scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-scale reference genome assembly and annotation of *Prunus scopulorum*

Tianfa Guo[1,2,3], Xingyan Li[1,3], Dongzi Zhu[1,3], Bingxue Shen[1], Po Hong[1], Lisi Zhang[1], Shan Sun[1], Jiawei Wang[1]✉, Qian Qiao[1]✉ & Qingzhong Liu[1]✉

Prunus scopulorum **is a valuable Chinese wild cherry germplasm resource. It can survive under relatively cold climate conditions and is highly adaptable to the environment. Here, we successfully constructed a high-quality whole-genome map of *P. scopulorum* at the chromosome level. The genome size is 248.6 Mb, with a scaffold N50 of 27.83 Mb. A total of ~111.76 Mb of repetitive sequences were identified, accounting for 44.96% of the total genome, of which 1.34% were long interspersed nuclear elements (LINEs), 18.29% were long terminal repeats (LTRs), and 0.22% were short interspersed nuclear elements (SINEs). The benchmarking universal single-copy orthologues (BUSCOs) value reached 97.0%, and the LTR assembly index (LAI) reached the reference genome level (17.04). Further analysis led us to predict the presence of 32,542 protein-coding genes, 99.41% of which are functionally annotated. This high-quality genome assembly provides valuable resources for understanding species evolution in the genus *Prunus* and promoting the study of important traits in *P. scopulorum*.**

Background & Summary

Prunus scopulorum is a species of cherry with a special geographical distribution, that is confined mainly to the Qinling Mountains in China and its surrounding areas and is distributed mainly in the valley forests of Shaanxi, Gansu, Hubei, Sichuan and Guizhou Provinces[1]. Therefore, it is also called Qinling Cherry. On the basis of its scientific classification, it belongs to the *Cerasus* subgenus of the genus *Prunus*. The name of this species dates back to 1912, when it was first proposed by Koehne in 'Plantae wilsonianae'[2]. It grows naturally on the sides of ravines and sunny mountain slopes at altitudes of 700–1200 m. It adapts to high altitudes and cold climates, can survive relatively cold climate conditions, and has strong adaptability to the environment. In terms of its morphological characteristics (Fig. 1), *P. scopulorum* is similar to common Chinese cherry species, with typical features of Chinese cherries, such as round or oval fruits and five-petaled petals[3]. However, owing to its unique living environment and geographical distribution, *P. scopulorum* differs from other Chinese cherry species in terms of fruit size, colour, flavour, etc., especially because the fruit has a unique bitter taste, which is very different from that of the common Chinese cherry and the European sweet cherry[3–6]. This difference may be the result of long-term natural selection to adapt to a particular ecological environment. *P. scopulorum* is a cherry species with a special distribution in China that not only contributes to increasing the understanding of cherry diversity but also provides references for cherry cultivation and ecological research.

With the latest technological progress in Illumina HiSeq and Hi-C sequencing technologies, we have achieved a high-quality chromosome-level assembly of *P. scopulorum*. The whole-genome assembly size of *P. scopulorum* is 248.6 Mb, which is anchored to 8 chromosomes, with a scaffold N50 of 27.83 Mb. The successful assembly of the genome will greatly facilitate research on cherry diversity and adaptability and molecular-assisted breeding efforts.

[1]Shandong Institute of Pomology, Taian, Shandong, 271000, China. [2]College of horticulture and forestry, Tarim University, Aral, Xinjiang, 843300, China. [3]These authors contributed equally: Tianfa Guo, Xingyan Li, Dongzi Zhu. ✉e-mail: wangjw-sdip@qq.com; qiaoq404@163.com; qzliu001@126.com

