scientific data

OPEN

DATA DESCRIPTOR

Telomere-to-Telomere gap-free genome assembly of the allotetraploid wild raspberry *Rubus alceifolius* Poir

Mengzhuo Zheng[1], Guilian Sun[1], Xianyan Han[1], Yan Sun[1], Shoucheng Liu[1], Xiaopeng Li[1], Jiyuan Guan[2], Qunying Zhang[2] & Junhui Zhou[1]✉

The genus *Rubus* serves as an ideal model for exploring plant genome evolution due to its complex polyploidization history, frequent interspecific hybridization, and exceptional adaptive capacity. Given its potential medicinal value and recognized invasive ecological threats, the wild allotetraploid *Rubus alceifolius* Poir. represents a scientifically important taxon. Here, we assembled a telomere-to-telomere gap-free genome assembly of *R. alceifolius* Poir. using PacBio HiFi, ONT, Hi-C and Illumina short reads sequencing. The assembly spans 462.05 Mb, anchored into 14 pseudochromosomes with a scaffold N50 of 33.09 Mb, and includes all 28 telomeric regions. Genome annotation revealed that 44.95% of the assembly comprises repetitive sequences, and 46,698 protein-coding genes were predicted, with 93.90% functionally annotated. Notably, a large-scale inversion structural variant spanning 12.44 Mb (6.65–19.09 Mb) was identified on chromosome Chr2 of the subgenome A. This high-quality genome assembly fills a critical gap in genomic resources for polyploid *Rubus* species and provides an essential foundation for elucidating adaptive evolution mechanisms within the genus, as well as advancing molecular breeding strategies for raspberry improvement.

Background & Summary

The genus *Rubus*, belonging to Rosaceae family, encompasses approximately 740 species distributed across all continents except Antarctica[1–3]. Its commercially significant edible aggregate fruits (raspberries) hold substantial agroeconomic importance in over 50 countries, with primary production concentrated in North America, Europe, and Russia[4]. Raspberries, identified as third-generation functional fruits, contain abundant bioactive compounds associated with substantial health-promoting properties, including nutritional and therapeutic effects[5,6]. Despite extensive germplasm diversity, cultivated raspberries represent only a limited subset. Most wild taxa, especially those widely distributed in Asia, remain underutilized in raspberry breeding. These wild relatives harbor valuable agronomic traits with considerable potential for advancing precision molecular breeding[7].

The genus *Rubus* exhibits considerable variation in ploidy level. With a basic chromosome number of 7, ploidy levels range from diploid (2n = 2x = 14) to tetradecaploid (2n = 14x = 98), reflecting extensive intra- and interspecific polyploidy[8]. A complex evolutionary dynamic, characterized by frequent interspecific hybridization, polyploidization, and introgression, significantly complicates taxonomic delineation and phylogenetic reconstruction within the genus. Crucially, polyploidization is recognized as a key adaptive mechanism underpinning habitat diversification in Rubus. This process enhances adaptive radiation, facilitating global colonization and generating reticulate patterns of evolution[9].

Rubus alceifolius Poir. is a tetraploid (2n = 4x = 28) species native to Southeast Asia[10]. It has become naturalized as an invasive weed in Indian Ocean islands (e.g., Mayotte, Réunion) and Australia, posing significant threats to native ecosystems[10–13]. Notably, its roots, leaves, and fruits accumulate bioactive antioxidants—including anthocyanins, β-carotene, and flavonoids—implicating potential medicinal value[14–16]. Consequently,

[1]Peking University Institute of Advanced Agricultural Sciences, Shandong Laboratory of Advanced Agricultural Sciences in Weifang, Shandong Provincial Key Laboratory of Precision Molecular Crop Design and Breeding, Shandong, 261325, China. [2]Botanic Garden of Guizhou Province, Guiyang city, Guizhou Province, 550081, China. ✉e-mail: junhui.zhou@pku-iaas.edu.cn

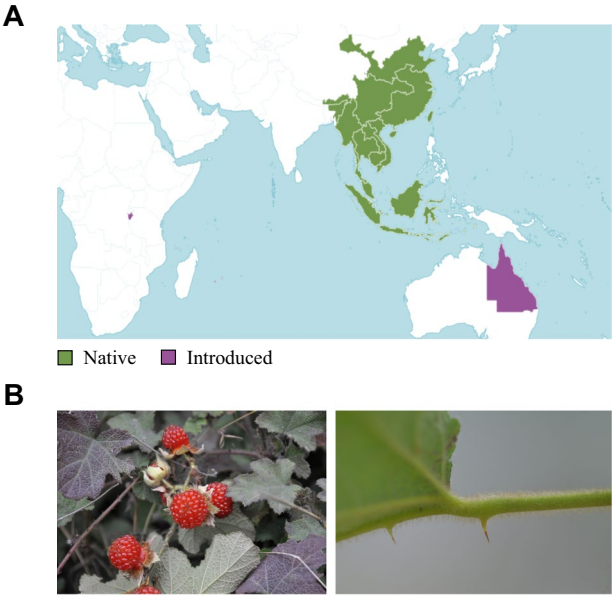

Fig. 1 Global distribution and morphological characteristics of *R. alceifolius* Poir. (**A**) Species distribution map showing the native range (green, East and Southeast Asia) and introduced range (purple, Africa and Oceania). (**B**) The morphology of mature fruits, leaves and stems of *R. alceifolius* Poir.

R. alceifolius Poir. serves as both an ideal model for studying adaptive evolution and polyploid genome dynamics in *Rubus*, and a critical genetic resource for breeding stress-resistant cultivars and mining pharmacologically active traits.

The allotetraploid status of *R. alceifolius* Poir. was confirmed through integrated k-mer frequency analysis and flow cytometry. A telomere-to-telomere (T2T) gap-free genome assembly was successfully generated through a hybrid approach combining PacBio high-fidelity (HiFi) sequencing, Oxford Nanopore (ONT) ultra-long reads, Illumina short-read sequencing, and chromatin conformation capture (Hi-C) technologies. The assembled genome comprises 462.05 Mb with a scaffold N50 length of 33.09 Mb and is anchored into 14 pseudochromosomes. Genome annotation revealed that 44.95% of the sequence comprises repetitive elements, while structural annotation predicted 46,698 protein-coding genes, of which 93.90% were functionally annotated. Subgenome comparison and synteny analysis identified a large-scale inversion of 12.45 Mb on subgenome A chromosome Chr2. The T2T genome assembly of *R. alceifolius* Poir. not only fills a critical gap in genomic resources for polyploid *Rubus* species, but also provides a pivotal genetic foundation for elucidating adaptive evolution mechanisms and advancing precision molecular breeding in raspberries.

