scientific data

OPEN

DATA DESCRIPTOR

The chromosome-level genome assembly of *Prunus cerasifera* 'Atropurpurea'

Hanlei Li[1,2], Linchang Dong[1], Mingjun Zhang[1,2], Jinrong Li[1,2], Yi Zheng[1,2] ✉, Qinghua Dong[1] ✉ & Sen Wang[1,2] ✉

Prunus cerasifera **'Atropurpurea' (Purpleleaf Plum), known for its unique purple-red foliage, is an important ornamental plant that enhances the aesthetic value of urban greening. To explore the molecular mechanisms underlying leaf color changes, this study assembled the Purpleleaf Plum genome, providing new insights for related research. We used HiFi sequencing data to assemble its genome. After chromosome anchoring, the final genome size was 244.89 Mb, with a contig N50 of 26.60 Mb, and approximately 97.10% of sequences were anchored to 8 chromosomes. Genome annotation identified 28,231 protein-coding genes, with LTR transposons comprising 27.93% of the genome. BUSCO assessment revealed a genome completeness of 98.9%. Telomeric repeat analysis identified 14 telomeres, with six chromosomes capped by double telomeres and two chromosomes containing a single telomere. This high-quality Purpleleaf Plum genome provides a solid foundation for future gene function analysis, cultivar improvement, and genetic research, offering valuable resources for related fields.**

Background & Summary

Prunus cerasifera 'Atropurpurea', commonly known as Purpleleaf Plum, is a bud sport of *P. cerasifera* characterized by high anthocyanin accumulation in its leaves, resulting in purple or red foliage. The earliest known cultivar was introduced from Iran to Europe in the 19th century and has since been widely used in landscaping for its year-round purple-red foliage[1–3]. However, issues like leaf color fading and 're-greening' over time reduce its ornamental value and limit its application[4]. To address these issues, researchers have begun to explore the molecular mechanisms underlying the color changes in the leaves of Purpleleaf Plum, along with the related genes and regulatory pathways, in order to provide a theoretical basis for the improvement of Purpleleaf Plum varieties and broader landscaping applications.

The purple coloration of Purpleleaf Plum leaves is primarily determined by the proportion and distribution of chlorophyll, carotenoids, and anthocyanins, with anthocyanins playing a key role[5,6]. High anthocyanin content results in bright purple or red leaves, while reduced content or disrupted synthesis leads to green or darker hues. Genes such as *ANS*, *C4H*, and *CHS* are involved in anthocyanin biosynthesis and are regulated by transcription factors such as MYB[7], bHLH[8], and WD40[9]. Chlorophyll biosynthesis also influences leaf color stability, with genes like *HEMA1* and *CAO* contributing to pigment variation[10].

Through genome assembly and related analyses, such as gene co-expression network analysis, comparative genomics, and gene-metabolite correlation networks, key genes related to anthocyanin biosynthesis, chloroplast development, and leaf color transition have been identified and studied in species such as *Quercus dentata*[11], *Ananas comosus* var. *bracteatus* f. *tricolor*[12], and *Acer pseudosieboldianum*[13]. These studies revealed the key genes and regulatory mechanisms closely associated with leaf color changes. Therefore, this study provides new data references for leaf color research through the genome assembly of Purpleleaf Plum, laying the foundation for further exploration of the molecular mechanisms underlying leaf color formation in this species. It also provides a new research basis for breeding varieties with stable leaf color and resistance to senescence.

We performed *de novo* assembly of the Purpleleaf Plum genome using PacBio HiFi sequencing, obtaining approximately 16.22 Gb of high-quality reads with an average length of 16,508 bp. K-mer analysis estimated the

[1]Beijing Key Laboratory for Agriculture Application and New Technique, College of Plant Science and Technology, Beijing University of Agriculture, Beijing, 102206, China. [2]Bioinformatics Center, Beijing University of Agriculture, Beijing, 102206, China. ✉e-mail: yz@moilab.net; dongbac@126.com; wangsen10@mails.ucas.ac.cn

Assembled contigs size (Mb)	252.20
Contig N50(Mb)/L50	26.60/4
Contig N90(Mb)/L90	15.13/9
Chromosome genome(Mb)	244.89
Chromosome number	8
Unanchored size(Mb)	7.31
Anchoring rate(%)	97.10
Complete Busco(%)	98.90
Complete and single-copy BUSCOs (C)	96.80
Complete and duplicated BUSCOs (S)	2.10

Table 1. Genome assembly of Purpleleaf Plum.

