicina </i> , a typical diploid plum species, plays a predominant role in modern commercial plums production. Here we selected <i> P. salicina </i> for whole-genome sequencing and presented a chromosome-level genome assembly through the combination of PacBio sequencing, Illumina Sequencing and Hi-C technology. <b> Findings: </b> The assembly had a total size of 284.2 Mb, <a class="ext-link" href="" data-jats-ext-link-type="uri"> with contig N50 of 1.8Mb and scaffold N50 of 32.3Mb </a> . 96.56% of the assembled sequences were anchored onto 8 pseudochromo
Corresponding Author:	Yehua He Soth China Agricultural University Guangzhou, Guangdong Province CHINA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Soth China Agricultural University
Corresponding Author's Secondary Institution:	
First Author:	Chaoyang Liu
First Author Secondary Information:	
Order of Authors:	Chaoyang Liu
	Chao Feng
	Weizhuo Peng
	Jingjing Hao
	Jianjun Pan
	Yehua He
Order of Authors Secondary Information:	
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics	Yes

Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the "Availability of Data and Materials" section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

The chromosome-level draft genome of a diploid plum (*Prunus salicina*)

Chaoyang Liu[1, 3*], Chao Feng[2,*], Weizhuo Peng[1,3], Jingjing Hao[1,3],
Jianjun Pan[4], Yehua He[1,3]

[1] Key Laboratory of Biology and Germplasm Enhancement of Horticultural Crops in South China, Ministry of Agriculture, South China Agricultural University, Guangzhou 510642, China

[2] Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China

[3] Maoming Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Maoming 525000, China

[4] Agricultural Technology Extension Center of Conghua District, Guangzhou. Guangzhou 510900, Guangdong Province, China

*Equal contribution

Corresponding author: Yehua He (email: heyehua@hotmail.com)

Abstract

Background: Plums are the economically important Rosaceae fruit crops and include dozens of species distributed across the world. They are the most taxonomically diverse within the *Prunus* genus and hold the center of the *Prunus* genetic stage. However, limited genomic information is available for the genetic studies and breeding programs of plums. *Prunus salicina,* a typical diploid plum species, plays a predominant role in modern commercial plums production. Here we selected *P. salicina* for whole-genome sequencing and presented a chromosome-level genome assembly through the combination of PacBio sequencing, Illumina Sequencing and Hi-C technology. **Findings:** The assembly had a total size of 284.2 Mb, with contig N50 of 1.8Mb and scaffold N50 of 32.3Mb. 96.56% of the assembled sequences were anchored onto 8 pseudochromosomes and a total of 24,448 protein-coding genes were identified. Phylogenetic analysis showed that *P. salicina* had closer relationship with *P. mume* and *P. armeniaca*, with *P. salicina* diverging from their common ancestor approximately 9.05 million years ago (Mya). 146 gene families were expanded during *P. salicina* evolution, and some cell wall-related GO terms were significantly enriched. It was noteworthy that members in the DUF579 family, a new class involved in xylan biosynthesis, were significantly expanded in *P. salicina,* which provided new insight into the xylan metabolism in plums. **Conclusions:** We constructed the first high-quality chromosome-level plum genome using PacBio, Illumina and Hi-C technologies. This work provides a valuable resource for facilitating plum breeding programs and studying the genetic diversity mechanisms of plums and *Prunus* species.

Background

Plums are the economically important Rosaceae fruit crops and produced throughout the world. About 12.6 million tons of plums (include sloes) are produced per year (FAOSTAT 2018, http://faostat.fao.org/), and the fruits are widely used for fresh consumption and processing like canning and beverages [1].There are 19-40 species of plums distributed across Asia, Europe and America. It is considered that plums hold the center of the *Prunus* genetic stage for the largest diversity of any subgenus and a link between the major subgenera [2].

Prunus salicina, commonly called the Japanese plum or Chinese plum, is an important diploid (2x=2n=16) plum species that predominates in the modern commercial production of plums (Fig. 1). *P. salicina* originate in China and its fruits are mostly used for fresh consumption for their characteristic taste [3]. Cultivars of *P. salicina* have wide variability in phenology, fruit size and shape, flavour, firmness, aroma, texture, phenolic composition, antioxidant activity and both skin and pulp color [4].

However, the genetic and genomic information for *P. salicina* as well as most plum species were scarce [5]. The availability of a fully sequenced and annotated genome will help to measure and characterize the genetic diversity and determine how this diversity relates to the tremendous phenotypic diversity among plum cultivars. The genomic information is essential to support many of the studies involved in fundamental questions about plums biology and genetics. Moreover, genome-based tools could be developed to improve breeding works of plums, which were usually hindered by the high degree of heterozygosity, self-incompatible and long juvenile stage [2, 5, 6].

The fruit firmness, one of the most important indices of plum quality, is closely associated with cell wall compositions [2]. Xylan is a major component of secondary cell walls [7], and the xylan metabolism is involved in various aspects of plant growth and development like fruit ripening and softening [8]. Previous studies showed that the plum species presented more xylose (the main component of xylan) compared to other *Prunus* species, and was also one of the richest natural sources of xyliot [9, 10]. The relatively high levels of xylan-related metabolites implying the possible special mechanism of the xylan metabolism in plum, and the available plum genomic information will be helpful to better understand the mechanism at molecular level.

Genome resources are already available for a number of Rosaceae fruit crops [11], including apple [12-14], peach [15], pear [16-18], strawberry [19, 20], almond [21], black raspberry [22], sweet cherry [23], apricot [24], loquat [25] and *Prunus mume* [26]. However, whole-genome sequencing and chromosome-level assembly for plums have not been reported until now. In this study, a typical diploid plum species, *P. salicina*, was selected for the whole-genome sequencing as a genomic reference. A high-quality chromosome-level *de novo* genome assembly of *P. salicina* was generated using an integrated strategy that combines PacBio sequencing, Illumina sequencing and Hi-C technology. The assembly has a total size of 284.2 Mb with contig N50 of 1.8Mb and scaffold N50 of 32.3 Mb, and vast majority (96.56%) of the assembled sequence was anchored onto 8 pseudochromosomes. The availability of the high-quality chromosome-scale genome sequences not only provides fundamental knowledge regarding plum biology but also presents a valuable resource for genetic diversity analysis and breeding programs of plums and other *Prunus* crops.

Sample collection, library construction and sequencing

P. salicina (*P. salicina* L. cv. ‘Sanyueli’) samples from Guangzhou, Guangdong Province, China (113°22'4" N, 23°9'5" E) were selected for genome sequencing. The samples were kept at the Horticultural Germplasm Conversation Center of South China Agricultural University (SCAU) for breeding and research. Total genomic DNA was extracted from fresh young leaves of 5-year-old *P. salicina* tree using the CTAB method [27].

A combination of PacBio single-molecule real-time (SMRT) sequencing, Illumina’s paired-end sequencing and Hi-C technology was applied. For PacBio sequencing, SMRT libraries were constructed using the PacBio 20-kb protocol (https://www.pacb.com/). A total of ~53.0 Gb long-sequencing reads were generated by PacBio Sequel platform. After removing adaptors within sequences, about 52.9 Gb (169.7 × coverage) subreads were obtained (Table S1). The subreads have a mean length of 13.2 kb (Table S2). The Illumina DNA paired-end libraries were constructed with an insert size of 350 bp, and sequencing was performed on the Illumina HiSeq 4000 platform according to the manufacturer's instructions; a total of approximately 26.6 Gb (85.4 × coverage) short sequencing reads was obtained (Table S1). Reads with adaptors, with unknown bases (N) than 10% and with low-quality bases (≤ 5) more than 50% were filtered out to gain the clean data for further analysis.