Methods

Sample collection. *R. alceifolius* Poir., native to East and Southeast Asia, has been introduced as an invasive alien species to Africa (Burundi, Comoros, Mauritius, Réunion) and Oceania (Queensland, Australia)[17] (Fig. 1A). Experimental materials were collected from the raspberry germplasm repository at the Guizhou Botanical Garden, China (106.73°E, 26.63°N). The fruits are subglobose aggregate drupelets that ripen to deep red at maturity, while the leaves and stems exhibit dense yellow tomentum and sparse, rigid prickles (Fig. 1B). For genomic sequencing, young leaf tissues were selected for PacBio HiFi, ONT, Hi-C, and Illumina sequencing. Additionally, root, stem, and leaf tissues were used for transcriptome sequencing and full-length transcriptome sequencing.

Library construction and sequencing. For PacBio HiFi sequencing, genomic DNA extracted via CTAB underwent quality assessment. A 15 kb library was constructed using the SMRTbell Express Template Prep Kit 3.0 (Pacific Biosciences, CA, USA), involving DNA shearing, damage repair, end repair, hairpin adapter ligation, size selection, and library purification. Following quality control, the resultant SMRTbell library was sequenced on a PacBio Revio platform (Pacific Biosciences, CA, USA) using a single 25 M SMRT Cell.

To generate ultra-long Oxford Nanopore reads, DNA was extracted using an optimized CTAB protocol and sequenced on the Oxford Nanopore Technologies platform. Following bead purification and Qubit quantification, samples underwent damage repair and end-prep using the NEBNext system, followed by adapter ligation and library construction. The final library was loaded onto a PromethION flow cell (Oxford Nanopore Technologies, Oxford, UK) for 72 hours of high-throughput sequencing.

The Hi-C library was constructed through the following steps: cells were crosslinked with formaldehyde, and genomic DNA was digested with the 4-cutter restriction enzyme Dpn II. DNA ends were filled in and labeled with biotin, followed by ligation of the resulting blunt-end fragments. After purification, the DNA was randomly sheared into 300–500 bp fragments. Biotin-labeled ligation junctions were then enriched using streptavidin-coated magnetic beads, followed by end-repair, A-tailing, and adapter ligation to generate sequencing-ready libraries. After quality control of the libraries using Qubit 2.0, an Agilent 2100 instrument

Library	Platform	Bases (Gb)	ReadsNumber	Mean read length (bp)	Average coverge (×)
HiFi	PacBio Revio	24.48	1,482,829	16510.66	52.99
ONT	ONT PromethION 48	21.19	431,885	49059.31	45.86
Hi-C	NovaSeq X Plus	39.89	265,941,092	150	86.33
DNA	NovaSeq X Plus	13.51	105,150.620	150	29.23
RNA	NovaSeq X Plus	55.78	372,056,546	150	/
ISO	PacBio Revio	9.99	6,890,434	1449.38	/

Table 1. Summary statistics of sequencing data after quality control for *R. alceifolius* Poir.

(Agilent Technologies, CA, USA), and qPCR, 150 bp paired-end sequencing was performed on the Illumina Novaseq X Plus platform.

For Illumina next-generation sequencing (NGS), a 350 bp paired-end library was constructed following manufacturer protocols (Illumina). DNA fragmentation was conducted by sonication, followed by end-polishing, A-tailing, and full-length adapter ligation. PCR-amplified products were purified using the AMPure XP bead system. Library size distribution was assessed via an Agilent 2100 Bioanalyzer, and quantification was performed by real-time PCR prior to sequencing on the Illumina NovaSeq X Plus platform.

For total RNA sequencing, RNA samples passing quality control were enriched for eukaryotic mRNA using Oligo(dT) magnetic beads. The enriched mRNA was fragmented and utilized as a template for first-strand cDNA synthesis with random hexamer primers. During second-strand synthesis, dATP, dCTP, dGTP, and dTTP were incorporated to construct the double-stranded cDNA library. The resulting dsDNA was purified, followed by end repair and adenylation (A-tailing). Sequencing adapters were then ligated to the repaired ends. Fragments of target size were selected, and the cDNA library was amplified via PCR. Finally, the amplified library was sequenced on the Illumina NovaSeq X Plus platform.

In order to generate full-length transcriptome data, total RNA samples meeting quality criteria (RIN $\geq$ 7.0, input 300 ng) were subjected to mRNA enrichment via Oligo(dT) magnetic beads. The enriched mRNA was fragmented and reverse-transcribed using the Iso-Seq Express 2.0 Kit, which employs template-switching oligonucleotides (TSO) to generate full-length cDNA. PCR amplification was performed with barcoded ISO-Seq primers (bc01-bc12), and the products were purified using SMRTbell magnetic beads. Purified cDNA underwent eight parallel Kinnex PCR reactions with specific primer mixes to generate directionally fragmented DNA sequences. The amplified products were pooled and ligated to Kinnex barcoded adapters (bc01-bc04), forming linear arrays. Subsequent steps included DNA repair, nuclease treatment, and two rounds of SMRTbell bead purification. The final Kinnex library was adjusted to a concentration <60 ng/μL, validated for size distribution via the Femto Pulse system, and sequenced on the PacBio Revio platform.

Data preprocessing and quality control. To ensure data integrity across multiple sequencing platforms, rigorous quality control procedures were implemented prior to downstream analysis. For PacBio HiFi sequencing data, duplicate reads were removed using pbmarkdup (v1.1.0), followed by adapter sequence removal with HiFiAdapterFilt (v3.0)[18]. ONT reads underwent preprocessing via the HERRO[19]. Adapters were trimmed using Porechop (v0.2.4), duplex reads were identified with duplex_tools (v0.3.3) to mitigate systematic errors, and all-to-all read alignments were generated using the create_batched_alignments.sh script. Base error correction was performed using a convolutional neural network (CNN) and Transformer encoder within HERRO inference module, leveraging the pre-trained model model_R10_v0.1.pt to generate high-accuracy FASTA sequences. For next-generation sequencing (NGS) datasets, including Hi-C, whole-genome resequencing, and RNA-seq, adapter trimming and quality filtering were conducted using fastp (v0.24.0)[20], with comprehensive quality assessment via FastQC (v0.12.1) and aggregated reporting through MultiQC (v1.28). Sample identity validation across all sequencing modalities was confirmed using the NTSM (v1.2.1)[21], ensuring biological consistency. Following these stringent quality control steps, high-fidelity datasets were obtained: 24.28 GB of PacBio HiFi reads, 21.19 GB of Oxford Nanopore ultra-long reads, 39.89 GB of Hi-C data, 13.51 GB of resequencing data, 55.78 GB of RNA-seq data, and 9.99 GB of full-length transcriptome data (Table 1).