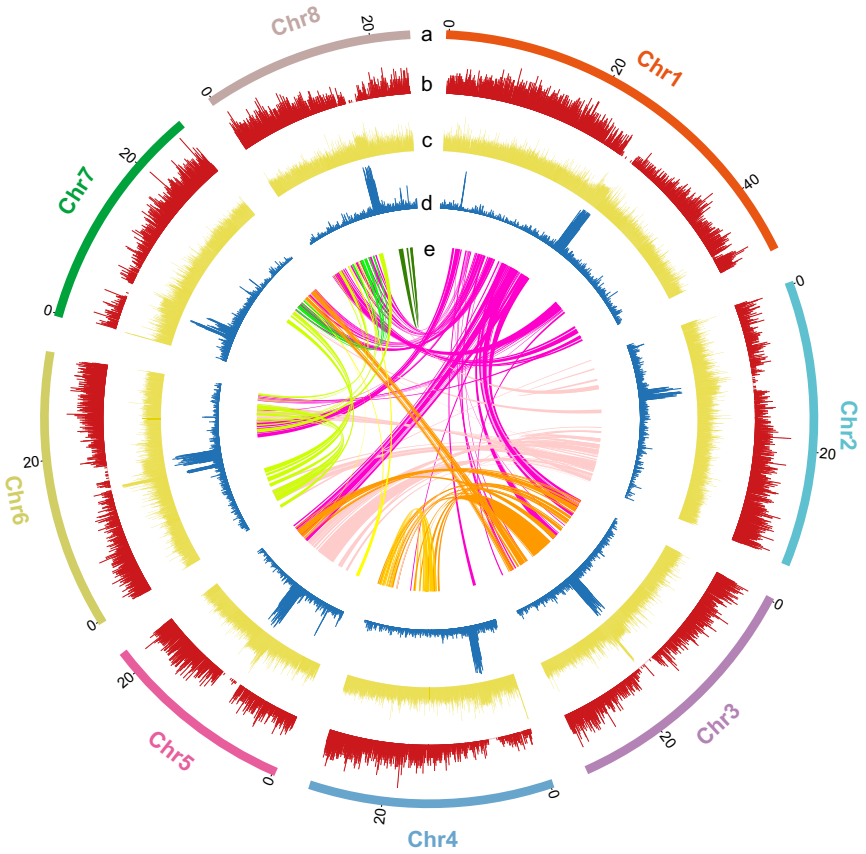

Fig. 1 Genome-wide circos map in Purpleleaf Plum. (**a**) Chromosome ID and length. (**b**) Density of protein-coding genes. (**c**) Density of LTR elements. (**d**) GC content. (**e**) Paralog synteny relationships.

genome size to be approximately 242 Mb, with a sequencing depth of around 65×. The assembled genome size was 252.20 Mb, with a contig N50 of 26.60 Mb. After chromosome anchoring, the final genome size was determined to be 244.89 Mb, with an anchoring rate of 97.10% (Table 1). The assembly quality was assessed using BUSCO, which indicated 98.9% completeness. Repeat annotation revealed that the genome has a high level of repetitiveness, with LTR transposons accounting for 27.93%, DNA transposons constituting 11.86%, and a GC content of 38.32%. The final annotation identified 28,231 protein-coding genes (Fig. 1), and BUSCO evaluation of the annotation showed a completeness score of 98.6%.

Methods

Plant materials. PacBio HiFi sequencing: Fresh young leaves were collected from a 3-year-old Purpleleaf Plum tree (Fig. 2a) at the Fruit Tree Germplasm Repository, Beijing University of Agriculture, Beijing, China (36°49′N, 128°37′E).