The Hi-C library was prepared using the standard procedures. The young leaves of the same *P. salicina* tree were used as starting materials. Nuclear DNA from young leaves was cross-linked in situ, extracted, and digested with DpnII restriction endonuclease. The 5' overhangs of the digested fragments were biotinylated, and the resulting blunt ends were ligated. The cross-links were reversed after ligation, proteins were removed to release the DNA molecules. The purified DNA was sheared to a mean fragment size of 350 bp and ligated to adaptors, followed by purification through biotin-streptavidin-mediated pull down. Finally, the library was sequenced on Illumina HiSeq 4000 platform to produce 59.1 Gb (189.5 × coverage) Hi-C sequencing data (Table S1). After filtering adapter contamination and low-quality reads, a total of 56.1 Gb clean data were obtained for further assisting genome assembly. The quality of Hi-C sequencing was evaluated with HiCUP [28], and the effect rate was approximately 28.10% (Table S3).

In order to generate the RNA-seq data needed for the genome annotation stage, a total of six tissues, including leaf, flower, branch, young fruit pericarp, young fruit pulp and matured fruit, were sampled from the same *P. salicina* tree. Total RNA was extracted from six tissues using E.N.Z.A.® Plant RNA kit (OMEGA). RNA-seq libraries were constructed and sequenced by Illumina Hiseq 4000 in paired-end 150bp mode, and a total of ~46.7 Gb transcriptome data were produced (Table S4).

De novo assembly of the *P. salicina* genome

In the genome assembly process, Illumina sequencing data were used for the genome survey and polishing of preliminary contigs, PacBio long reads were used for contig assembly and Hi-C reads were used for chromosome-level scaffolding.

Sequencing data from the Illumina library were used to estimate the genomic information of *P. salicina* with the k-mer based method. Quality-filtered reads were subjected to 17-mer frequency distribution analysis using SOAPdenovo (SOAPdenovo, RRID: SCR_010752) [29]. Based on the total number of k-mers (19,341,904,177), the estimated *P. salicina* genome size was calculated to be approximately 311.82 Mb (Figure S1). The heterozygous and repeat sequencing ratios were 0.70% and 54.49%, respectively (Table S5).

The *de novo* assembly of the *P. salicina* genome was carried out using the FALCON assembler (FALCON, RRID: SCR_016089) [30], followed by the polishing with Quiver [31] and Pilon (Pilon, RRID: SCR_014731) [32]. The PacBio subreads were subsequently processed by a self-correction of errors using FALCON [30]. Based on the overlap-layout-consensus algorithm, the detection of overlaps among input reads and the assembly for the final string graph [33] were performed using FALCON pipeline [30]. The draft assembly was further polished using Quiver [31]. Finally, the Illumina reads were mapped back to the assembly and the remaining errors were corrected by Pilon [32]. These processes yielded a draft *P. salicina* genome assembly with a total length of 284.2 Mb (Table 1).

Clean Hi-C reads were aligned to the assembled genome with BWA aligner (BWA, RRID: SCR 010910) using default parameters [34]. Only uniquely aligned read pairs whose mapping quality more than 20 were remained for further analysis. Invalid read pairs, including dangling-end and self-cycle, relegation, and dumped products, were filtered by HiCUP [28]. 88.9% of uniquely mapped read pairs were valid interaction pairs (Table S3), which were used to cluster, order, and orient the assembly contigs onto pseudochromosomes by Lachesis (LACHESIS, RRID:SCR_017644) [35]. The Juicebox [36] was applied to build the interaction matrices and complete the visual correction. As shown in Fig. 1, the assembled sequences were anchored onto the 8 pseudochromosomes with lengths ranging from 23.70 to 54.53 Mb (Table S6). The total length of pseudochromosomes accounted for 96.56% of the genome sequences (Figure 1), with contig N50 of 1.78 Mb and scaffold N50 of 32.32 Mb (Table1; Table S7).

Assessing the completeness of the genome assembly

To assess the quality of the genome assembly, the pair-end short reads were aligned to the assembled genome with BWA [34]. The mapping rate was 96.93% and a total of 98.81 % assembled genome was covered by the reads and the mapping coverage with at least 4×, 10×, 20× was 98.48 %, 98.06% and 97.13%, respectively (Table1; Table S8). RNA-seq reads from six tissues of *P. sacilina* were mapped against our assembly using Hisat with default parameters [37], and the percentage of aligned reads ranged from 92.44% to 95.25% (Table1; Table S4). The SNPs were counted to evaluate the accuracy of the genome assembly, a total of 3668 homozygous SNPs were identified, accounting for only 0.0015% of the reference genome (Table S9). The low rate of homozygous SNPs suggested that the assembly had a high base accuracy. The completeness of the assembly was accessed using both Core Eukaryotic Genes Mapping Approach (CEGMA, RRID: SCR_015055) [38] and Benchmarking Universal Single-Copy Orthologs (BUSCO, RRID: SCR_015008) [39] approaches. 234 Core Eukaryotic Genes (CEGs) out of the complete set of 248 CEGs (94.35%) were covered by the assembly, and 229 (92.34%) of these were complete (Table1; Table S10). BUSCO analysis based on single copy orthologs set showed that 95.7% of the expected genes were identified as complete, 1.3% were fragmented, and only 3.0% were missing (Table1; Table S11). These results suggested that the genome assembly was complete and robust.

Annotation

Repeat annotation

To annotate repeat elements in the *P. salicina* genome, a combined strategy based on homology searching and *de novo* prediction was applied. For homology-based prediction, interspersed repeats were identified using RepeatMasker (http://www.repeatmasker.org) (RepeatMasker, RRID: SCR_012954) and RepeatProteinMask (RepeatProteinMask, RRID: SCR 012954) [40] to search against the Repbase database [41]. For *de novo* prediction, RepeatScout (http://www.repeatmasker.org/) (RepeatScout, RRID:SCR 014653) [42], RepeatModeler (http://www.repeatmasker.org/RepeatModeler/) RepeatModeler (RRID:SCR_015027), and LTR_Finder (http://tlife.fudan.edu.cn/tlife/ltr_finder/) (LTR_Finder, RRID:SCR_015247) [43] were used to identify *de novo* involved repeats. Tandem repeats were also *de novo* predicted using Tandem Repeats Finder (TRF) [44]. Overall, the results found that 48.28% of the assembly was covered with transposable elements (TE). Among of them, long terminal repeat (LTR) retrotransposons represented the greatest proportion, making up 42.10% of the genome (Table1; Table S12). Tandem duplicates occurred for 9.8% of the genes and were preferentially enriched in transferase activity and phloem development (Table 1; Figure S2). The TE percentage and density of duplicates resulted from tandem duplications were shown in Figure 1.

