scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly and annotation of the endangered plant *Primula kwangtungensis* (Primulaceae)

Tongjian Liu[1,8], Yunling Xiao[1,2,8], Xing Wu[1,3,8], Sheng Chen[4], Sisi Chen[5], Itay Mayrose[6], Gang Hao[5] ✉, Ruijiang Wang[1,7] ✉ & Yuan Xu[1,3] ✉

***Primula kwangtungensis*, a perennial herbaceous plant endemic to the Nanling Mountains in southern China, is rare and sparsely distributed, making it a conservation priority. This study reports a high-quality chromosome-level genome assembly of approximately 1.29 Gb, achieved through BGI-DIPSEQ, Nanopore, and Hi-C sequencing. Hi-C data anchored 88.68% of contigs to nine chromosomes, with contig N50 of 5.87 Mb and scaffold N50 of 144.82 Mb. The assembly achieved a BUSCO completeness score of 95.8%, indicating high accuracy. A total of 31,717 protein-coding genes were identified, with 97.46% functionally annotated, and repetitive sequences accounted for 63.25% of the genome. This assembly provides crucial data for understanding species diversity mechanisms in the Nanling region and the comparative genomics studies among Primulaceae. It offers essential genetic resources for biodiversity research and aids in developing conservation strategies for rare and endangered plant species.**

Background & Summary

The Nanling Mountains, located in Southern China, are characterized by unique geological formations and represent a significant biodiversity hotspot, providing essential habitats for numerous species[1]. Among the diverse plant species in this region, members of the genus *Primula* are particularly notable for their floral dimorphism known as distyly[2]. Distyly is a reproductive strategy where two distinct floral morphs coexist within a population, differing in the reciprocal positioning of anthers and stigmas: pin (long-styled) flowers and thrum (short-styled) flowers[3]. *Primula kwangtungensis* W. W. Smith stands out as a typical distylous plant (Fig. 1b, c), making it an important model for studying the genetic and evolutionary mechanisms of this reproductive strategy. In addition, the species was first described in 1937 and belongs to the family Primulaceae and the section *Carolinella* of the genus *Primula*[4]. As a perennial herb, *P. kwangtungensis* predominantly grows in the karst limestone and Danxia landscapes of the Nanling Mountains (Fig. 1a). These distinctive habitats provide a unique ecological niche but also severely limit its distribution. This species is exceedingly rare in the wild, with no confirmed collections for over seventy years. It wasn't until 2013, during a field investigation, that we rediscovered and collected specimens of this species on the central Nanling Mountains.

Subsequent studies revealed that over half of the remaining wild populations consist of less than 50 mature individuals, severely threatening its survival[5]. According to the "50/500 rule" (i.e., the Minimum Viable Population theory), when the effective population size (Ne) is less than 50, the species faces a higher

[1]Key Laboratory of National Forestry and Grassland Administration on Plant Conservation and Utilization in Southern China, Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, P. R. China. [2]University of Chinese Academy of Sciences, Beijing, 100049, P. R. China. [3]South China National Botanical Garden, Guangzhou, 510650, P. R. China. [4]Guizhou Institute of Biology, Guiyang, 550009, P. R. China. [5]College of Life Sciences, South China Agricultural University, Guangzhou, 510642, P. R. China. [6]School of Plant Sciences and Food Security, Tel Aviv University, Ramat Aviv, Tel Aviv, 69978, Israel. [7]State Key Laboratory of Plant Diversity and Specialty Crops, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, P. R. China. [8]These authors contributed equally: Tongjian Liu, Yunling Xiao, Xing Wu. ✉e-mail: haogang@scau.edu.cn; wangrj@scbg.ac.cn; xuyuan@scbg.ac.cn