Chromosomal ploidy determination. Following quality control of HiFi sequencing data, k-mer spectra were analyzed using Fastk (k = 21). The ploidy of *R. alceifolius* Poir. was revealed to be an AABB allotetraploid through Smudgeplot (v0.4.0)[22](Fig. 2A). Model fitting of the k-mer spectrum by GenomeScope (v2.0)[23] estimated the haploid genome size at 224.23 Mb, with a fitting error of less than 0.8% (Fig. 2B). For flow cytometry validation, tissue samples were placed in a plastic culture dish containing 500 μL of Nuclei Extraction Buffer (CyStain UV Precise P Kit), finely chopped with a sharp blade, and incubated for 60 seconds. The resulting homogenate was then filtered through a 50 μm CellTrics filter. Subsequently, 2000 μL of DAPI Staining Buffer was added, and the mixture was incubated in the dark for 2 minutes. Nuclei suspensions were analyzed using a CyFlow Space Flow Cytometer (Sysmex Partec, Muenster, Germany) and the corresponding FloMax software. Diploid *Rubus hirsutus* Thunb. was employed as an internal reference standard. Flow cytometry analysis conclusively identified *R. alceifolius* Poir. as a tetraploid species, with an estimated genome size of 427.07 Mb (Fig. 2C,D).

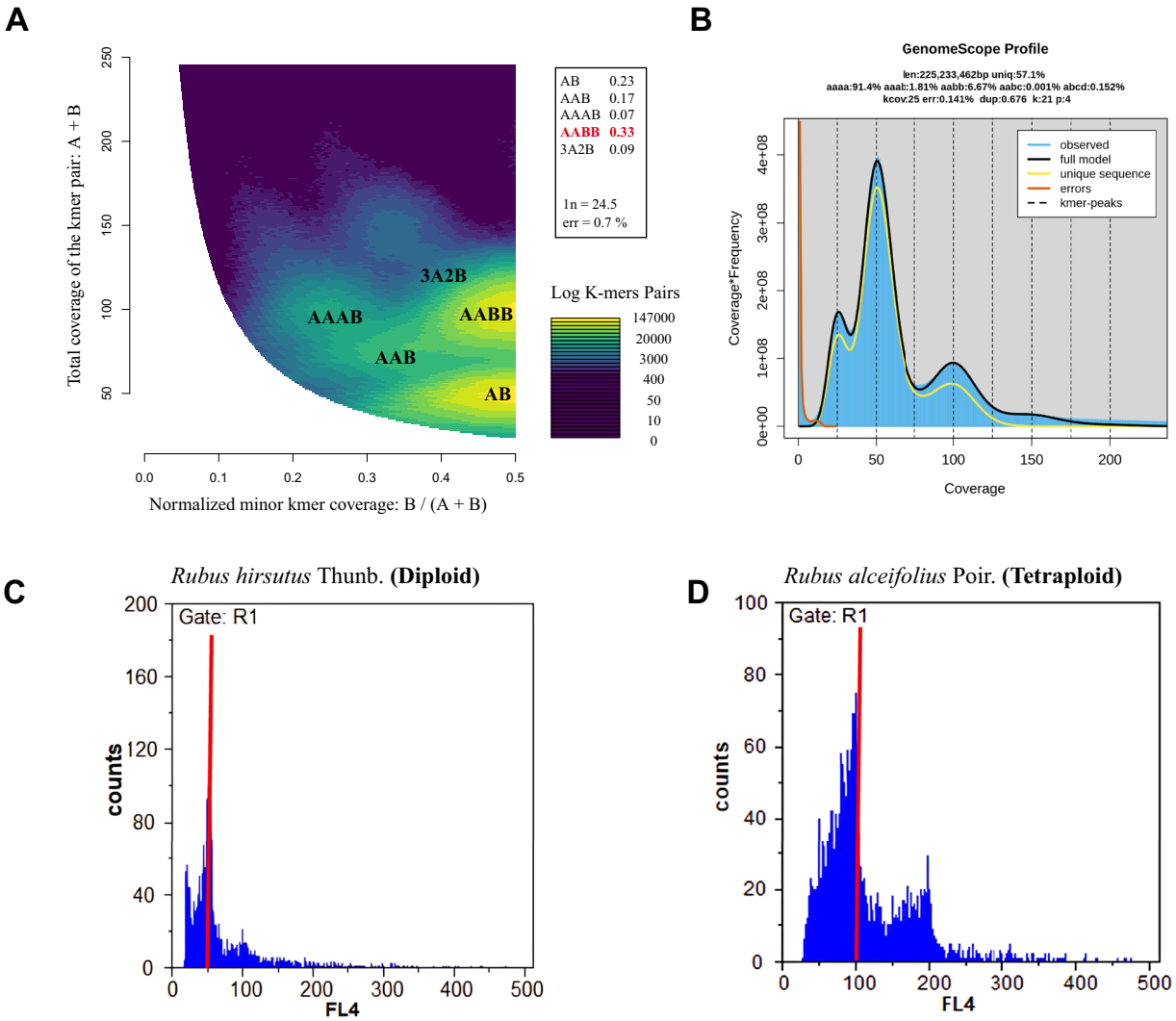

Fig. 2 Genome characteristics of *R. alceifolius* Poir. (**A**) Ploidy estimation via Smudgeplot. (**B**) Genome characteristic assessment using Genomescope 2.0. (**C**) Ploidy level of the control sample (diploid *Rubus hirsutus*) by flow cytometry. (**D**) Ploidy determination of *R. alceifolius* Poir. by flow cytometry.

Genome assembly. Genome assembly of *R. alceifolius* Poir. was performed using PacBio HiFi reads, ONT reads, and Hi-C reads. Initial assemblies were generated independently using hifiasm (v0.25.0-r726)[24], verkko (v2.2.1)[25], and Flye (v2.9.5-b1801)[26], yielding assemblies of 467.90 Mb, 447.87 Mb, and 611.22 Mb, respectively (Table 2). Based on assembly continuity metrics, the hifiasm assembly was selected as the primary draft for constructing a T2T genome. Chromosome-scale scaffolding involved two main steps: First, Hi-C data were aligned to this draft assembly using chromap (v0.2.7-r494)[27]. Second, the aligned Hi-C data were processed using YaHS (v1.2.2)[28] to order, orient, and cluster the scaffolds into chromosome-scale structures based on Hi-C interaction patterns. Normalized contact matrices were generated from the Hi-C alignments using Juicer (v1.19.02)[29] and manually curated within the JuiceBox visualization platform. This process resulted in the construction of 14 pseudochromosomes (Fig. 3B). To enhance genome completeness and achieve T2T standards, sequence gaps within the scaffolds were filled. This was accomplished by integrating information from HiFi reads, ONT reads, and the verkko and Flye assemblies using both the GapFiller module of quarTeT (v1.0)[30] and the patch module of RagTag (v2.1.0)[31]. During the post-assembly polishing stage, we first constructed a k-mer database from HiFi reads using meryl (v1.4.1) and extracted highly repetitive k-mers for subsequent alignment. HiFi reads were then mapped to the initial assembly using winnowmap (v2.03), and the alignment results were sorted and indexed with samtools (v1.21) to produce a high-quality BAM file. In parallel, k-mer databases were generated from NGS data using yak (v0.1). Finally, NextPolish2 (v0.2.1) was employed to integrate the HiFi alignment results and NGS k-mer databases to polish the initial assembly, yielding the final polished genome sequence[32] (Table 3).