Gene annotation

A combination of three approaches, including homology-based prediction, *de novo* prediction and transcriptome-based prediction, was used to predict the protein-coding genes within *P. salicina* genome. For homology-based prediction, the homologous protein sequences of *Prunus persica*, *Prunus avium*, *Prunus mume*, *Pyrus bretschneideri*, *Malus domestica*, *Fragaria vesca* and *Arabidopsis thaliana* were obtained from NCBI database and mapped onto the *P. salicina* genome using TblastN (TBLASTN; RRID:SCR_011822) (E-value ≤ 1e-5) [45], and then the matching proteins were aligned to the homologous genome sequences for accurate spliced alignments with GeneWise (GeneWise, RRID:SCR 015054) [46] to define gene models. For *de novo* prediction, Augustus (Augustus, RRID: SCR_008417) [47], GlimmerHMM (GlimmerHMM, RRID: SCR_002654) [48], SNAP (SNAP, RRID:

SCR 002127) [49], Geneid (GeneID, RRID: SCR 002473) [50] and Genescan (GENSCAN, RRID: SCR_012902) [51] were used to predict the coding regions of genes. For transcriptome-based predictions, RNA-seq data from six tissues were used for genome annotation, processed by HISAT2 (HISAT2, RRID: SCR_015530) [37] and Stringtie (StringTie, RRID: SCR_016323) [52]. RNA-seq data were also *de novo* assembled with Trinity (Trinity, RRID: SCR_013048) [53]. The assembled sequences were aligned against *P. salicina* genome with PASA (Program to Assemble Spliced Alignment, PASA, RRID: SCR_014656) [54], and the effective alignments were assembled to gene structures. Gene models predicted by all of the methods were integrated by EVidenceModeler (EVidenceModeler, RRID: SCR_014659) [54]. To update the gene models, PASA was further used to generate UTRs [54]. Finally, 24,448 non-redundant protein-coding genes with an average transcript size of 2,988.45bp were predicted in the *P. salicina* genome (Table 2), and the gene density was shown in Figure 1.

The functional annotation of protein-coding genes within *P. salicina* genome was carried out by aligning protein sequences against SwisssProt [55] and NR databases using BLASTp (with a threshold of E-value ≤ 1e-5). The protein motifs and domains were annotated by searching against InterPro (InterPro, RRID: SCR 006695) [56] and Pfam (Pfam, RRID: SCR_004726) database [57] with InterProScan (InterProScan, RRID: SCR_005829) [58]. Gene Ontology (GO) terms for each gene were retrieved according to the corresponding InterPro entry. KEGG pathway was mapped by the constructed gene set to identify the best match for each gene [59]. Overall, 23,931 (97.90%) protein-coding genes were successfully annotated (Table 1; Table S13).

Non-coding RNA annotation

The tRNAs were predicted using the program tRNAscan-SE (tRNAscan-SE, RRID: SCR 010835) [60], and rRNA genes were annotated using BLASTN (BLASTN, RRID:SCR_001598) tool with E-value of 1e-5 against rRNA sequences from several relative plant species. miRNA and snRNA were identified by searching against the Rfam (Rfam, RRID:SCR_007891) database [61] with default parameters using the INFERNAL software (INFERNAL, RRID:SCR 011809) [62]. A total of 627 miRNA, 960 tRNA, 273 rRNA and 2023 snRNA in the *P. salicina* genome were finally identified (Table S14).

Gene family identification and phylogenetic analysis of *P. salicina*

OrthoFinder version 2.3.3 (OrthoFinder, RRID:SCR_017118) [63] was used to identify the orthogroups among *P. salicina* and 16 other sequenced rosids, including *P. armeniaca*, *P. mume*, *P. persica*, *P. dulcis*, *P. avium*, *P. yedoensis*, *M. domestica*, *P. bretschneideri*, *Pyrus communis*, *F. vesca*, *Potentilla micrantha*, *Rosa chinensis*, *Rosa multiflora*, *Rubus occidentalis*, *Morus notabilis* and *A. thaliana*. As a result, 15,751 orthogroups containing 23,265 genes were found in *P. salicina*. Moreover, 1,010 genes which were specific to *P. salicina* were identified. A comparison of the predicted proteomes among the 17 species indicated that 9,616, 10,447, 11,098, 13,963 and 15,512 orthogroups were shared between *P. salicina* and Rosids, Rosales, Rosaceae, Amygdaloideae and *Prunus*, respectively.

For phylogenetic construction, proteins of single-copy orthogroups (i.e., the orthogroups which contain none or only one genes for each species) presented in at least 70% of species were selected and aligned with MAFFT version 6.846b (MAFFT, RRID: SCR 011811) [64]. After determination of the best substitution model for each orthogroup with IQ-TREE version 1.7-beta12 (IQ-TREE, RRID: SCR_017254) [65], the maximum likelihood phylogenetic tree across the 17 plant species was constructed using IQ-TREE with the parameter (-p -bb 1000), setting *A. thaliana* as outgroup. The divergence time of each node in the phylogenetic tree was estimated with Bayesian Evolutionary Analysis Sampling Trees (BEAST, RRID: SCR_010228) [66].Two fossil constraints and a secondary calibration node were applied. The fossil *Prunus wutuensis* (age: Early Eocene, minimum age of 55.0 Mya) and the fossil *Rubus acutiformis* (age: Middle Eocene, minimum age of 41.3Mya) were placed at the stem *Prunus* and *Rubus*, respectively [67]. For the secondary calibration node, the divergence of Rosoideae and Amygdaloideae at 100.7 Mya was dated according to Xiang et al. [67]. The Markov chain Monte Carlo was reported 10,000,000 times with

1000 steps. The phylogenetic tree indicated that *P. salicina* diverged from the ancestor of *P. mume* and *P. armeniaca* approximately 9.05 Mya, from the ancestor of *P. persica* and *P.dulcis* 11.12 Mya (Figure 2).

A collinear analysis of the three closely related *Prunus* species (*P. salicina*, *P. armeniaca*, and *P. mume*) was performed using MCScan (minspan=100; MCScan, RRID: SCR_017650; http://chibba.pgml.uga.edu/mcscan2/), and the results showed that the three species exhibited high collinearity (Figure 3). A total of 16,827 and 12,426 *P. salicina* genes were located in collinear blocks between *P. salicina* and *P. armeniaca* and between *P. salicina* and *P. mume*, respectively. Fewer inversions were found in *P. salicina* vs *P. armeniaca* than in *P. salicina* vs *P. mume*.

Gene family expansion and contraction analysis

For gene family expansion analysis, the ancestral gene content of each cluster at each node was investigated with CAFÉ version 3.1 (CAFÉ, RRID: SCR_005983) [68], basing on the phylogeny and gene numbers per orthogroup in each species, the gene family expansions/contractions at each branch were determined with *p*-value < 0.001. The gene family analysis showed that during the evolution of *P. salicina*, 146 gene families were expanded and 500 families were contracted. The functional enrichment on Gene Ontology (GO) of those expanded gene families identified 60 significantly enriched GO terms (p-value < 0.05) (Table S15; Figure S3).