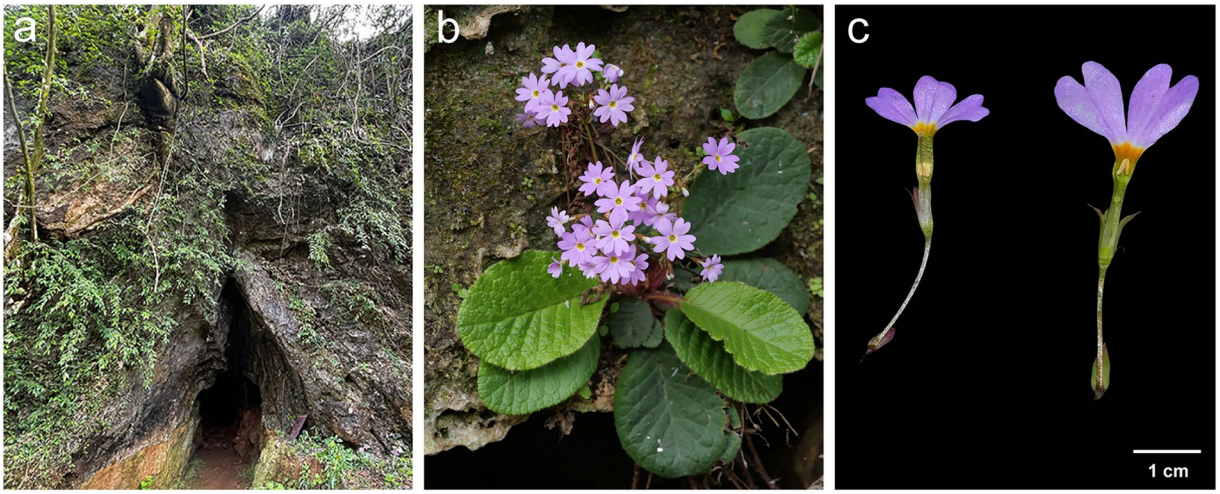

Fig. 1 The habitat and habit of *Primula kwangtungensis*. (**a**) Typical habitat. (**b**) Flowering individual. (**c**) Long-styled (left) and short-styled (right) flowers.

risk of inbreeding depression[6]. Based on current data and criteria established by the International Union for Conservation of Nature (IUCN), *P. kwangtungensis* is classified as "Critically Endangered" (CR)[7]. Although some research has focused on the phylogenetic relationship[8], pollen morphology[9], and chloroplast genome characteristics[5] of *P. kwangtungensis*, its complete genomic information and genetic traits remain poorly understood.

In this study, we have successfully assembled the chromosomal-level genome of *P. kwangtungensis* using a combined sequencing approach, including Oxford Nanopore Technologies (ONT), BGI-DIPSEQ, and Hi-C methods. The final assembled genome has a total length of approximately 1.29 Gb, with a scaffold N50 of 144.82 Mb. Using Hi-C data, scaffolds were further clustered and ordered into nine pseudochromosomes ($2n = 18$, ranging in length from 132.6 Mb to 153.8 Mb). In BUSCO evaluations, using the eudicots_odb10 single-copy ortholog datasets, the genome exhibited 95.8% completeness scores. Repetitive DNA sequences account for 63.25% of the total genome length, and a total of 31,717 protein-coding genes were annotated, with 97.46% of these genes functionally annotated. According to the published genomic data of species of Primulaceae *s.l.*, the genome size, scaffold N50, and contig N50 of *P. kwangtungensis* represent the largest values among them (Table S1)[10–14].

In summary, given its unique ornamental value, distinct habitat, and status as an endemic species of the Nanling Mountains, *P. kwangtungensis* serves as an ideal model for studies on species adaptive evolution. The analysis of its whole-genome characteristics not only provides crucial foundational data for understanding the mechanisms of species diversity in the Nanling region but also offers essential information for future research into the genetic architecture of the endangerment of *P. kwangtungensis* and evolution of the *Primula* genus.

Methods

Sample preparation, library construction and sequencing. In Yizhang County, located in Chenzhou City, Hunan Province, China, on the central Nanling Mountains, we collected a range of fresh samples from *P. kwangtungensis* for DNA and RNA extraction. These included mature leaves, roots, long-styled and short-styled flowers, and fruits. All samples were selected from healthy, disease-free plants. After collection, the samples were immediately wrapped in aluminum foil, labeled for identification, and then stored in liquid nitrogen.