RNA-seq: Fresh mature leaves were collected from three distinct phenotypes: Red-leaf phenotype, Green-leaf phenotype, and Purple-leaf phenotype.

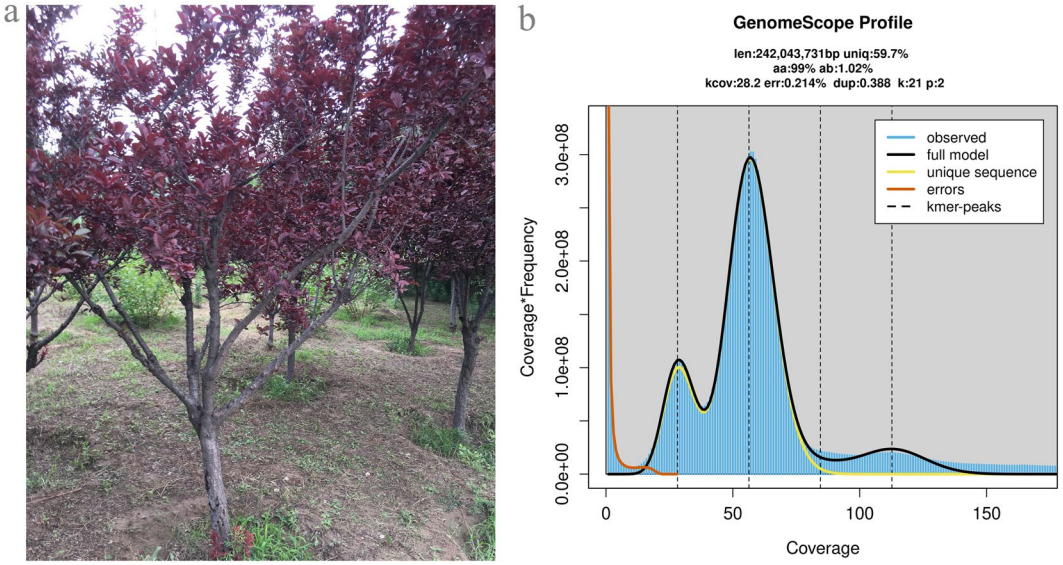

Fig. 2 Tree and genome complexity of Purpleleaf Plum. (**a**) A photograph of the Purpleleaf Plum tree in the feld. (**b**) Genome survey results based on K-mer analysis. Genome size and heterozygosity rate were estimated using GenomeScope2.

DNA extraction and genome sequencing. Genomic DNA (gDNA) was extracted from Purpleleaf Plum leaf tissues using the cetyltrimethylammonium bromide (CTAB) method. DNA purity and concentration were assessed using Qubit and agarose gel electrophoresis. High quality gDNA (≥10 μg, ≥100 ng/μl) was prepared for library construction.

Pulsed-field electrophoresis was used to evaluate 10 μg gDNA, most of DNA fragments should be longer than 40 Kb. The gDNA was subsequently sheared to ~15 kb fragments using a Megaruptor (Diagenode) to obtain appropriately sized fragments for library construction. Sheared DNA was purified and concentrated using AMPure® PB Beads (PacBio, 100-265-900) to remove small fragments and contaminants. The DNA ends were then subjected to damage repair, end repair, and A-tailing to ensure molecular integrity for downstream ligation. Hairpin adapters were ligated to the prepared DNA to form SMRTbell templates. Enzymatic digestion was performed using the SMRTbell® Express Template Prep Kit 2.0 (Pacific Biosciences, PN 101-853-100) to remove unligated or damaged DNA molecules. Size selection was carried out using the BluePippin (Sage Science) system to enrich DNA fragments of the desired length and enhance sequencing quality. The resulting library was sequenced on the PacBio Sequel II platform. A 30-hour sequencing run yielded approximately 16.22 Gb of high-quality, long-read data for downstream genomic analyses.