It was noteworthy that genes from the expanded families were enriched in a series of cell wall related processes, such as ‘cell wall polysaccharide metabolic process (GO: 0010383)’, ‘hemicellulose metabolic process (GO: 0010410)’ and ‘regulation of cellular biosynthetic process (GO: 0031326)’. Specially, genes in ‘xylan biosynthetic process (GO: 0045492)’, which corresponded to the DUF579 family [69], were significantly expanded. Further investigation showed that the major copy differences were found in Clade II, which consisted of orthologs of IRX15/IRX15L [69], with seven members in *P. salicina* and only two to four members in other *Prunus* species (Figure 4). It was reported that IRX15 and IRX15L defined a new class of genes involved in xylan biosynthesis [70, 71]. The species-specific expansion of this new subclade might contribute to the relatively high content of xylan-related metabolites (like xylose and xyliot) in plum [9, 10], which provided new insight into the xylan metabolism in plum.

Moreover, the FRS (FAR1-related sequence) gene family, which played multiple roles in a wide range of cellular processes [72], was also significantly expanded in the phylogeny (GO: 000945), and the family expansion may be related to the genetic and phenotypic diversity in *P. salicina*.

The positive selection analysis

The ratios of nonsynonymous to synonymous substitutions (Ka/Ks) were calculated for all the 2,314 single-copy orthologs of the sequenced *Prunus* species using the Codeml program with the free-ratio model as implemented in the PAML (PAML, RRID:SCR_014932) package. A total of 213 positively selected genes (PSGs) were obtained in *P. salicina*, which were enriched in the 'monooxygenase activity (GO: 0004497)' and 'enzyme inhibitor activity (GO: 0004857)' (Figure S4). It was noteworthy that the category 'monooxygenase activity' was also found in the enriched GO terms for the expanded gene families in *P. salicina*, which might provide valuable candidate genes for further functional investigations.

Conclusions

To our knowledge, this is the first report of the chromosome-level genome assembly of plums using Illumina and PacBio sequencing platforms with Hi-C technology. The assembly had a total size of 284.2 Mb, the contig and scaffold N50 reached 1.8 and 32.3 Mb, respectively. A total of 24,448 protein-coding genes were predicted, and 97.9% (23,931 genes) of which have been annotated. Phylogenetic analysis indicated that *P. salicina* was closely related to *P. mume* and *P. armeniaca*, and collinear analysis showed that these three species exhibited high collinearity. Expanded gene families in *P. salicina* were significantly enriched in several cell-wall related processes. Remarkably, the *P. salicina*-specific expansion of the xylan biosynthesis-related DUF579 family provided new insight into the xylan metabolism in plums. Given the economic and evolutionary importance of *P. salicina*, the genomic data in this study offer a valuable resource for facilitating plum breeding programs and studying the genetic basis for agronomic and adaptive divergence of plum and *Prunus* species.

Availability of supporting data and materials

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession WERZ00000000. The version described in this paper is version WERZ01000000. The raw sequencing data are available through the NCBI Sequence Read Archive (SRA) via accession numbers from SRR10233497 to SRR10233505, via the Project PRJNA574159 (Reviewer link: https://dataview.ncbi.nlm.nih.gov/object/PRJNA574159?reviewer=dkism9m6v4lriar12reb0gh59u). The transcriptome data are available through the NCBI SRA (from SRR10235674 to SRR10235679) (Reviewer link: https://dataview.ncbi.nlm.nih.gov/object/PRJNA576011?reviewer=el1jecd7btt3e3kodpc46jiko5). All the annotation tables containing results of an analysis of the draft genome are available at Figshare (https://doi.org/10.6084/m9.figshare.9973469).

Abbreviations

BLAST: Basic Local Alignment Search Tool; BEAST: Bayesian Evolutionary Analysis Sampling Trees; bp: base pair; BUSCO: Benchmarking Universal Single-Copy Orthologs; CEGMA: Core Eukaryotic Genes Mapping Approach; CTAB: cetyltrimethylammonium bromide; EVM: EVidenceModeler; Gb: gigabase pair; GO: Gene Ontology; Hi-C: high-throughput chromosome conformation capture; kb: kilobase pair; KEGG: Kyoto Encyclopedia of Genes and Genomes; Mb: megabase pair; miRNA: microRNA; Mya: million years ago; NCBI: National Center for Biotechnology Information; PacBio: Pacific Biosciences; PAML: phylogenetic analysis by maximum likelihood; PASA: Program to Assemble Spliced Alignments;

RNA-seq: RNA sequencing; rRNA: ribosomal RNA; SMRT: single-molecule real-time; SnRNA，small nuclearRNA; SNP: single-nucleotide polymorphism; TRF: Tandem Repeats Finder; tRNA: transfer RNA.

Funding

This work was financially supported by The Industry University Research Collaborative Innovation Major Projects of Guangzhou Science Technology Innovation Commission (201704020021) and Modern Agricultural Industry Technology System of Guangdong Province (2016LM1128).

Author Contributions

Y.H.H. conceived the study. C.Y.L. and C.F. performed bioinformatics analysis. W.Z.P., J.J.H. and J.J.P. collected the samples and extracted the DNA. C.Y. L. and C. F. wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Figure Legends

Figure 1 The genome and photograph of *Prunus salicina*. Landscape of the *P. salicina* genome, comprising 8 pseudochromosomes that cover ~96.56% of assembly (A); Concentric circles, from outermost to innermost, showing TE percentage (red; B); gene density (green; C); density of duplicates resulted from tandem duplications (blue; D); (E) photograph of *Prunus salicina*.

Figure 2 Evolution of *Prunus salicina* genome and orthogroups. (A) The phylogeny, divergence time and orthogroup expansions/contractions for 17 rosids. The tree was constructed by maximum likelihood method using 341 single copy orthogroups. All nodes have 100% bootstrap support. Divergence time was estimated on a basis of three calibration points (blue circles). Blue bar indicates 95% HPD (highest posterior density) for each node. The numbers in red and green indicate the numbers of orthogroups that have expanded and contracted along particular branches, respectively. (B) The comparison of genes among 17 rosids. The grey bars indicate the genes belonging to 9,616 rosids-shared orthogroups in each of 17 rosids. The grey + green bars indicate the genes belonging to 10,447 rosales-shared orthogroups in each of 16 rosales. The grey + green + pink bars indicate the genes belonging to 11,098 Rosaceae-shared orthogroups in each of 15 Rosaceae. The grey + green + pink + yellow bars indicate the genes belonging to 13,963 rosaceae-shared orthogroups in each of ten Amygdaloideae. The grey + green + pink + yellow + blue bars indicate the genes belonging to 15,512 *Prunus*-shared orthogroups in each of seven *Prunus* species. The red and stripe bars indicate the genes in species-specific orthogroups and unassigned genes, respectively. The white bars indicate the remaining genes for each genome.

Figure 3 Collinear genes located in the pseudochromosomes of *P. salicina*, *P. mume* and *P. armeniaeca*. The inverted regions were highlighted with green color.

Figure 4 The significant expansion of the DUF579 family members in *P. salicina*. (A) Phylogenetic tree of DUF579 proteins of *P. salicina* (red cicle), *P. persica* (hollow inverted triangle), *P. mume* (solid triangle), *P. armeniaeca* (hollow diamond), *P. dulcis* (solid diamond) and *A. thaliana* (solid square). The DUF579 family members were achieved using Pfam PF04669 domain as a query to search against their respective genomes. (B) The summary of the numbers of clade members in DUF579 family.