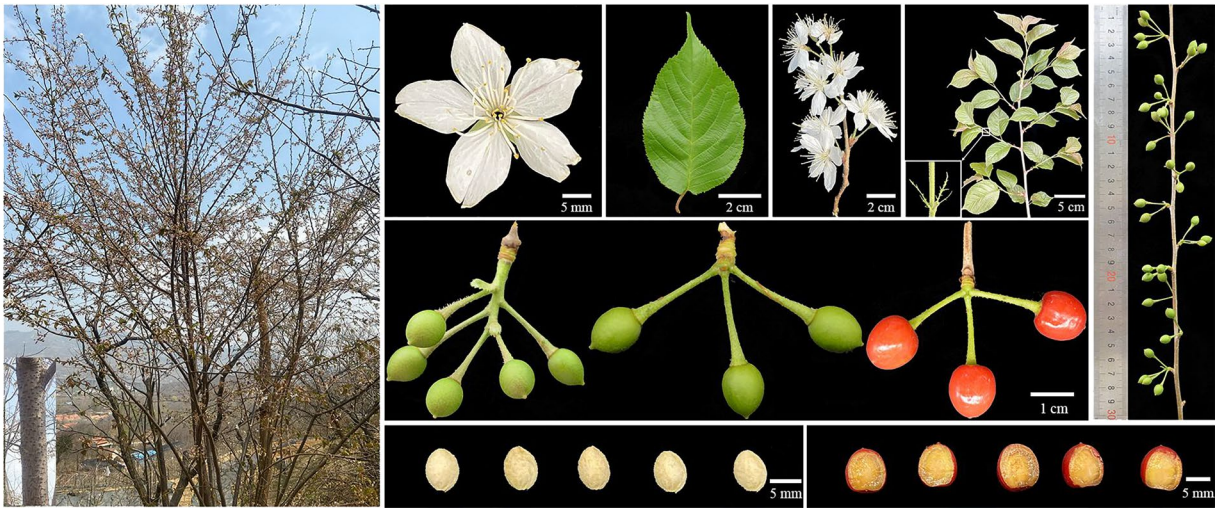

Fig. 1 Phenotypic characteristics of *P. scopulorum*. Note: Flowers, leaves, stems, fruits and seeds were collected between March and May 2024.

Property	min	max
Homozygous (aa)	98.15%	98.19%
Heterozygous (ab)	1.81%	1.85%
Genome Haploid	230,528,139 bp	231,478,697 bp
Genome Repeat Length	78,168,183 bp	78,490,501 bp
Genome Unique Length	152,359,956 bp	152,988,195 bp
Model Fit	67.35%	80.90%
Read Error Rate	0.238%	0.238%

Table 1. Genome survey of *P. scopulorum* (k-mer = 31).

Methods

Sample collection and sequencing. The *P. scopulorum* plant (Psco-1) preserved in our cherry germplasm resource nursery (Tai'an city, Shandong Province, China (117°14′, 36°18′)) was used as experimental material (Fig. 1). The plant was bred from the seeds of the *P. scopulorum* population collected from the Qinling Mountains. Fresh young leaves were subjected to DNA extraction for genome analysis. The samples were quickly frozen in liquid nitrogen and stored at −80 °C for subsequent genome sequencing and Hi-C analysis.

We used a modified CTAB Plant DNA Kit (Aidab Biotechnologies Co., Ltd., Beijing, China) to extract genomic DNA following the manufacturer's guidelines. The concentration and purity of the extracted DNA were evaluated using an IMPLEN P330-31 (Thermo Fisher Scientific, Waltham, MA, USA).

The genomic short-read library, genomic long-read library, short-read transcriptome library, and full-length transcriptome library were constructed following the standard protocol provided by the sequencer manufacturer and were described in our previous study[7,8]. The Hi-C library was constructed following the protocols proposed by Wang *et al.*[7]. The qualified genomic short-read library, Hi-C library and short-read transcriptome library were sequenced using a HiSeq 4000 sequencer (Illumina Inc., San Diego, CA, USA) with the PE150 model.