For ONT library construction and sequencing, DNA was extracted from fresh, young leaves of *P. kwangtungensis*. The plant used for genome sequencing was a long-styled flower individual. The leaves were ground into a fine powder using liquid nitrogen to maintain the biological integrity and preserve DNA quality. High-quality DNA libraries were constructed with the LSK108 kit (SQK-LSK108, Oxford), and sequencing was carried out on a Nanopore GridION X5 sequencer using five flow cells for real-time single-molecule sequencing[15]. This approach provided high throughput and long read lengths, resulting in approximately 104 Gb of sequence data.

For short-read sequencing, DNA was extracted from fresh young leaves of *P. kwangtungensis* using the CTAB method[16] and subsequently sequenced on the BGI-DIPSEQ platform, producing approximately 147 Gb of 100 bp paired-end reads with an insert size of around 250 bp[17]. To ensure the accuracy of subsequent analyses including genome size estimation and ONT assembly polishing, only high-quality reads were retained for further processing. These reads were defined as those trimmed of adapters and quality-filtered with Fastp v0.23.1[18] under thresholds of Phred score ≥ 20, N content $\leq 5\%$, and length ≥ 50 bp, with additional validation by FastQC v0.12.1[19]. For genome annotation, total RNA was extracted from various tissues of *P. kwangtungensis* using the TIANGEN kit with DNase I. Paired-end libraries with a 250 bp insert size were constructed using the NEBNext Ultra™ RNA Library Prep Kit, and sequenced on the BGI-DIPSEQ platform, generating about 6 Gb of 100 bp paired-end data per tissue. The raw data were filtered for quality using Trimmomatic v0.39[20], with the following parameters: ILLUMINACLIP:TruSeq 3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36.