Additional files

Table S1 Statistics of *P. salicina* genome sequencing data.

Table S2 Statistics of characteristics of PacBio long-read.

Table S3 Statistics of Hi-C sequencing data.

Table S4 Summary of the transcriptome and their mapping rate on the genome assembly.

Table S5 Estimation of the genome size using k-mer analysis.

Table S6 Summary of assembled 8 chromosomes of *P. Salicina.*

Table S7 Summary of the genome assembly of *P. Salicina.*

Table S8 Statistics of mapping ratio in genome.

Table S9 Number and density of SNPs in *P. salicina* genome.

Table S10 Assessment of CEGMA.

Table S11 Summary of BUSCO analysis results according to prediction.

Table S12 Detailed classification of repeat sequences.

Table S13 Statistics of functional annotation.

Table S14 Summary of non-coding RNA.

Table S15 List of the Gene ontology terms significantly enriched in the expanded gene families of *P. salicina*

Figure S1 17-mer frequency distribution in *P. salicina* genome.

Figure S2 Gene ontology enrichment of the tandemly duplicated genes in *P. salicina*.

Figure S3 Gene ontology enrichment of *P. salicina*-expanded genes.

Figure S4 Gene ontology enrichment of the positively selected genes in *P. salicina*.

Table 1 Summary of genome assembly and annotation for *P. salicina*

	Number or percentage
Assembly feature	
Total length of scaffolds (bp)	284,209,110
Number of scaffolds	75
N50 of scaffolds (bp)	32,324,625
Total length of contigs (bp)	284,189,410
Number of contigs	272
N50 of contigs (bp)	1,777,944
Mapping rate by reads from short-insert libraries	96.93%
Core Eukaryotic Genes Mapping Approach (CEGMA) evaluation	94.35%
Benchmarking Universal Single-Copy Orthologs (BUSCO) evaluation	95.70%
RNA-Seq evaluation	92.44-95.25%
Genome annotation	
Percentage of transposable elements (TE)	48.28%
Percentage of long terminal repeat (LTR) retrotransposon	42.10%
No. of predicted protein-coding genes	24,448
No. of genes annotated to public database	23,930 (97.90%)
No. of genes annotated to GO database	13,484 (55.20%)
No. of genes duplicated by tandem duplications	2,384(9.8%)

Table 2 Statistics of predicted protein-coding genes.

Gene set		Number	Average transcript length (bp)	Average CDS length (bp)	Average exons per gene	Average exons length (bp)	Average intron length (bp)
De novo prediction	Augustus	23,592	2,627.71	1167.83	4.80	243.43	384.45
	GlimmerHMM	39,985	5,450.51	747.07	3.14	238.12	2200.59
	SNAP	24,882	2,876.50	728.45	4.22	172.73	667.66
	Geneid	33,780	3,829.40	899.99	4.44	202.74	851.78
	Genscan	21,882	8,251.09	1355.87	6.34	213.98	1292.13
Homolog prediction	*Pyrus bretschneideri*	20,265	3,119.83	1356.17	4.74	286.35	472.06
	Malus domestica	20,010	2,920.17	1361.30	4.65	292.56	426.72
	Prunus mume	23,064	3,038.66	1346.19	4.78	281.67	447.84
	Prunus persica	28,915	2,296.51	1099.56	4.06	270.55	390.64
	Arabidopsis thaliana	28,284	2,071.73	973.28	3.67	265.51	412.07
	Fragaria vesca	22,927	2,994.24	1380.61	4.59	300.66	449.24
	Prunus avium	22,715	3,077.20	1351.28	4.74	284.86	461.03
RNA-seq	PASA	196,264	3,913.86	1008.68	5.16	195.60	698.88
	Transcripts	42,450	11,076.28	2360.92	6.85	344.83	1490.64
EVM		27,981	2,736.70	1061.73	4.57	232.52	469.68
PASA-update*		27,594	2,784.15	1092.82	4.64	235.59	464.83
Final set*		24,448	2,988.45	1157.42	4.97	233.09	461.72

* UTR regions were contained

References

1. Roussos PA, Efstathios N, Intidhar B, Denaxa N-K and Tsafouros A. Plum (*Prunus domestica* L. and *P. salicina* Lindl.). In: Monique Simmonds VRP, editor. Nutritional Composition of Fruit Cultivars. Elsevier; 2016. p. 639 - 666.
2. Topp BL, Russell DM, Neumüller M, Dalbó MA and Liu W. Plum. In: Maria Luisa Badenes DHB, editor. Fruit Breeding. Springer; 2012. p. 571-621.
3. Hartmann W and Neumüller M. Plum breeding. In: Shri Mohan Jain PMP, editor. Breeding Plantation Tree Crops: Temperate Species. Springer; 2009. p. 161-231.
4. Okie W and Hancock J. Plums. In: Hancock JF, editor. Temperate Fruit Crop Breeding. Springer Science & Business Media; 2008. p. 337-358.
5. Esmenjaud D and Dirlewanger E. Plum. In: Kole C, editor. Genome Mapping and Molecular Breeding in Plants. Springer; 2007. p. 119-135.
6. Guerra M and Rodrigo J. Japanese plum pollination: A review. SCI Hortic-Amsterdam 2015;**197**:674-686.
7. Rennie EA and Scheller HV. Xylan biosynthesis. Curr Opin Biotech 2014;**26**:100-107.
8. Brummell DA and Schröder R. Xylan metabolism in primary cell walls. NZ J Forestry Sci. 2009;**39**:125-143.
9. Renard CMGC and Ginies C. Comparison of the cell wall composition for flesh and skin from five different plums. Food Chem 2009;**114**(3):1042-1049.
10. Arcaño YD, García ODV, Mandelli D, Carvalho WA and Pontes LAM. Xylitol: A review on the progress and challenges of its production by chemical route. Catal Today 2020;**344**:2-14.
11. Aranzana MJ, Decroocq V, Dirlewanger E, Eduardo I, Gao ZS, Gasic K, et al. *Prunus* genetics and applications after de novo genome sequencing: achievements and prospects. Hortic Res 2019;**6** (1):1-25.
12. Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, et al. The genome of the domesticated apple (*Malus*× *domestica* Borkh.). Nat Genet 2010;**42** (10):833-839.

13. Chen X, Li S, Zhang D, Han M, Jin X, Zhao C, et al. Sequencing of a wild apple (*Malus baccata*) genome unravels the differences between cultivated and wild apple species regarding disease resistance and cold tolerance. G3: Genes, Genomes, Genet 2019;**9** (7):2051-2060.

14. Zhang L, Hu J, Han X, Li J, Gao Y, Richards CM, et al. A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. Nat Commun 2019;**10** (1):1-13.

15. Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, et al. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet 2013;**45**(5):487-494.

16. Wu J, Wang Z, Shi Z, Zhang S, Ming R, Zhu S, et al. The genome of the pear (*Pyrus bretschneideri* Rehd.). Genome Res 2013;**23**(2):396-408.

17. Chagné D, Crowhurst RN, Pindo M, Thrimawithana A, Deng C, Ireland H, et al. The draft genome sequence of European pear (*Pyrus communis* L.‘Bartlett’). PloS One 2014;**9** (4):e92644.