Repetive elements annotation. Genome-wide repetitive element annotation was performed using an integrated approach combining *de novo* prediction and homology-based alignment. *De novo* prediction was executed using RepeatModeler (v2.0.6)[33], which employs RECON and RepeatScout to generate a species-specific

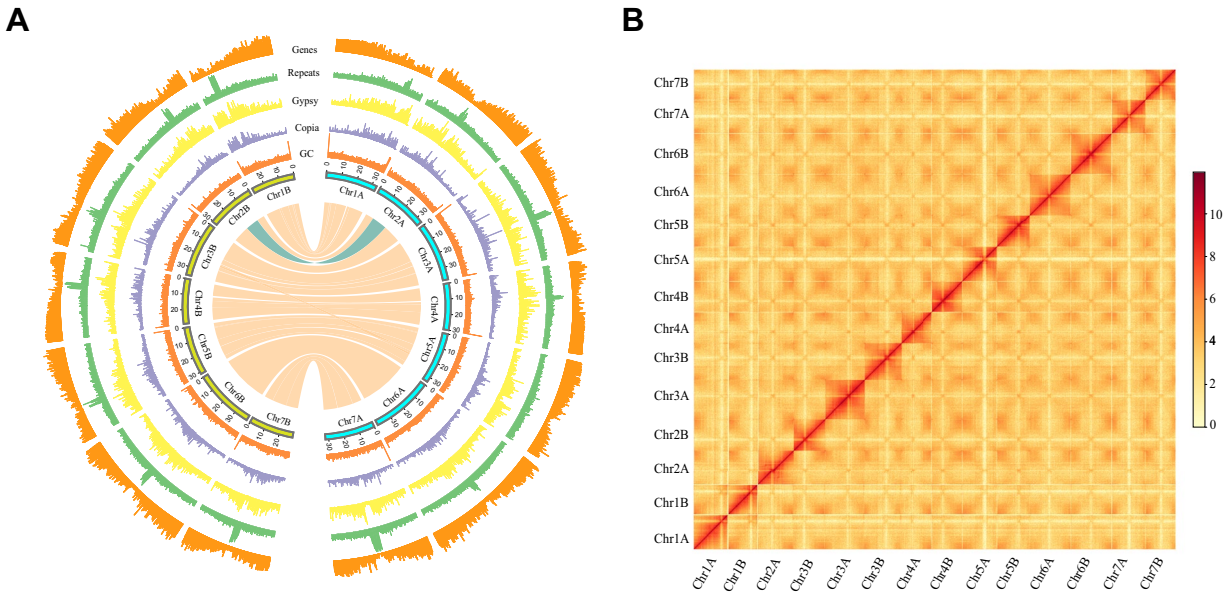

Fig. 3 *R. alceifolius* Poir. T2T genome assembly. (**A**) *R. alceifolius* Poir. T2T genome landscape. Orange lines indicate syntenic regions, while the green line highlights the inversion event. (**B**) Hi-C contact matrix among 14 pseudochromosomes of *R. alceifolius* Poir.

Assembly Stage	Scaffolds Num	Total length (bp)	N50 (bp)	N90 (bp)	Largest sequence (bp)
Hifiasm Assembly	70	467,897,634	31,122,603	12,190,348	38,012,094
Verkko Assembly	195	447,868,893	17,883,921	5,476,588	34,808,986
Flye Assembly	10574	611,219,582	442,626	20,194	12,652,535
HiC-Scaffolding	14	459,223,311	32,444,880	27,580,166	39,271,289
T2T Assembly	14	462,050,512	33,093,563	28,743,846	39,608,517

Table 2. Summary Statistics of *Rubus alceifolius* Poir. genome across different assembly stages.

Chr	Chr length (bp)	Telomere Status	Telomere Left Num	Telomere Left Direction	Telomere Right Num	Telomere Right Direction
Chr1A	33,093,563	both	255	+	214	–
Chr2A	33,995,934	both	240	+	263	–
Chr3A	37,997,701	both	251	+	199	–
Chr4A	30,070,748	both	229	+	204	–
Chr5A	33,145,425	both	213	+	220	–
Chr6A	39,608,517	both	249	+	256	–
Chr7A	32,144,107	both	284	+	231	–
Chr1B	28,000,368	both	306	+	222	–
Chr2B	30,378,977	both	459	+	243	–
Chr3B	35,457,332	both	212	+	229	–
Chr4B	29,079,538	both	260	+	236	–
Chr5B	31,324,413	both	229	+	257	–
Chr6B	39,010,043	both	284	+	223	–
Chr7B	28,743,846	both	199	+	141	–

Table 3. Summary statistics of *R.alceifolius* Poir. T2T genome assembly and telomeres information.

repeat library. The predicted sequences were subsequently merged with the RepBase database[34]. RepeatMasker (v4.1.2)[35] was then utilized to systematically characterize repetitive elements across the entire genome. The *R.alceifolius* Poir. genome consists of multiple classes of repetitive sequences, including DNA transposons (0.98%), LINEs (0.82%), LTR retrotransposons (20.45%), and simple repeats (1.01%) (Table 4).

Type	Num	Length (bp)	Percentage in Genome (%)
DNA transposons	9,238	4,534,536	0.98
LINE	8,984	3,804,512	0.82
LTR	117,983	94,503,745	20.45
Simple repeats	112,792	4,647,778	1.01
Unknown	370,254	100,214,730	21.69
Total	619,251	207,705,301	44.95

Table 4. Summary statistics of repetitive elements of *R.alceifolius* Poir. genome.

Features	Value
Number of Genes	46,698
Average Length of Genes (bp)	3409.37
Number of Exons	400,412
Average Length of Exons (bp)	301.38
Number of CDS	67,955
Average Lengthof CDS (bp)	230.60

Table 5. Summary statistics of protein-coding genes.

Metrics	BUSCO	Compleasm
Complete and Single-copy Gens	11.75%	15.65%
Complete and Duplicated Genes	87.01%	83.15%
Fragmented Genes	0.11%	0.34%
Missing Genes	1.13%	0.86%

Table 6. BUSCO and Compleasm result of protein-coding genes.