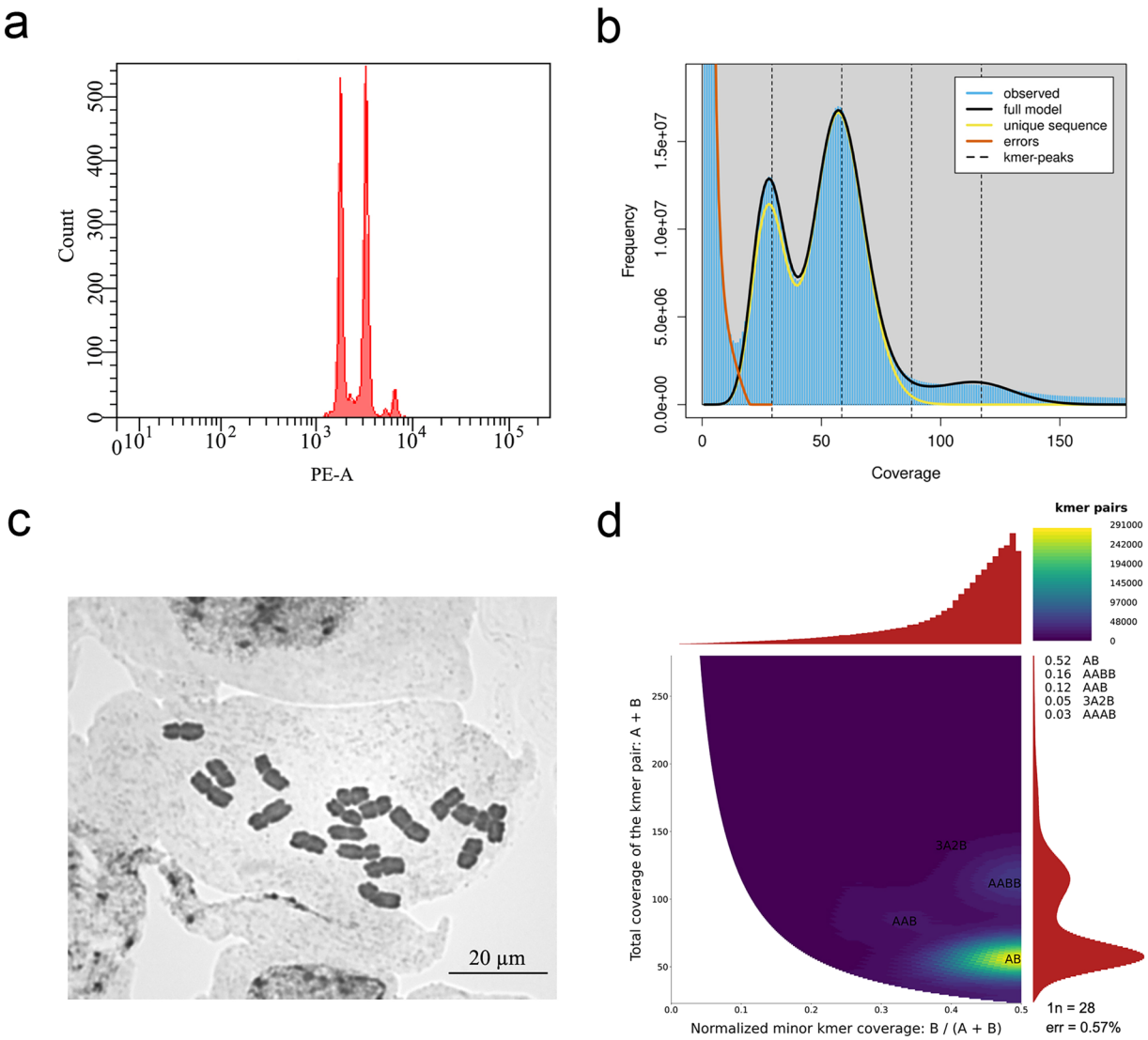

Fig. 2 Genomic survey assessment of *Primula kwangtungensis*. (**a**) Flow cytometry result for *P. kwangtungensis*, using *Solanum lycopersicum* as the internal reference. (**b**) Genome survey based on k-mer (k = 21) analysis. (**c**) Photomicrograph of somatic metaphase chromosomes. (**d**) Assessment of the ploidy level of *P. kwangtungensis* using Smugeplot.

Employing Hi-C technology for library construction and sequencing[21], we aimed to achieve a chromosome-level high-quality genome of *P. kwangtungensis*. Plant cells were first fixed with formaldehyde to preserve DNA-protein interactions, followed by cell lysis. DNA was then digested with the restriction enzyme DpnII, which recognizes the GATC site. Biotin labels were introduced post-digestion to link adjacent DNA fragments. After lysis and digestion, the Hi-C library was carefully quality-controlled to meet the required standards. High-throughput sequencing was performed on the Illumina platform, generating approximately 128 Gb of 150 bp paired-end read data.

Genome size, chromosome ploidy level estimation. Flow cytometry[22] and k-mer analysis[23] were employed to estimate the genome size of *P. kwangtungensis*. Flow cytometry, with *Solanum lycopersicum*[24] as the reference species, utilized a mixed detection approach, recording a minimum of 1128 nuclei events for *P. kwangtungensis* and 1047 events for the tomato reference, with coefficients of variation (CV%) less than 6% in all three biological replicates. This resulted in a genome size estimate of 1.46–1.51 Gb (Fig. 2a; Table S2). Additionally, high-quality reads from BGI-DIPSEQ were filtered using SOAPnuke v2.1.8[25], followed by genome survey analysis with 21-mer frequency distribution in Jellyfish v2.3.0[26] and GenomeScope v2.0[26]. The results indicated a genome size of approximately 1.25 Gb, a heterozygosity rate of 1.02%, and repeat content of 57.2% (Fig. 2b). Lastly, Smudgeplot v0.4.0[27] was employed to estimate ploidy, revealing that *P. kwangtungensis* is likely diploid, consistent with our cytological observations of chromosome count (2n = 2X = 18; Fig. 2c, d).