18. Dong X, Wang Z, Tian L, Zhang Y, Qi D, Huo H, et al. *De novo* assembly of a wild pear (*Pyrus betuleafolia*) genome. Plant Biotechnol J 2020;**18**(2):581-595

19. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, et al. The genome of woodland strawberry (*Fragaria vesca*). Nat Genet 2011;**43**(2):109-116.

20. Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, McKain MR, et al. Origin and evolution of the octoploid strawberry genome. Nat Genet 2019;**51**(3):541-547.

21. Alioto T, Alexiou KG, Bardil A, Barteri F, Castanera R, Cruz F, et al. Transposons played a major role in the diversification between the closely related almond and peach genomes: Results from the almond genome sequence. Plant J 2020;**101**(2):455-472.

22. VanBuren R, Bryant D, Bushakra JM, Vining KJ, Edger PP, Rowley ER, et al. The genome of black raspberry (*Rubus occidentalis*). Plant J 2016;**87**(6):535-547.

23. Shirasawa K, Isuzugawa K, Ikenaga M, Saito Y, Yamamoto T, Hirakawa H, et al.

The genome sequence of sweet cherry (*Prunus avium*) for use in genomics-assisted breeding. DNA Res 2017;**24**(5):499-508.

24. Jiang S, An H, Xu F and Zhang X. Chromosome-level genome assembly and annotation of the loquat (*Eriobotrya japonica*) genome. GigaScience 2020;**9**(3):giaa015.

25. Jiang F, Zhang J, Wang S, Yang L, Luo Y, Gao S, et al. The apricot (*Prunus armeniaca* L.) genome elucidates Rosaceae evolution and beta-carotenoid synthesis. Hortic Res 2019;6 (1):1-12.

26. Zhang Q, Chen W, Sun L, Zhao F, Huang B, Yang W, et al. The genome of *Prunus mume*. Nat Commun 2012;**3**:1318.

27. Lodhi MA, Ye G-N, Weeden NF and Reisch BI. A simple and efficient method for DNA extraction from grapevine cultivars and *Vitis* species. Plant Mol Biol Rep. 1994;**12** (1):6-13.

28. Wingett S, Ewels P, Furlan-Magaril M, Nagano T, Schoenfelder S, Fraser P, et al. HiCUP: pipeline for mapping and processing Hi-C data. F1000Res 2015;**4**:1310.

29. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience 2012;**1** (1):2047-217X-1-18.

30. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 2016;**13** (12):1050-1054.

31. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 2013;**10**(6):563-569.

32. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One 2014;**9** (11):e112963.

33. Myers EW. The fragment assembly string graph. Bioinformatics. 2005;**21** (suppl_2):ii79-ii85.

34. Li H and Durbin R. Fast and accurate short read alignment with

Burrows-Wheeler transform. Bioinformatics 2009;**25** (14):1754-1760.

35. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO and Shendure J. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. Nat Biotechnol 2013;**31** (12):1119-1125.

36. Robinson JT, Turner D, Durand NC, Thorvaldsdottir H, Mesirov JP and Aiden EL. Juicebox. js provides a cloud-based visualization system for Hi-C data. Cell Syst 2018;**6**(2):256-258. e1.

37. Kim D, Langmead B and Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 2015;**12** (4):357-360.

38. Parra G, Bradnam K and Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 2007;**23** (9):1061-1067.

39. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;**31**(19):3210-3212.

40. Tarailo‐Graovac M and Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinf 2009;**25** (1):4.10. 1-4.10. 4.

41. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O and Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 2005;**110** (1-4):462-467.

42. Price AL, Jones NC and Pevzner PA. *De novo* identification of repeat families in large genomes. Bioinformatics 2005;**21** (suppl_1):i351-i358.

43. Xu Z and Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 2007;**35** (suppl_2):W265-W268.

44. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999;**27** (2):573-580.

45. Gertz EM, Yu Y-K, Agarwala R, Schäffer AA and Altschul SF. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol 2006;**4** (1):1-14.

46. Birney E, Clamp M and Durbin R. GeneWise and genomewise. Genome Res

2004;**14** (5):988-995.

47. Stanke M, Steinkamp R, Waack S and Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 2004;**32** (suppl_2):W309-W312.

48. Majoros WH, Pertea M and Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 2004;**20** (16):2878-2879.

49. Korf I. Gene finding in novel genomes. BMC Bioinf 2004;**5** (1):59.

50. Blanco E, Parra G and Guigó R. Using geneid to identify genes. Curr Protoc Bioinf 2007;**18** (1):4.3. 1-4.3. 28.

51. Burge C and Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997;**268** (1):78-94.

52. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT and Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 2015;**33**(3):290-295.

53. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 2013;**8** (8):1494-1512.

54. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 2008;**9** (1):R7.

55. Bairoch A and Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;**28** (1):45-48.

56. Mulder N and Apweiler R. InterPro and InterProScan: tools for protein sequence classifcation and comparison. Methods Mol Biol 2007; **396**:59-70.

57. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res 2013;**42** (D1):D222-D230.

58. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 2014;**30** (9):1236-1240.

59. Kanehisa M and Goto S. KEGG: kyoto encyclopedia of genes and genomes.

Nucleic Acids Res 2000;**28** (1):27-30.

60. Lowe TM and Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 1997;**25** (5):955-964.

61. Griffiths-Jones S, Bateman A, Marshall M, Khanna A and Eddy SR. Rfam: an RNA family database. Nucleic Acids Res 2003;**31**(1):439-441.

62. Nawrocki EP and Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 2013;**29** (22):2933-2935.

63. Emms DM and Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 2015;**16** (1):157.

64. Katoh K and Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013;**30** (4):772-780.

65. Nguyen L-T, Schmidt HA, Von Haeseler A and Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 2015;**32** (1):268-274.

66. Drummond AJ and Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 2007;**7** (1):1-8.

67. Xiang Y, Huang C-H, Hu Y, Wen J, Li S, Yi T, et al. Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. Mol Biol Evol 2017;**34** (2):262-281.

68. De Bie T, Cristianini N, Demuth JP and Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics 2006;**22** (10):1269-1271.

69. Temple H, Mortimer JC, Tryfona T, Yu X, Lopez ‐ Hernandez F, Sorieul M, et al. Two members of the DUF 579 family are responsible for arabinogalactan methylation in Arabidopsis. Plant Direct 2019;**3** (2):e00117.

70. Jensen JK, Kim H, Cocuron JC, Orler R, Ralph J and Wilkerson CG. The DUF579 domain containing proteins IRX15 and IRX15-L affect xylan synthesis in Arabidopsis. Plant J 2011;**66** (3):387-400.

71. Brown D, Wightman R, Zhang Z, Gomez LD, Atanassov I, Bukowski JP, et al. Arabidopsis genes IRREGULAR XYLEM (IRX15) and IRX15L encode DUF579‐containing proteins that are essential for normal xylan deposition in the secondary cell wall. Plant J 2011;**66** (3):401-413.

72. Ma L and Li G. FAR1-related sequence (FRS) and FRS-related factor (FRF) family proteins in *Arabidopsis* growth and development. Front Plant Sci 2018; **9**:692.