RNA extraction, library preparation, and sequencing. Leaf samples were collected from Purpleleaf Plum. Total RNA was extracted using the standard TRIzol reagent protocol. RNA integrity was assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA). After confirming RNA quality, 3 μg of total RNA was used for transcriptome library construction. The RNA concentration of the library was initially quantified using the Qubit® RNA Assay Kit on a Qubit® 3.0 Fluorometer and diluted to 1 ng/μl. Sequencing was performed on the Illumina platform, generating approximately 58.36 Gb of paired-end reads.

Genome survey. The genome size and complexity of Purpleleaf Plum were assessed using a k-mer-based approach with HiFi reads. K-mer counting was performed by Jellyfish (v2.2.10)[14] with the parameters "count -t 32 -C -s 3 G -m 21", followed by genome profiling using Genomescope (v2.0)[15] with the parameters "-k 21 -p 2" to estimate genome size, heterozygosity rate, and repetitive sequence content. The genome was estimated at 242 Mb, with 40.3% repetitive sequences and 1.02% heterozygosity (Fig. 2b).

Genome assembly. High-quality HiFi reads were assembled into primary contigs by hifiasm (v0.16.1)[16] with the parameters "–primary -t 64". Subsequently, haplotypic duplication sequences were removed using purge_dups (v1.2.5) pipeline with the default parameters. To eliminate plastid sequences, the contigs were aligned to the mitochondrial (OK563724.1) and chloroplast (MZ128910.1) genomes using MUMmer (v 4.0.0rc1)[17] with the default parameters for plastid sequence removal, resulting in a total of 43 contigs. The final assembly spanned 252.20 Mb with a N90 of 15.13 Mb (L90 = 9). To resolve contig direction and assess telomere-to-telomere (T2T) completeness for chromosome-level assembly, telomeric sequence were identified. Tandem repeats were detected in contigs using Tandem Repeats Finder (v4.09.1)[18] with the parameters "2 7 7 80 10 50 500 -m -f -d". Contig termini containing the telomeric repeats $(TTTAGGG)_n$ (>1000 bp) were defined as telomeric region, corresponding to chromosome ends. This analysis showed that four contigs possessed telomeric regions at both ends, indicating that they represent complete T2T chromosomes (Fig. 3a). The remaining contigs, six of which contained telomeric regions, were then scaffolded using a reference-based approach. Using MUMmer with default parameters, the contig sequences were aligned against the two reference genomes, *P. salicina* 'Sanyueli' (GCA_019277915.1)[19]