Draft genome assembly and chromosome-level genome assembly. For genome assembly, we initially employed raw long-read sequence data from the ONT platform for *de novo* genome assembly using NextDenovo

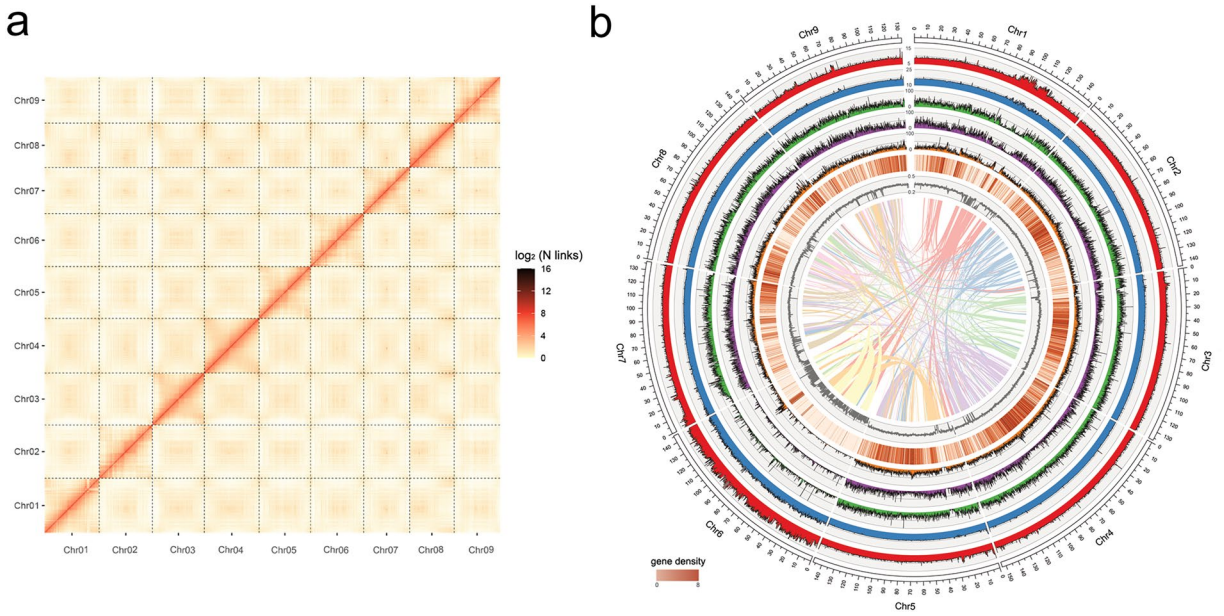

Fig. 3 Chromosomal contact map and genomic landscape of *Primula kwangtungensis*. (**a**) Hi-C contact heatmap showing the interaction frequency among the nine pseudochromosomes of *P. kwangtungensis*. The interaction intensity is represented by the color gradient from light yellow to dark red, corresponding to the $\log_2$-transformed number of Hi-C links. (**b**) Circos plot illustrating the genomic features of the *P. kwangtungensis* genome. From the outermost to the innermost circles: (1) chromosome coordinates (Mb); (2) ONT read coverage; (3) Illumina read coverage; (4) density of Copia elements; (5) density of Gypsy elements; (6) density of DNA transposon elements; (7) gene density; and (8) GC content. The innermost colored ribbons represent syntenic relationships among homologous chromosomal regions, revealing large-scale collinearity within the genome. All track-based statistics were calculated using 100-kb sliding windows.

v2.5.2[28]. During this process, the assembly quality of the contigs was optimized by adjusting the parameters "read_cutoff to 1,000 bases and seed_cutoff to 18,934 bases". To mitigate the high error rate associated with third-generation sequencing technologies, the assembled data from the previous step was polished in three rounds using high-accuracy short reads with NextPolish v1.4.1[29] to enhance the accuracy of the genomic sequences. Finally, redundant sequences in the genome assembly were removed using Purge dups v1.2.6[30] to obtain a draft genome.