Genome assembly and evaluation. We sequenced and assembled *P. scopulorum* using Illumina HiSeq, Oxford Nanopore technology (ONT) and Hi-C technology. A total of ~46.69 Gb of long sequencing reads were generated from ONT sequencing.. We first predicted the genome using K-mer analysis. The predicted genome size was approximately 231.48 Mb, with a repetition ratio of 33.90% and heterozygosity of 1.83% (Table 1; Fig. 2A). For Illumina sequencing, ~41.28 Gb of reads were obtained, with ~40.91 Gb of reads remaining after low-quality reads were filtered out, of which the Q30 was 38.11 Gb (93.18%) and the GC content was 39.53% (Table 2). During the process of genome assembly, preliminary NECAT (v0.01) (https://github.com/xiaochuanle/NECAT)[9] assembly was carried out using ONT data, then the primary assembly was polished by Medaka (v1.0.3) (https://github.com/nanoporetech/medaka)[10] and Nextpolish (v1.0.5) with Illumina data, finally Purge Haplotigs (v1.1.0)[11] was applied to generate 238.55 Mb of contigs. We then manually adjusted some errors in the automatic mount and assessed the reliability of the Hi-C-based chromosomal assembly using Juicebox (v2.15.00) (https://github.com/aidenlab/Juicebox)[12] and D-Genies (v1.5.0) (https://github.com/genotoul-bioinfo/dgenies/releases)[13]. A Hi-C interaction heatmap indicated that the *P. scopulorum* genome had no obvious assembly errors and contained eight clusters at the chromosomal level (Fig. 2B). The fragments were subsequently mounted to the chromosome level in combination with Hi-C (v0.9.1214) (https://github.com/tangerzhang/ALLHiC/wiki)[14] data to obtain 235.57 Mb scaffolds. Finally, gap filling was performed using the ONT data, and eight pseudochromosomes of