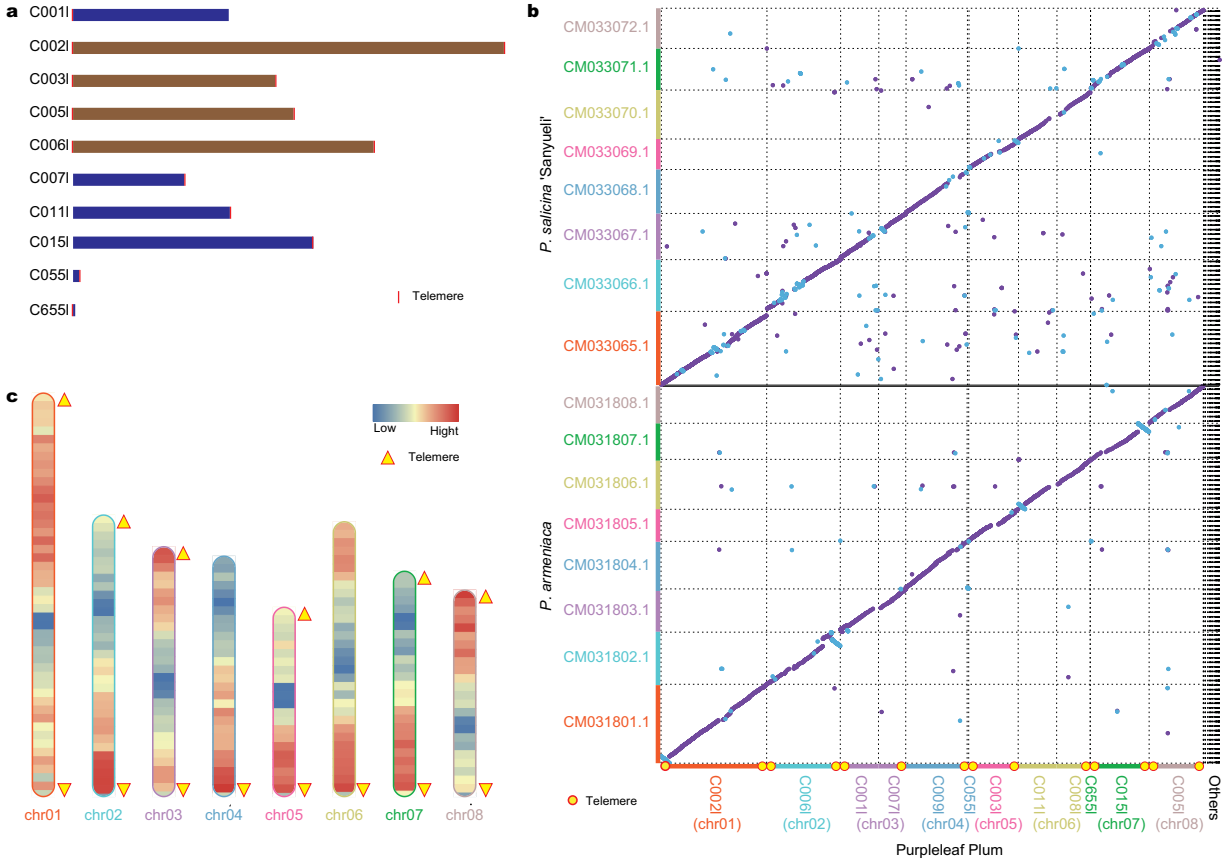

Fig. 3 Telomere and sequence collinearity. (**a**) Telomere analysis revealed ten contigs containing telomeric regions, with four contigs (brown) exhibiting terminal regions at both ends, corresponding to complete T2T chromosomes. (**b**) Sequence synteny between the contigs of Purpleleaf Plum and the two reference genomes of *P. armeniaca* and *P. salicina* 'Sanyueli'. (**c**) Chromosome telomere and gene density map of Purpleleaf Plum. Red and blue represent high-density and low-density regions of genes, respectively.

and *P. armeniaca* (GCA_018524995.1)[20]. Based on genomic synteny and contig orientation, eight contigs were anchored to four chromosomes (Fig. 3b). Among these, two chromosomes harbored terminal telomeres at both ends, while the other two contained telomeric repeats at only one end (Fig. 3c).

Genome annotation. The annotation of repetitive sequences was performed using a *de novo*[16] annotation approach. First, RepeatModeler (v2.0.1)[21] was used to build a repetitive sequence database with the parameters "-database -pa 5 -LTRStruct". Subsequently, EDTA[22] was employed for transposable element annotation, processing the genome to generate an initial transposable element library. Next, the classified repetitive element sequences generated by RepeatModeler were merged with those from EDTA, followed by further removal of redundancy, resulting in a non-redundant repetitive sequence library. Finally, RepeatMasker (v4.1.2)[23], with the parameters "-e rmblast -pa 36 -lib", was used to identify transposable elements and low-complexity repetitive sequences within the genome. The results showed that the genome has a high level of repetitiveness, with LTR accounting for 27.93% and DNA transposons constituting 11.86% (Table 2).