Building upon this preliminary assembly, Hi-C data were integrated to achieve chromosome-level genome assembly. First, the raw Hi-C data were quality-filtered using Sickle v1.33[31] and the filtered reads were aligned to the reference genome using Bowtie2 v2.4.4[32]. Subsequently, interaction information with the reference genome was generated using Juicer v1.6.0[33]. The contigs were then subjected to two rounds of error assembly identification and correction utilizing the interaction information from the Hi-C data through 3D-DNA v180922[34]. Following this, the corrected contigs were clustered, oriented, and ordered to generate scaffolds. After the split step of the process, manual corrections were applied to the scaffolding using Juicebox v2.04.06[35] to ensure accuracy. Subsequent assembly steps were performed using 3D-DNA to obtain a chromosome-level genome assembly[34], with a size of approximately 1.29 Gb and a scaffold N50 of 144.82 Mb (Fig. 3a). Whole-genome Hi-C contact heatmap was generated with HiCPlotter[36]. The plot of the distribution of genomic elements was generated using Circos v0.69-8[37] (Fig. 3b). Syntenic blocks were identified using MCScanX v1.0.0[38] with default parameters requiring five collinear genes, following BLASTP v 2.13.0[39] alignment using an e-value cutoff of 1e-10 and retention of the top five hits.

Identification of repetitive elements. Annotation of transposable elements (TEs) was conducted using a combination of homology-based searches and *de novo* prediction approaches. For homology-based identification, RepeatMasker v4.0.9[40] was employed to recognize TEs using the Repbase database. TE prediction was performed by constructing a custom repeat library with RepeatModeler v1.0.4[41], followed by detection using RepeatMasker. This analysis was further complemented by the use of LTR_Finder v1.0.5[42] and RepeatScout v1.0.5[43]. Tandem repeats within the genome were identified using Tandem Repeats Finder v4.0.9[44]. As a result, 815.61 Mb of repetitive sequences were annotated in the *P. kwangtungensis* genome, accounting for 63.25% of the total genome. This is consistent with observations in most angiosperms, with long terminal repeats (LTR) retrotransposons constituting the largest proportion at 43.35%[45] (Table 1).

Gene structure prediction. Prediction of protein-coding genes was conducted using a comprehensive approach that integrated transcriptome-based prediction, *de novo* prediction, and homology-based methods. Initially, RNA-Seq reads were quality-filtered using Trimmomatic v0.36[20], followed by *de novo* assembly of the processed reads with Trinity v2.1.1[46] to generate transcriptome sequences. The assembly results were subsequently

Type	Number of elements	Length (bp)	Percentage of sequence (%)
Retroelements	504415	587759211	45.58
SINEs	2821	535428	0.04
LINEs	31351	28238592	2.19
LTR elements	470243	558985191	43.35
DNA transposons	137888	55143908	4.28
Rolling-circles	10026	4954633	0.38
Small RNA	3812	1282132	0.10
Satellites	1440	1110229	0.09
Simple repeats	163550	12444901	0.97
Low complexity	27912	1454551	0.11
Unclassified	612786	172709563	13.39
Total	1255089	815612682	63.25

Table 1. Summary of the repetitive sequences in *Primula kwangtungensis* genome.