Protein-coding genes identification and functional annotation. Quality-controlled RNA-seq data were aligned to the repeat-softmasked genome using HISAT2 (v2.2.1)[36] and assembled into transcripts with StringTie (v2.15.2)[37]. Full-length transcript isoforms were clustered and refined using IsoSeq3 (v4.0.0)[38] to generate high-accuracy isoform sequences. Transcript evidence from both genome-guided Trinity (v2.15.2)[39] assemblies and full-length isoforms were integrated through PASApipeline (v2.5.3)[40] for comprehensive alternative splicing analysis. TransDecoder (v5.7.1) was employed to identify complete ORFs. For protein homology evidence, Rosaceae protein sequences were aligned to the genome using miniprot (v0.14)[41], retaining only high-confidence matches (>50% coverage). These evidence sources, combined with *Ab initio* predictions from Braker3 (v3.0.8)[42], were consolidated into weighted consensus gene models using EVidenceModeler (v2.0.0)[40]. Final gene structures were optimized via PASApipeline, yielding 46,698 protein-coding genes (mean length: 3,409.37 bp) comprising 400,412 exons (mean: 301.38 bp) and 67,955 coding sequences (mean CDS length: 230.60 bp) (Table 5). We found that the gene density on the long arms of chromosome 1 is higher than that on the short arms. This uneven distribution is likely related to the position of the centromere: regions closer to the centromere generally show lower gene density, whereas distal telomeric regions tend to be gene-rich. Because the centromere of chromosome 1 is located near the chromosomal end, the short arm does not fully regain the typical telomeric enrichment of genes, resulting in an apparent imbalance between the long and short arms. Assessment with BUSCO (v5.8.3)[43] and Compleasm (v0.2.6)[44] demonstrated 98.8% completeness of protein-coding genes (Table 6).

Protein-coding genes were functionally annotated using integrated protein databases including KEGG (https://www.kegg.jp/), GO (https://geneontology.org/), KOG (https://mycocosm.jgi.doe.gov/help/kogbrowser.jsf), eggNOG (http://eggnog5.embl.de/#/app/home), SwissProt (https://www.uniprot.org), and NR (https://ftp.ncbi.nlm.nih.gov/blast/db/). Reliable functional annotations were assigned to 43,851 genes (93.90% of all predicted protein-coding genes) (Table 7).

Non-coding RNA annotation. tRNAscan-SE (v2.0.12)[45] systematically identified tRNA genes with their structural features, while Infernal (v1.1.5)[46] was employed with the Rfam database to annotate other non-coding RNA families (rRNAs, snRNAs, and miRNAs). A total of 2,908 non-coding RNAs were identified, accounting for 0.28% (1.28 Mb) of the genome. These include 204 miRNAs, 841 tRNAs, 1,100 rRNAs, and 763 snRNAs. Notably, all ncRNAs were found to be complete, with no truncated sequences detected (Table 8).

Databases	Annotated Genes Number
KEGG	20,239
GO	21,479
KOG	21,966
eggNOG	41,882
SwissProt	30,823
NR	43,828
All	43,851

Table 7. Summary statistics of protein-coding genes.

Type	Annotated Genes Number	Total Length (bp)	Truncated Percentage (%)
tRNA	841	82,344	0
miRNA	204	25,604	0
rRNA	1,100	1,093,787	0
snRNA	763	81,227	0
Total	2,908	1,282,962	0

Table 8. Summary statistics of non-coding genes.

Subgenome identification. Due to the limited availability of high-quality genomic resources in *Rubus*, the precise ancestral origins and subgenome identification of allotetraploid *R. alceifolius* Poir. remain challenging to trace. To address this, we implemented a phylogenetic localization strategy to resolve subgenomic composition[47]. Homologous chromosome pairs were identified using the JCVI (v1.5.6)[48], with the diploid *Fragaria vesca* 'Yellow Wonder' genome[49] serving as the reference (Fig. 4A). Single-copy orthologous genes were identified using OrthoFinder (v3.0.1b1)[50], followed by multiple sequence alignment with MAFFT (v7.525)[51]. Conserved sites were filtered using Gblocks (v0.91b)[52], and conserved sequences from each chromosome were concatenated. Neighbor-joining phylogenetic trees were constructed in MEGA (v12)[53] using 'Yellow Wonder' as the outgroup to determine evolutionary relationships. Based on molecular evolutionary rate divergence, subgenomes were systematically assigned: those with shorter branch lengths were designated subgenome A, while those with longer branches were assigned subgenome B (Fig. 4B).

Subgenome synteny and variants detection. Comparisons between subgenomes A and B were conducted using MUMmer4 (v4.0.1)[54] and SyRI (v1.5.4)[55], with structural variants visualized via plotsr[56]. This analysis identified 1,270 homologous regions spanning 51.57% of the genome (Table 9). A prominent large-scale inversion (~12.45 Mb; coordinates: 6,647,243–19,094,252) was detected on chromosome Chr2 of subgenome A, which has not been described in previous research (Fig. 3A, Fig. 5A). To validate this structural variant, chromatin architecture was reconstructed for both subgenomes using HiC-Pro (v3.1.0)[57], with spatial interaction patterns visualized through HiCPlotter[58]. The HiC contact matrix revealed continuous interaction patterns within the region, with no disruptive signals at inversion boundaries, confirming this as a real structural variant (Fig. 5B).

Data Records

All high-throughput sequencing data generated in this study have been deposited in the National Center of Biotechnology Information (NCBI) database under the project accession number PRJNA1288027[59]. The genome assembly data (PKU_CYXGZ_1.0) is available under the accession number GCA_051990025.1. The original sequencing data from the genome libraries are available in the Sequence Read Archive (SRA) under the accession number SRP598905 comprising the following datasets: SRR34413162 (HiFi), SRR34413156 (Hi-C), SRR34413161 (ONT), SRR34413155 (Illumina), SRR34413149-SRR34413154, SRR34413158-SRR34413160 (RNA-seq) and SRR34413157 (ISO-seq). The annotation files are available from the figshare archive[60].

Technical Validation

To systematically evaluate the assembly quality of the *Rubus alceifolius* Poir. gap-free T2T genome, we employed a multidimensional quality assessment framework. Integrity metrics demonstrated exceptional completeness: Compleasm (v0.2.6)[44] assessed genome integrity at 99.39%, while BUSCO (v5.8.3)[43] achieved 99.1%. Continuity assessment indicated robust retrotransposon assembly, with a long terminal repeat retrotransposon assembly index (LAI)[61] of 27.12, significantly exceeding the gold standard threshold (LAI $\geq$ 20). Sequencing data support further validated assembly fidelity, with HiFi and ONT reads mapping rates[62] reaching 99.50% and 100.00%, respectively. Accuracy was confirmed by a genome quality value (QV)[63] of 61.66 (equivalent to a base error rate $< 10^{-6}$). Collectively, these results establish that the T2T assembly achieves high-quality standards across integrity, continuity, and accuracy dimensions (Table 10).