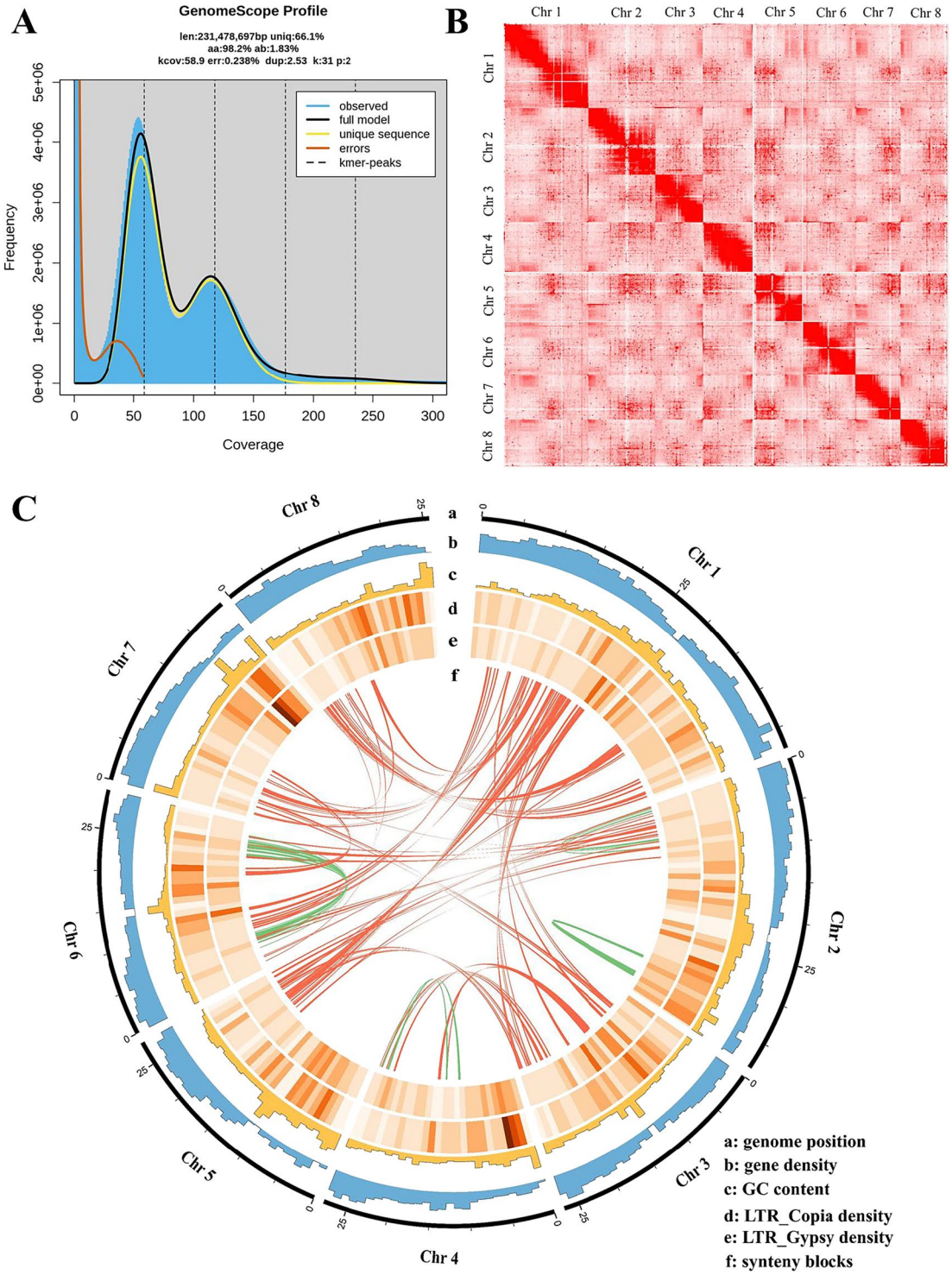

Fig. 2 Genome assembly of *P. scopulorum*. (**A**) K-mer distribution and statistics for the genome of *P. scopulorum;* (**B**) High-resolution Hi-C interaction heatmap of the *P. scopulorum* genome. (**C**) Summary of the *de novo P. scopulorum* genome assembly and sequencing analysis.

248.6 Mb were obtained. The GC content was 37.87%, and the scaffold N50 was 27.83 Mb (Table 1). We compared the genome assembly metrics among *P. scopulorum*, *P. serrulata*, and *P. avium* (Table 3). The results indicate that *P. scopulorum* had the lowest number of contigs and scaffolds and produced the longest number of contigs. Moreover, it had the greatest average coding sequence (CDS) length (3,365.38 bp), which exceeded those of the other two species. The quality of the assembly was evaluated using the Benchmarking Universal Single-Copy Orthologues (BUSCOs, (v5.0.2)) pipeline with the Embryophta_obd10 database[15]. Analysis of the final assembly indicated 97.0% completeness, with only 2.3% missing single-copy orthologues (Table 1). Long terminal repeat

Type	Before filtering	After filtering
Total reads	275.177308 M	273.246556 M
Total bases:	41.276596 G	40.906521 G
Q20 bases:	40.270187 G (97.561793%)	39.971050 G (97.713148%)
Q30 bases:	38.375068 G (92.970524%)	38.114698 G (93.175113%)
GC content:	39.55%	39.53%

Table 2. Statistics of the Illumina sequencing data for *Prunus scopulorum*.

Genome information	*P. scopulorum*	*P. serrulata*	*P. avium*
Genome size (Mb)	248.6	265.4	344.29
GC content (%)	37.87	38.15	38.43
Number of contigs	87	304	610
Contig N50 (Mb)	5.43	1.56	3.25
Longest contig (Mb)	22.73	7.34	13.6
Number of scaffolds	8	67	—
Scaffold N50 (Mb)	27.83	31.12	—
Longest scaffold (Mb)	47.56	49.87	—
Predicted chromosome genes	32,717	—	—
Gene density (per 100 kb)	9.72	8.08	11.17
Average CDS length (bp)	3,365.38	1,572.98	2,758.67
Average exon length (bp)	290.10	269.393	277.63
Complete BUSCO number	1565 (96.9%)	1528 (94.67%)	1572 (97.4%)
Complete single-copy BUSCOs	1517 (94.0%)	1352 (83.77%)	1472 (91.2%)
Complete duplicated BUSCOs	49 (3.0%)	176 (10.9%)	100 (6.2%)
Fragmented BUSCOs (%)	12 (0.7%)	30 (1.86%)	8 (0.5%)
Missing BUSCOs (%)	36 (2.3%)	56 (3.47%)	33 (2.1%)

Table 3. Statistics for the final *P. scopulorum, P. serrulata*, and *P. avium* genome assembly.