After masking repetitive sequences, protein-coding gene prediction was performed based on integration strategy, involving transcript-based gene prediction, homology-based gene prediction and ab initio gene prediction. (1) Transcript-based gene prediction. HISAT2 (v2.2)[24] with the parameters "-p 48" was used to align RNA-seq reads to the genome, producing alignment results. Next, StringTie (v1.3)[25] with the parameters "-l -p 32 -o -A" was used to assemble the transcripts from the alignment results and TransDecoder was used to generate ORF region. (2) Genome Threader[26] with default parameters was employed to perform structural annotation using the homologous proteins from *P. salicina* 'Sanyueli' and *P. armeniaca*, predicting the protein-coding regions of genes and determining their protein structures. (3) Helixer (https://www.plabipd.de/helixer_main.html) was used for online annotation, further enhancing the annotation information. Finally, manual filtering and correction were performed using Maker (v3.1)[27] to remove redundant data, resulting in non-redundant annotation results. We ultimately annotated a total of 28,231 genes.

Functional annotation of the protein-coding genes involved multiple tools to ensure comprehensive analysis. Protein sequences were aligned against the public databases (NR, Swiss-Prot[28], TrEMBL, and Arabidopsis) using Diamond (v2.0)[29] with an E-value cutoff of 1E-3, and 27,831, 22,249, 25,189, and 27,808 genes were annotated, respectively. InterProScan (v5.59)[30] was used to compare the protein sequences against the InterPro database,

Type	Number	Length(bp)	Percentage of Genome(%)
Retroelements	88,890	72,919,331	28.91
SINEs	3,515	257,177	0.10
LINEs	4,474	2,213,807	0.88
LTR elements	80,901	70,448,347	27.93
Copia	22,416	13,397,859	5.31
Gypsy	38,913	43,333,659	17.18
DNA transposons	122,811	29,910,989	11.86
Unclassified	70,368	12,994,647	5.15
Total interspersed		115,824,967	45.93

Table 2. Summary statistics of transposable elements annotated in the Purpleleaf Plum genome.

	Gene number	Percent (%)
GenBank-NR	27,831	98.58
A. thaliana	25,189	89.22
SwissProt	22,249	78.81
TrEMBL	27,808	98.50
KEGG	12,656	44.83
GO	20,125	71.29
Pfam	22,218	78.70
TF/TR	1,908	6.76
PK	1,243	4.40

Table 3. Functional annotation of protein-coding genes in the Purpleleaf Plum genome.

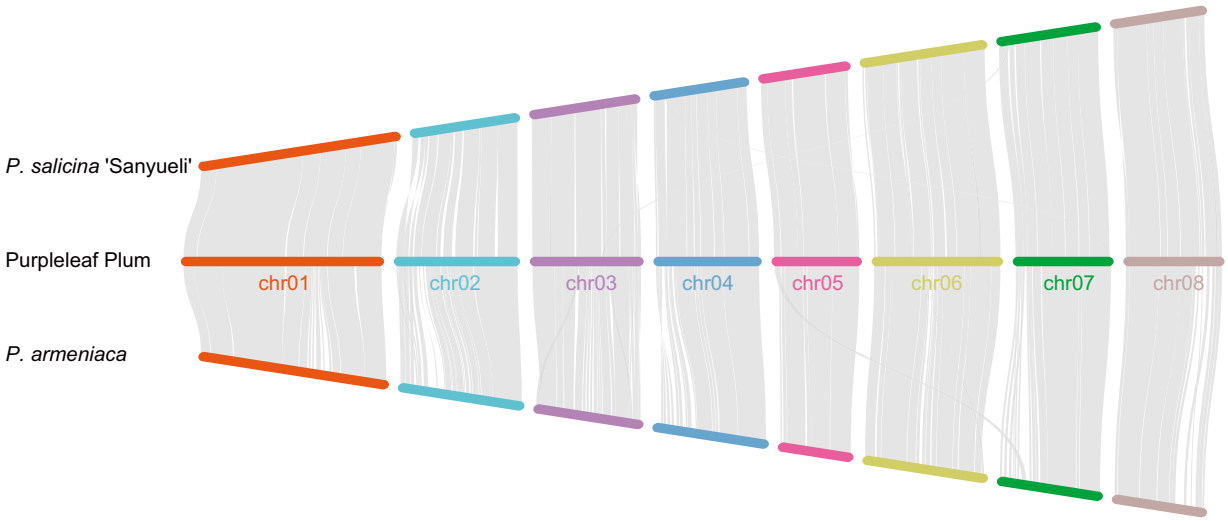

Fig. 4 Collinearity between Purpleleaf Plum and the reference genomes of *P. armeniaca* and *P. salicina* 'Sanyueli'. Synteny analysis was performed using JCVI to identify conservation of homologous genes and gene order.