Figure 1

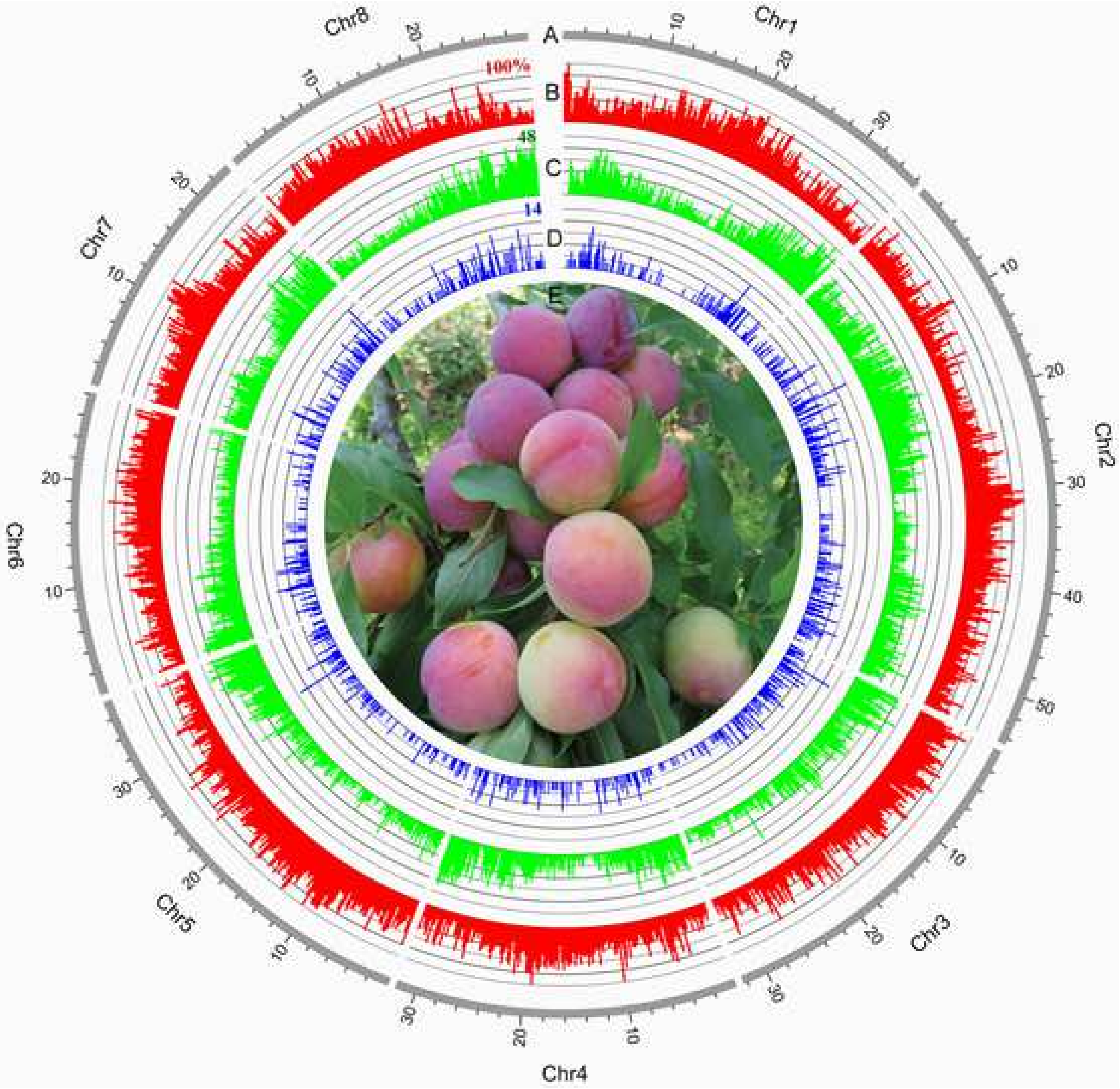

A
B
C
D
E
100%
48
14
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
10
20
30
40
50