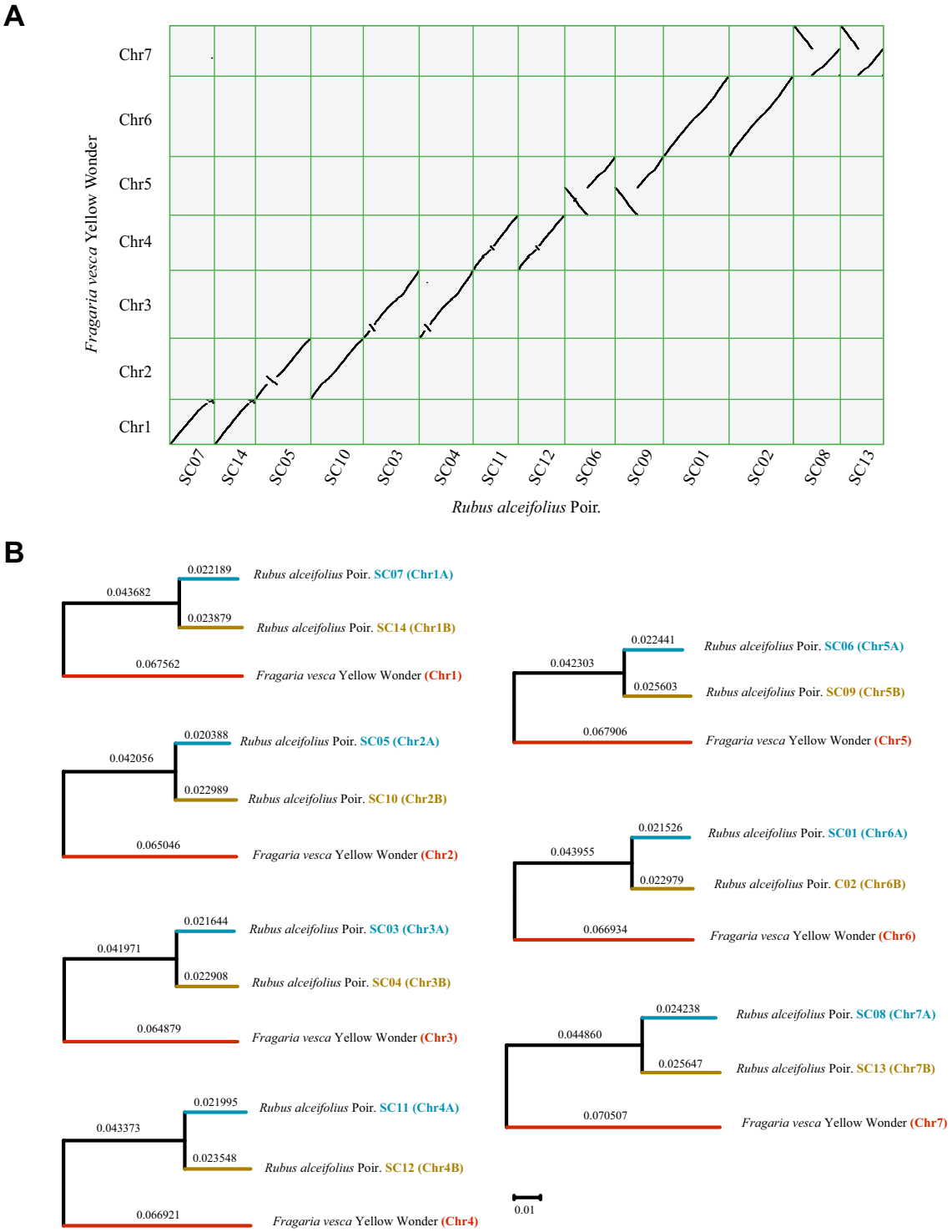

Fig. 4 *R. alceifolius* Poir. subgenome identification. (**A**) Chromosome synteny with diploid strawberry *Fragaria vesca* 'Yellow Wonder'. (**B**) Phylogenetic tree based on chromosome differentiation.

Data availability

The dataset has been deposited to National Center of Biotechnology Information (NCBI) as a BioProject under accession number PRJNA1288027.The T2T gap-free genome assembly of Rubus alceifolius Poir (PKU_CYXGZ_1.0) has been deposited in the NCBI GenBank database under the accession number GCA_051990025.1. Raw reads have been deposited under SRAs: SRR34413162 (HiFi), SRR34413156 (Hi-C), SRR34413161 (ONT), SRR34413155 (Illumina), SRR34413149-SRR34413154, SRR34413158-SRR34413160 (RNA-seq) and SRR34413157 (ISO-seq). The annotation files are available from the figshare archive (https://doi.org/10.6084/m9.figshare.31227940).

Type	Count	Subgenome A Length(bp)	Subgenome B Length(bp)
Syntenic regions	1,270	119,778,909	118,506,868
Inversions	38	15,717,861	15,412,096
Translocations	1,858	5,890,211	5,910,513
Duplications (Subgenome A)	595	2,761,984	\
Duplications (Subgenome B)	818	\	2,999,513
SNPs	768,367	768,367	\
Insertions	32,981	\	179,010
Deletions	32,797	185,959	\
CopyGains	6	\	1,627
CopyLosses	4	1,977	\
Highly Diverged	14,090	114,854,751	113,300,238

Table 9. Synteny and structural variation between subgenomes.