and protein functional domains and gene ontology (GO)[31] terms were identified. A total of 22,218 proteins contained Pfam domains, and 20,125 were assigned to the GO term. KEGG[32] pathway analysis linked 12,656 genes to known biochemical pathways, while iTAK[33] identified 1,908 transcription factors (TFs) and transcription regulators (TRs) across 92 gene families and 1,243 protein kinases (PKs) within 124 families. This robust annotation strategy revealed a total of 28,231 genes, offering insights into the genome's functional and regulatory landscape (Table 3).

Genome collinearity analysis. Using jcvi[34] for synteny analysis between Purpleleaf Plum and its two reference genomes, *P. armeniaca* and *P. salicina*, the karyotype layout reveals the structural conservation between the genomes. A total of 115 syntenic blocks and 18,740 gene pairs were identified between Purpleleaf Plum and

P. armeniaca, while 72 syntenic blocks and 16,731 gene pairs were identified between Purpleleaf Plum and *P. salicina* 'Sanyueli'. These results highlight the homologous gene regions across the genomes, demonstrating significant structural conservation (Fig. 4).

Data Records

Both the genomic HiFi sequencing data (~16.22 Gb) generated using the PacBio Sequel II system and the transcriptome sequencing data have been deposited in the NCBI Sequence Read Archive under the accession number SRP586190[35]. The assembled genome has been deposited in GenBank under the accession number GCA_050575025.1[36]. The chromosomal assembly and dataset of gene annotation have been deposited at Figshare[37].

Technical Validation

We evaluated the integrity of the genome using the BUSCO[38] pipeline based on the embryophyta_odb10 database. For the genome, the BUSCO assessment revealed that 98.9% of the 1597 highly conserved orthologs were present as complete genes. Among these, 96.8% were single-copy BUSCOs, and 2.1% were duplicated BUSCOs.

The BUSCO assessment for protein coding genes revealed that 98.6% of the 1614 highly conserved orthologs were present as complete genes. Among these, 84.2% were single-copy BUSCOs, and 14.4% were duplicated BUSCOs. Additionally, 0.7% were fragmented BUSCOs, and 0.7% were missing, demonstrating the high completeness of the annotation.

Code availability

No custom code was used during this study.The public softwares used in this work, were cited in the Methods section. If no detailed parameters were mentioned for a software, default parameters were applied with the guidance.

Received: 26 February 2025; Accepted: 7 July 2025;
Published online: 11 July 2025