Species	LAI	Contig N50 (Mb)
P. scopulorum	17.04	5.18
P. avium[7]	19.68	3.25
P. persica[15]	18.79	0.25
P. pusilliflora[16]	17.35	6.00
P. armeniaca[17]	16.29	1.02
P. yedoensis[18]	6.87	0.92
P. domestica[19]	2.27	1.74

Table 4. Long-terminal-repeat retrotransposon assembly index (LAI) analysis and contig N50s of different genome assemblies in *Prunus* species.

retrotransposons (LTR-RTs) were identified using LTR_FINDER scripts, and the LTR assembly index (LAI) was calculated for each genome using LTR Retriever (v2.9.0)[16]. The LAI of *P. scopulorum* was calculated to be 17.04, which reached the reference genome level ($10 \leq LAI < 20$). We found that the LAI was only slightly lower than that of the well-assembled *P. avium* (19.68)[7], *P. persica* (18.79)[17] and *P. pusilliflora* (17.35)[18] genomes but higher than that of the *P. armeniaca* (16.29)[19], *P. yedoensis* (6.87)[20], and *P. domestica* (2.27)[21] genomes, indicating that it has superior assembly quality (Table 4).

Genome annotation. Repetitive sequences in the *P. scopulorum* genome were initially identified using RepeatModeler (v2.0.1)[22] on the basis of the structural features of the sequences. The identified repetitive elements were subsequently classified using RepeatMasker (v4.0.9)[23] with DFAM (v3.151)[24] and the Repbase library (v20170127) From the assembled draft genome of *P. scopulorum*, a total of ~111.76 Mb of repetitive sequences were identified, accounting for 44.96% of the total genome, of which 1.34% were long interspersed nuclear elements (LINEs), 18.29% were LTRs, and 0.22% were short interspersed nuclear elements (SINEs) (Table S1).

The Funannotate pipeline (v1.8.3)[25] was used to predict and annotate the protein-coding genes (Scripts used were deposited to github repository). RNA-seq data of three stage of fruit was used to train the prediction database. The functional annotation of the protein-coding genes was carried out using the COG[26], EggNOG (v4.5.1)[27], Gene Ontology[8,28], Pfam (v32.0)[29] and InterProScan (v101.0)[30] databases. In this study,

Annotation database	Annotated number	Annotated ratio (%)
GO	18,780	57.71
InterPro	25,076	77.06
COG	27,104	83.29
Pfam	21,290	65.42
eggNOG	29,117	89.47
Total	32,542	—

Table 5. Information on the annotations of the predicted protein-coding genes.

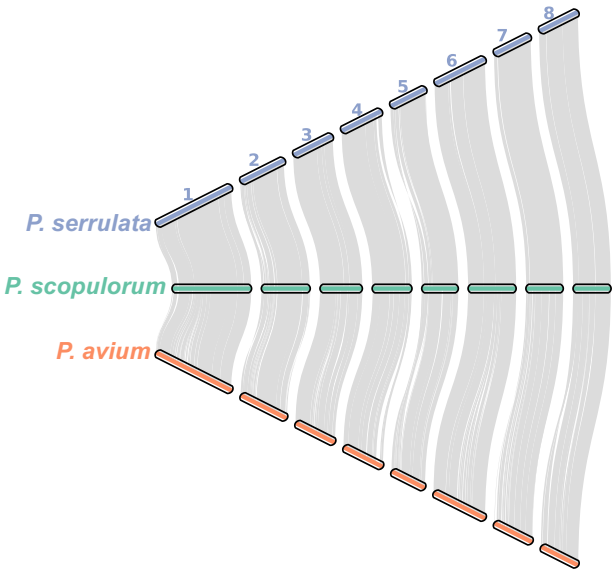

Fig. 3 Synteny analysis of the genomes of *P. scopulorum*, *P. serrulata*, and *P. avium* genomes.

a total of 32,717 genes were assigned to chromosomes on the basis of a combination of transcriptome-based, homology-based, and ab initio predictions. The average transcript length and coding sequence size were 3,365.38 bp and 290.10 bp, respectively, and the average exon length was 454.10 bp (Table 3). We functionally annotated 32,542 genes using the EggNOG (29,117), Pfam (21,290), Clusters of Orthologous Groups of proteins (COG) (27,104), Gene Ontology (GO) (18,780), and InterPro (25,076) databases (Table 5). We also identified more than one and a half thousand RNA genes, mainly rRNAs (274) and tRNAs (723), and several snRNAs (75) from the *P. scopulorum* genome (Table S1). In light of the observed gene density, repeat density, and GC content, a circos map of the genome of *P. scopulorum* was constructed (Fig. 2C).