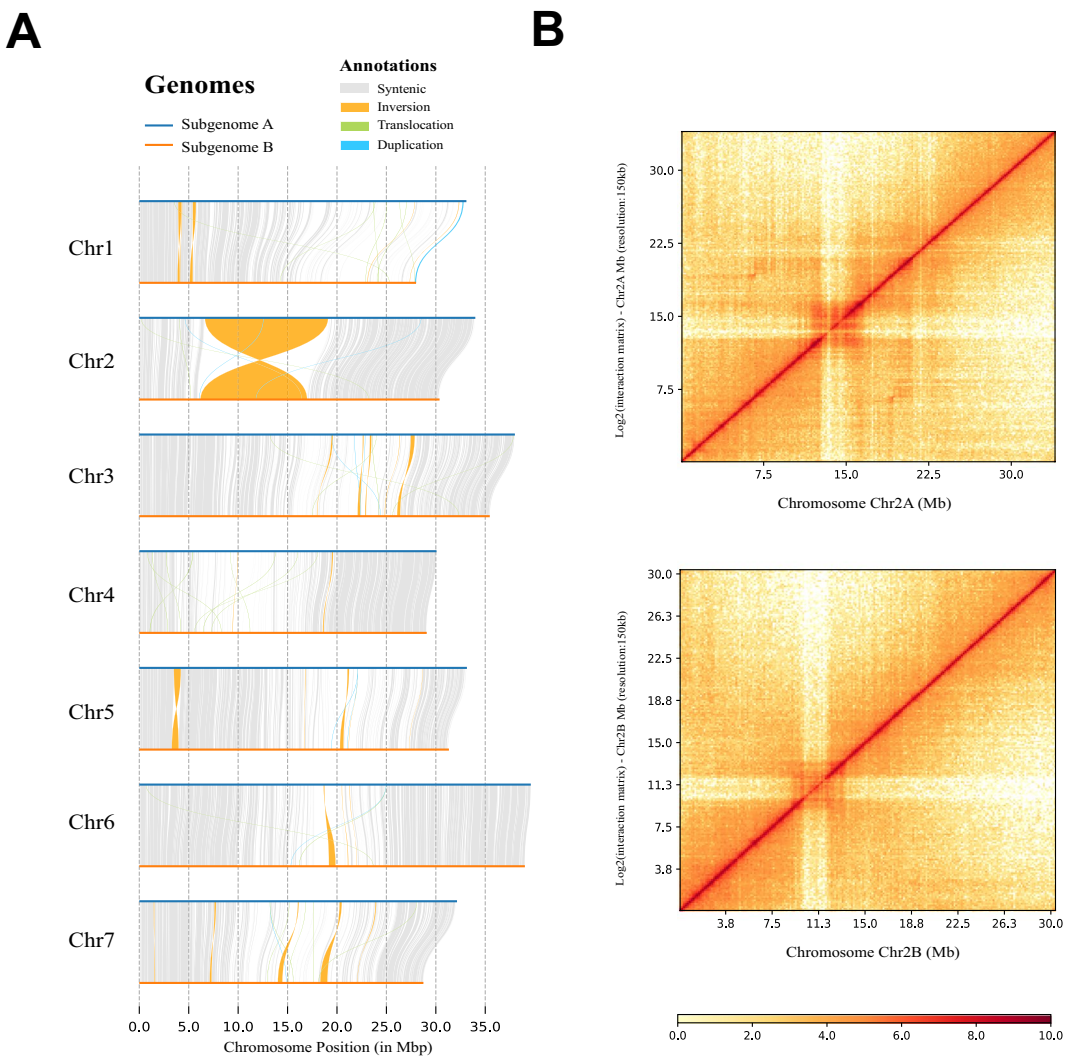

Fig. 5 Subgenome synteny and structural variants in *Rubus alceifolius Poir*. (**A**) Synteny between subgenomes A and B. (**B**) Hi-C contact matrix of Chr2A and Chr2B.

All supporting data have also been archived in the National Genomics Data Center (NGDC) under BioProject PRJCA042841. We confirm all data are accurate and available.

Metrics	Result
Compleasm	99.39%
BUSCO	99.1%
LAI	27.12
HiFi mapping	99.50%
ONT mapping	100.00%
QV	61.6623

Table 10. Summary statistics of *Rubus alceifolius* Poir. genome evaluation.

Code availability

Programs used in data processing were executed with default parameters unless otherwise specified in the Methods section. No custom code was employed for these analyses.

Received: 11 July 2025; Accepted: 27 March 2026;
Published online: 04 April 2026

References

1. Focke, W. O. *Species ruborum: Monographiae generis rubi prodromus*, (E. Schweizerbart, 1914).
2. Gu, Y., Jin, W., Zhao, C.M. & He, S.A. Rubus Resources in Fujian And Hunan Provinces. 117–126 (International Society for Horticultural Science (ISHS), Leuven, Belgium, 1993).
3. Hummer, K. E. Rubus Diversity. *HortScience HortSci* **31**, 182–183, https://doi.org/10.21273/HORTSCI.31.2.182 (1996).
4. FAO. Food and Agriculture Organization of the United Nations.. (FAOSTAT, 2023).
5. Moyer, R. A., Hummer, K. E., Finn, C. E., Frei, B. & Wrolstad, R. E. Anthocyanins, Phenolics, and Antioxidant Capacity in Diverse Small Fruits: Vaccinium, Rubus, and Ribes. *Journal of Agricultural and Food Chemistry* **50**, 519–525, https://doi.org/10.1021/jf011062r (2002).
6. Xue-yong, W. Research Advances on Raspberry and Blackberry. *Journal of Anhui Agricultural Sciences*, (2010).
7. Darrow, G. *Blackberry and raspberry improvement* (1937).
8. Thompson, M. M. Chromosome Numbers of Rubus Species at the National Clonal Germplasm Repository. *HortScience HortSci* **30**, 1447–1452, https://doi.org/10.21273/hortsci.30.7.1447 (1995).
9. Meng, Q., Manghwar, H. & Hu, W. Study on Supergenus Rubus L.: Edible, Medicinal, and Phylogenetic Characterization. *Plants (Basel)* **11**, https://doi.org/10.3390/plants11091211 (2022).
10. Amsellem, L., Noyer, J. L. & Hossaert-McKey, M. Evidence for a switch in the reproductive biology of Rubus alceifolius (Rosaceae) towards apomixis, between its native range and its area of introduction. *Am J Bot* **88**, 2243–2251 (2001).
11. Baret, S., Nicolini, E., Le Bourgeois, T. & Strasberg, D. Developmental patterns of the invasive bramble (Rubus alceifolius Poiret, Rosaceae) in Reunion island: an architectural and morphometric analysis. *Ann Bot* **91**, 39–48, https://doi.org/10.1093/aob/mcg006 (2003).
12. Amsellem, L., Noyer, J. L., Le Bourgeois, T. & Hossaert-McKey, M. Comparison of genetic diversity of the invasive weed Rubus alceifolius poir. (Rosaceae) in its native range and in areas of introduction, using amplified fragment length polymorphism (AFLP) markers. *Mol Ecol* **9**, 443–455, https://doi.org/10.1046/j.1365-294x.2000.00876.x (2000).
13. Amsellem, L., Chevallier, M. H. & Hossaert-McKey, M. Ploidy level of the invasive weed Rubus alceifolius (Rosaceae) in its native range and in areas of introduction. *Plant Systematics and Evolution* **228**, 171–179 (2001).
14. Laldinchhana, Rajat, S. D., Kawl, T., Hauzel, L. & Lalduhsanga, P. Evaluation of the Food and Nutrition Value of Rubus alceifolius Poir. Fruits of Mizoram, India. *Current Nutrition & Food Science* **16**, 554–562, https://doi.org/10.2174/1573401315666190502162837 (2020).
15. Zhao, J. *et al*. Total alkaloids of Rubus alceifolius Poir inhibit tumor angiogenesis through suppression of the Notch signaling pathway in a mouse model of hepatocellular carcinoma. *Mol Med Rep* **11**, 357–361, https://doi.org/10.3892/mmr.2014.2702 (2015).
16. Tran, C. H. *et al*. Optimization Conditions of Ultrasound-Assisted Extraction for Phenolic Compounds and Antioxidant Activity from Rubus alceifolius Poir Leaves. *Int J Food Sci* **2023**, 7576179, https://doi.org/10.1155/2023/7576179 (2023).
17. POWO. Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew. Published on the Internet;. Vol. 2025 (2025).
18. Sim, S. B., Corpuz, R. L., Simmonds, T. J. & Geib, S. M. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC Genomics* **23**, 157, https://doi.org/10.1186/s12864-022-08375-1 (2022).
19. Stanojević, D., Lin, D., Nurk, S., Florez de Sessions, P. & Šikić, M. Telomere-to-Telomere Phased Genome Assembly Using HERRO-Corrected Simplex Nanopore Reads. *bioRxiv*, 2024.2005.2018.594796 https://doi.org/10.1101/2024.05.18.594796 (2024).
20. Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048, https://doi.org/10.1093/bioinformatics/btw354 (2016).
21. Chu, J., Rong, J., Feng, X. & Li, H. ntsm: an alignment-free, ultra-low-coverage, sequencing technology agnostic, intraspecies sample comparison tool for sample swap detection. *Gigascience* **13**, https://doi.org/10.1093/gigascience/giae024 (2024).
22. 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, https://doi.org/10.1038/s41467-020-14998-3 (2020).
23. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204, https://doi.org/10.1093/bioinformatics/btx153 (2017).
24. 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, https://doi.org/10.1038/s41592-020-01056-5 (2021).
25. Rautiainen, M. *et al*. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* **41**, 1474–1482, https://doi.org/10.1038/s41587-023-01662-6 (2023).
26. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**, 540–546, https://doi.org/10.1038/s41587-019-0072-8 (2019).
27. Zhang, H. *et al*. Fast alignment and preprocessing of chromatin profiles with Chromap. *Nat Commun* **12**, 6566, https://doi.org/10.1038/s41467-021-26865-w (2021).
28. Zhou, C., McCarthy, S.A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, https://doi.org/10.1093/bioinformatics/btac808 (2023).
29. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).

30. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic Res* **10**, uhad127, https://doi.org/10.1093/hr/uhad127 (2023).
31. Alonge, M. *et al*. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol* **23**, 258, https://doi.org/10.1186/s13059-022-02823-7 (2022).
32. Hu, J., *et al*. NextPolish2: A Repeat-aware Polishing Tool for Genomes Assembled Using HiFi Long Reads. *Genomics Proteomics Bioinformatics* **22**, https://doi.org/10.1093/gpbjnl/qzad009 (2024).
33. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
34. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11, https://doi.org/10.1186/s13100-015-0041-9 (2015).
35. Tempel, S. Using and Understanding RepeatMasker. in *Mobile Genetic Elements: Protocols and Genomic Applications* (ed. Bigot, Y.) 29–51 (Humana Press, Totowa, NJ, 2012).
36. Thakur, V. RNA-Seq Data Analysis for Differential Gene Expression Using HISAT2-StringTie-Ballgown Pipeline. *Methods Mol Biol* **2812**, 101–113, https://doi.org/10.1007/978-1-0716-3886-6_5 (2024).
37. Shumate, A., Wong, B., Pertea, G. & Pertea, M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLoS Comput Biol* **18**, e1009730, https://doi.org/10.1371/journal.pcbi.1009730 (2022).
38. Guizard, S., *et al*. nf-core/isoseq: simple gene and isoform annotation with PacBio Iso-Seq long-read sequencing. *Bioinformatics* **39**, https://doi.org/10.1093/bioinformatics/btad150 (2023).
39. Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644–652, https://doi.org/10.1038/nbt.1883 (2011).
40. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
41. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, https://doi.org/10.1093/bioinformatics/btad014 (2023).
42. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644, https://doi.org/10.1093/bioinformatics/btn013 (2008).
43. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**, 227–245, https://doi.org/10.1007/978-1-4939-9173-0_14 (2019).
44. Huang, N. & Li, H. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics* **39**, https://doi.org/10.1093/bioinformatics/btad595 (2023).
45. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res* **49**, 9077–9096, https://doi.org/10.1093/nar/gkab688 (2021).
46. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935, https://doi.org/10.1093/bioinformatics/btt509 (2013).
47. Zhang, J. *et al*. The genome of Cleistogenes songorica provides a blueprint for functional dissection of dimorphic flower differentiation and drought adaptability. *Plant Biotechnol J* **19**, 532–547, https://doi.org/10.1111/pbi.13483 (2021).
48. Tang, H. *et al*. JCVI: A versatile toolkit for comparative genomics analysis. *Imeta* **3**, e211, https://doi.org/10.1002/imt2.211 (2024).
49. Joldersma, D., Sadowski, N., Timp, W. & Liu, Z. Assembly and annotation of Fragaria vesca 'Yellow Wonder' genome, a model diploid strawberry for molecular genetic research. *Fruit Research* **2**, 1–5, https://doi.org/10.48130/FruRes-2022-0013 (2022).
50. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238, https://doi.org/10.1186/s13059-019-1832-y (2019).
51. Katoh, K. & Standley, D. M. MAFFT: iterative refinement and additional methods. *Methods Mol Biol* **1079**, 131–146, https://doi.org/10.1007/978-1-62703-646-7_8 (2014).
52. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564–577, https://doi.org/10.1080/10635150701472164 (2007).
53. Kumar, S., *et al*. MEGA12: Molecular Evolutionary Genetic Analysis Version 12 for Adaptive and Green Computing. *Mol Biol Evol* **41**, https://doi.org/10.1093/molbev/msae263 (2024).
54. Marcais, G. *et al*. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* **14**, e1005944, https://doi.org/10.1371/journal.pcbi.1005944 (2018).
55. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* **20**, 277, https://doi.org/10.1186/s13059-019-1911-0 (2019).
56. Goel, M. & Schneeberger, K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics* **38**, 2922–2926, https://doi.org/10.1093/bioinformatics/btac196 (2022).
57. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**, 259, https://doi.org/10.1186/s13059-015-0831-x (2015).
58. Akdemir, K. C. & Chin, L. HiCPlotter integrates genomic data with interaction matrices. *Genome Biol* **16**, 198, https://doi.org/10.1186/s13059-015-0767-1 (2015).
59. *NCBI Bioproject* https://www.ncbi.nlm.nih.gov/bioproject/1288027 (2025).
60. Zheng, M. T2T genome assembly of Rubus alceifolius Poir. *figshare* https://doi.org/10.6084/m9.figshare.31227940 (2025).
61. Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol* **176**, 1410–1422, https://doi.org/10.1104/pp.17.01310 (2018).
62. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
63. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).

Acknowledgements

We thank Taishan Scholars Program (J.Z.), the Dean's Foundation from Peking University Institute of Advanced Agricultural Sciences (PKU-IAAS), the funding from Shandong Laboratory of Advanced Agricultural Sciences at Weifang (SYS202206), the funding from Guizhou Provincial Germplasm Repository of Berry Fruits and Guizhou Provincial Scientific and Technological Program (QKHFQ[2024]004-1) for financial support.

Author contributions

J.Z. conceived and supervised the study; X.H., Y.S. and J.G. collected samples; G.S. performed the experiments; M.Z., S.L. and X.L. conducted the genome assembly and bioinformatics analysis; M.Z. drafted the manuscript, J.Z. revised the manuscript, and all authors read and approved the final manuscript.

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/s41597-026-07160-z.

Correspondence and requests for materials should be addressed to J.Z.

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.