References

1. Tutin, T. G. (ed.) Flora Europaea, Plantaginaceae to Compositae (and Rubiaceae) **Vol. 4** (Cambridge Univ. Press, 1964).
2. Chatziathanasiou, S., Kitikidou, K. & Milios, E. Crown width-tree height models for *Magnolia grandiflora, Prunus cerasifera*, and *Acer negundo* growing in cities in northeastern Greece. *Land* **13**, 1579 (2024).
3. Duchovskis, P., Stanys, V., Sasnauskas, A. & Bobinas, C. Cold resistance of *Prunus domestica* L. and *Prunus cerasifera* Ehrh. Lithuania. *Acta Hort.* **734**, 299–303 (2007).
4. Shi, B. S. Study on physiological changes and influencing factors of leaf color of *Prunus cerasifera*. Ph.D. thesis, Northeast Forestry University (2006).
5. Wang, X. A. *et al*. Secondary metabolites in leaves of purple-leaf plum and red-leaf peach. *Fujian J. Agric. Sci.* **33**, 1039–1042 (2018).
6. Chen, X., Xie, J., Yue, Y. Z., Yang, X. L. & Wang, L. G. Advances in research on leaf coloratioin mechanism of colored leaf plants. *Acta Bot. Boreali-Occident. Sin.* **40**, 358–364 (2020).
7. Wang, S. *et al*. Apple MdMYB306-like inhibits anthocyanin synthesis by directly interacting with *MdMYB17* and *MdbHLH33*. *Plant J.* **110**(4), 1021–1034 (2022).
8. Wang, W. *et al*. The basic helix-loop-helix transcription factor *PpeUNE12* regulates peach ripening by promoting polyamine catabolism and anthocyanin synthesis. *Plant Physiol. Biochem*. 109537 (2025).
9. Zhu, C. *et al*. WD40 protein *OsTTG1* promotes anthocyanin accumulation and CBF transcription factor-dependent pathways for rice cold tolerance. *Plant Physiol.* **197**, kiae604 (2025).
10. Li, W., Yang, S., He, Z. & Jin, B. Research advances in the regulatory mechanisms of leaf coloration. *Acta Hort. Sin.* **44**, 1811–1824 (2017).
11. Wang, W. B. *et al*. Chromosome-scale genome assembly and insights into the metabolome and gene regulation of leaf color transition in an important oak species, Quercus dentata. *New Phytologist*. **238**, 2016–2032 (2023).
12. Feng, L. *et al*. The highly continuous reference genome of a leaf-chimeric red pineapple (*Ananas comosus* var. *bracteatus* f. *tricolor*) provides insights into elaboration of leaf color. *G3.* **12**, jkab452 (2022).
13. Li, X. *et al*. Chromosome-level genome assembly for *Acer pseudosieboldianum* and highlights to mechanisms for leaf color and shape change. *Front. Plant Sci.* **13**, 850054 (2022).
14. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k -mers. *Bioinformatics* **7**, 764–770 (2011).
15. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
16. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat Methods.* **18**, 170–175 (2021).
17. Marçais, G. *et al*. MUMmer4, A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
18. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
19. Huang, Z., Shen, F., Chen, Y., Cao, K. & Wang, L. Chromosome-scale genome assembly and population genomics provide insights into the adaptation, domestication, and flavonoid metabolism of Chinese plum. *Plant J.* **108**, 1174–1192 (2021).
20. Groppi, A. *et al*. Population genomics of apricots unravels domestication history and adaptive events. *Nat. Commun.* **12**, 3956 (2021).
21. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
22. Ou, S. *et al*. Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline. *Genome Biol* **20**, 275 (2019).
23. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
24. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
25. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
26. McGuffin, L. J. & Jones, D. T. Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* **19**, 874–881 (2003).

27. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform.* **12**, 1–14 (2011).
28. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
29. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods.* **18**, 366–368 (2021).
30. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
31. Ashburner, M. *et al.* Gene ontology, tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
32. Kanehisa, M. & Goto, S. KEGG, kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
33. Zheng, Y. *et al.* iTAK, a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant.* **9**, 1667–1670 (2016).
34. Tang, H. *et al.* JCVI, A versatile toolkit for comparative genomics analysis. *iMeta* **3**, e211 (2024).
35. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP586190 (2025).
36. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_050575025.1 (2025).
37. Li, H. The genome and annotation of *Prunus cerasifera*. *Figshare* https://doi.org/10.6084/m9.figshare.27897762.v2 (2025).
38. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

Acknowledgements

This work was supported by grants from the Scientific Research Improvement Project of Beijing University of Agriculture (GL2012005) to D.Q., the R&D Program of the Beijing Municipal Education Commission (KM202310020010) to S.W., and the Beijing University of Agriculture (Start-up fund) to Y.Z.

Author contributions

Y.Z., S.W. and Q.H.D. conceived the project and designed the experiments; L.D. collected and prepared the samples; S.W. H.L. and J.L. performed the genome assembly, gene annotation, and other bioinformatics analyses; H.L. drafted the manuscript; H.L., S.W., Y.Z., Q.D. and M.Z. revised the manuscript. All authors contributed to the article and approved the submitted version.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.Z., Q.D. or S.W.