Intergenomic synteny. *P. serrulata* and *P. avium* were selected for intergenomic synteny analysis with *P. scopulorum*. As shown in Fig. 3, significant genomic synteny was detected among the *P. scopulorum*, *P. serrulata*, and *P. serrulata* genomes, with excellent collinearity confirming the high quality of the *P. scopulorum* intergenomic assembly.

Data Records

Raw sequencing data and genome assembly data of this study have been deposited into the European Nucleotide Archive (ENA) https://identifiers.org/ena.embl:ERP180767 (2025) and https://identifiers.org/insdc.gca:GCA_977002725 (2025). The data that support the findings of this study also deposited into the CNGB Sequence Archive (CNSA) of the China National GenBank DataBase (CNGBdb). The details of data assembly and annotations are described in Tables 3 and 5. The publication of this data set further enriches the cherry genome database, and complete annotation information is conducive to molecular breeding and germplasm diversity[31,32].

The genome assembly and annotate script used were deposited to https://github.com/wangjw065/Genome-assembly-and-annotation-of-Prunus-scopulorum. Other commands and pipelines related to data processing were executed strictly following the manuals and protocols provided by the respective bioinformatics software tools.

Technical Validation

To comprehensively assess the quality of this genome, first, we used BUSCO, a tool that enables high-quality assessment of genome assembly integrity, and second, we used LTR_retriever, a program that supports multi-threaded execution, the ability to perform *de novo* identification enables the assessment of high-quality genomic annotation information. For details, see the Methods description.

At the same time, we also conducted the collinearity of *P. scopulorum*, *P. serrulata*, and *P. avium* (Fig. 3). This also verified the quality of the *P. scopulorum* genome that we assembled.

Data availability

Raw sequencing data and genome assembly data of this study have been deposited into the European Nucleotide Archive (ENA) https://identifiers.org/ena.embl:ERP180767 (2025) and https://identifiers.org/insdc.gca:GCA_977002725 (2025). The data that support the findings of this study also deposited into the CNGB Sequence Archive (CNSA) of the China National GenBank DataBase (CNGBdb).

Code availability

The genome assembly and annotate script used were deposited to https://github.com/wangjw065/Genome-assembly-and-annotation-of-Prunus-scopulorum. Other commands and pipelines related to data processing were executed strictly following the manuals and protocols provided by the respective bioinformatics software tools.

Received: 25 February 2025; Accepted: 19 January 2026;
Published online: 09 March 2026