Figure 2

Click here to access/download;Figure;Figure 2-ok.tif

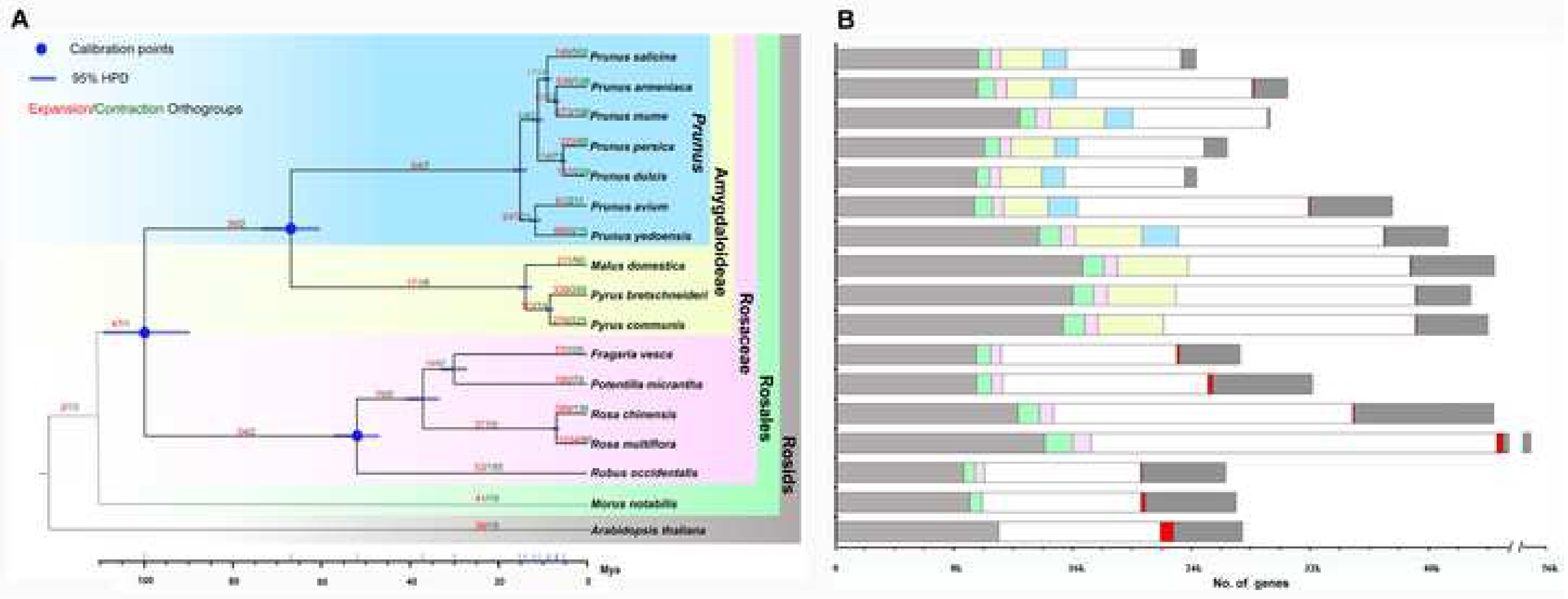


Figure 3

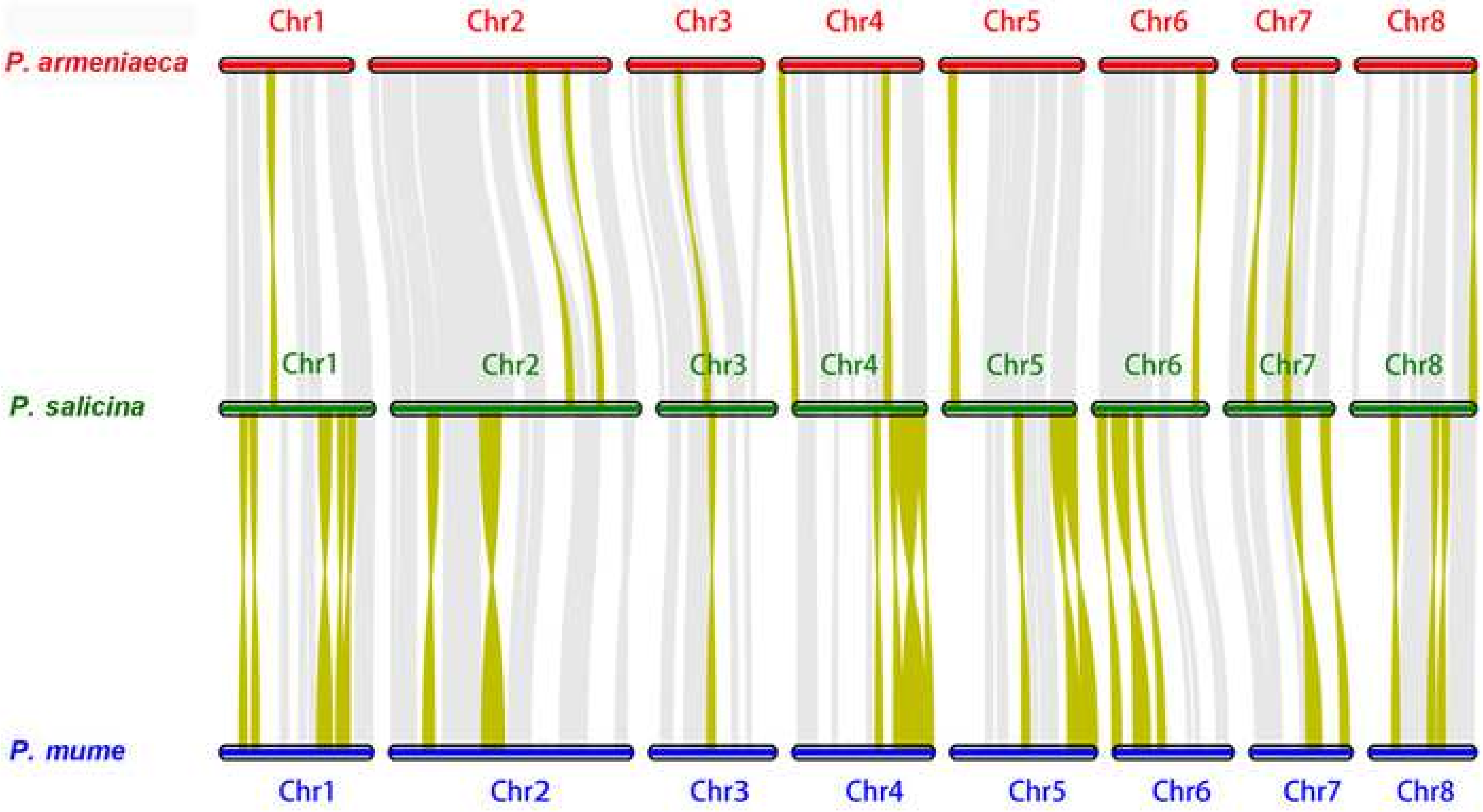
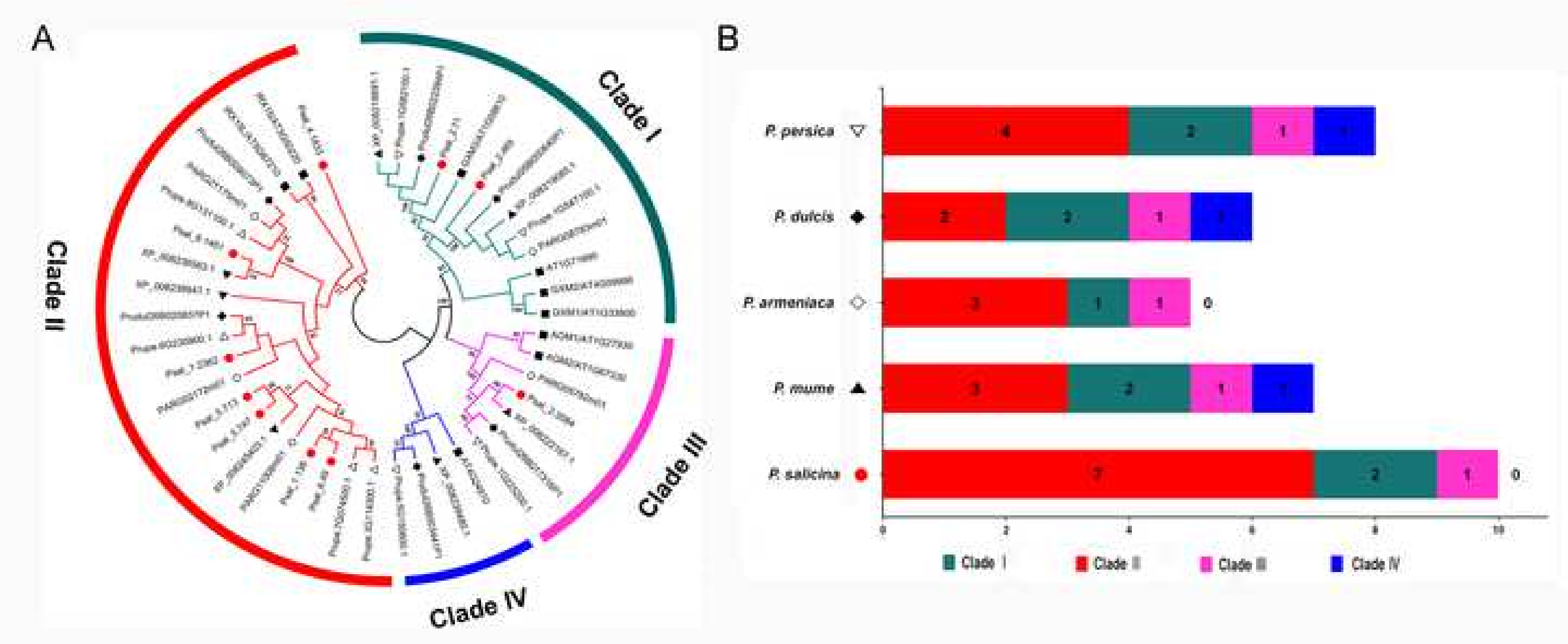

Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
P. armeniaeca
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
P. salicina
P. mume
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8

Figure 4


A
Clade I
Clade II
Clade III
Clade IV
B
P. persica ▽
P. dulcis ◆
P. armeniaca ◇
P. mume ▲
P. salicina ●
4
2
1
1
2
2
1
1
3
1
1
0
3
2
1
1
7
2
1
0
0
2
4
6
8
10
Clade Ⅰ
Clade Ⅱ
Clade Ⅲ
Clade Ⅳ

Supplementary Tables

Supplementary Figures