References

1. L LC, B B. Flora of China: 9: Beijing:Science Press,Missouri:Missouri Botanical Garden Press; (2003).
2. Koehne, E. Plantae wilsonianae. Part 1, vol. 59. Cambridge: The University Press; 1912.
3. Wang, Y. *et al*. Phenotyping in flower and main fruit traits of Chinese cherry [*Cerasus pseudocerasus* (Lindl.) G.Don]. *Scientia Horticulturae* **296**, 110920 (2022).
4. Wang, Y. *et al*. Inheritance analysis of fruit-related traits in Chinese cherry [*Cerasus pseudocerasus* (Lindl.) G. Don] breeding progenies. *Scientia Horticulturae* **307**, 111519 (2023).
5. Wang, Y. *et al*. Survey on Intra-specific Crossing and F1 Seedling Cultivation in Seven Combinations of Chinese Cherry. *The Horticulture Journal* **91**(3), 267–275 (2022).
6. Chen, T. *et al*. Investigation, collection and preliminary evaluation of genetic resources of Chinese cherry [*Cerasus pseudocerasus* (Lindl.) G. Don]. *Journal of Fruit Science* **08**(33), 917–933 (2016).
7. Wang, J. *et al*. Chromosome-scale genome assembly of sweet cherry (*Prunus avium* L.) cv. Tieton obtained using long-read and Hi-C sequencing. *Horticulture Research* **7**(1), 122 (2020).
8. Qiao, Q. *et al*. Chromosome-level genome assembly of the red mutant of Chinese chestnut (*Castanea mollissima*) provides new insights into anthocyanin accumulation. *Scientia Horticulturae* **338**, 113683 (2024).
9. Chen, Y. *et al*. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nature Communcations* **12**(1), 60 (2021).
10. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**(7), 2253–2255 (2020).
11. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**(1), 460 (2018).
12. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems* **3**(1), 99–101 (2016).
13. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958 (2018).
14. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature Plants* **5**(8), 833–845 (2019).
15. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**, 227–245 (2019).
16. Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiology* **176**(2), 1410–1422 (2018).
17. Verde, I. *et al*. The Peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC Genomics* **18**(1), 225 (2017).
18. Jiu, S. *et al*. Chromosome-scale genome assembly of *Prunus pusilliflora* provides novel insights into genome evolution, disease resistance, and dormancy release in *Cerasus* L. *Horticulture Research* **10**(5), uhad062 (2023).
19. Jiang, F. *et al*. The apricot (*Prunus armeniaca* L.) genome elucidates Rosaceae evolution and beta-carotenoid synthesis. *Horticulture Research* **6**(1), 128 (2019).
20. Baek, S. *et al*. Draft genome sequence of wild *Prunus yedoensis* reveals massive inter-specific hybridization between sympatric flowering cherries. *Genome Biology* **19**(1), 127 (2018).
21. Zhebentyayeva, T. *et al*. Genetic characterization of worldwide *Prunus domestica* (plum) germplasm using sequence-based genotyping. *Horticulture Research* **6**(1), 12 (2019).
22. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**(17), 9451–9457 (2020).
23. Zhi, D., Raphael, B. J., Price, A. L., Tang, H. & Pevzner, P. A. Identifying repeat domains in large genomes. *Genome Biol* **7**(1), R7 (2006).
24. Hubley, R. *et al*. The Dfam database of repetitive DNA families. *Nucleic acids research* **44**(D1), D81–89 (2016).
25. Love, J. nextgenusfs/funannotate: funannotate v1. 7.2. Zenodo. In.; (2019).
26. Koonin, E. V. *et al*. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol* **5**(2), R7 (2004).
27. Huerta-Cepas, J. *et al*. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol* **34**(8), 2115–2122 (2017).
28. Harris, M. A. *et al*. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**(Database issue), D258–261 (2004).
29. El-Gebali, S. *et al*. The Pfam protein families database in 2019. *Nucleic Acids Res* **47**(D1), D427–d432 (2019).
30. Paysan-Lafosse, T. *et al*. InterPro in 2022. *Nucleic Acids Res* **51**(D1), D418–d427 (2023).
31. Liu, Q. *et al*. *European Nucleotide Archive*. https://identifiers.org/ena.embl:ERP180767 (2025).
32. Liu, Q. *et al*. *European Nucleotide Archive* https://identifiers.org/insdc.gca:GCA_977002725 (2025).

Acknowledgements

This study was supported by the Shandong Provincial Natural Science Foundation (ZR2023QC054 and ZR2021MC117), the Key R&D Program of Shandong Province, China (2022TZXD006 and 2020LGGC008), and the Yantai Seed Project 'Creation of the Sweet Cherry Breeding Team'.

Author contributions

J.W., Q.L. and Q.Q. conceived and designed the experiments; Q.Q., T.G. and X.Y. collected the samples and drafted the manuscript; D.Z., X.Y. and J.W. conducted the assembly and annotations and contributed to the sequencing data analyses; and B.S., P.H., L.S. and S.S. collected the samples and performed the phenotyping. All the authors provided final approval for publication.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-06657-x.

Correspondence and requests for materials should be addressed to J.W., Q.Q. or Q.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.