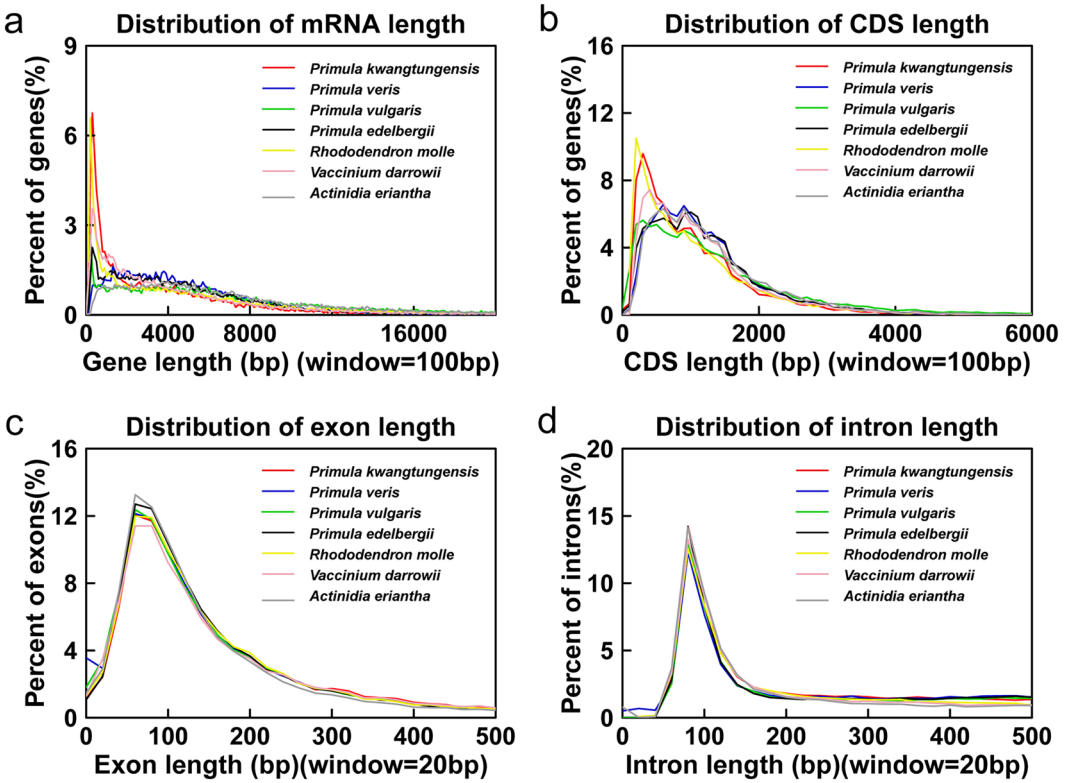

Fig. 4 Comparison of the distribution of gene elements for each gene among *Primula kwangtungensis* and six species in Ericales. (**a**) Gene length. (**b**) CDS length. (**c**) Exon length. (**d**) Intron length. The x-axis represents the length and the y-axis represents the density of genes.

spliced and aligned using PASA v2.4.1[47] to complete transcriptome-based gene prediction. Additionally, protein sequences from closely related species within the same order as *P. kwangtungensis* were downloaded from public databases to construct a custom homology protein database. To streamline the genome annotation process, BRAKER v3.0.6[48] was employed to predict gene models on the reference genome, which had been masked for repetitive sequences using RepeatMasker. This prediction process integrated transcriptome data, protein sequences from closely related species downloaded from the NCBI database or other data websites designated by researchers, including *P. veris*[10], *P. vulgaris*[11], *P. edelbergii*[13], *Rhododendron molle*[49] (GCA_025413875.1), *Vaccinium darrowii*[50] (GCA_020921065.1), and *Actinidia eriantha*[51] (GCF_019202715.1). To eliminate the interference of transposable elements, the predicted gene set was filtered using TEsorter 1.4.6[52], BRAKER3, and GlimmerHMM v3.0.4[53] to remove protein-coding genes containing transposable element sequences. Finally, EVidenceModeler v2.0.0[54] was used to integrate the three gene sets into a consensus gene set, which was further refined by PASA v2.5.0[54], resulting in a final high-quality gene set supported by *de novo* prediction, homologous proteins, and RNA-Seq data. Through comparative analysis of gene, CDS, exon, and intron lengths, as well as the average

Type	Database	Annotated number	Percent (%)	Annotated number and proportion
eggnog-mapper	GO	12032	37.94	26008 (82.00%)
	KEGG_EC	5343	16.85	
	KEGG_Pathway	7423	23.40	
	KEGG_Knum	11648	36.72	
	COG	9580	30.20	
	KOG	27119	85.50	
	Pfam_domain	25577	80.06	
Interproscan	Panther	26156	82.47	27596 (87.01%)
	Pfam	23302	73.47	
	Interpro	24741	78.01	
	RefSeq	24132	76.08	
Total	—	30912	97.46	—

Table 2. Numbers and proportions of functionally annotated protein-coding genes in the genome of *Primula kwangtungensis*.

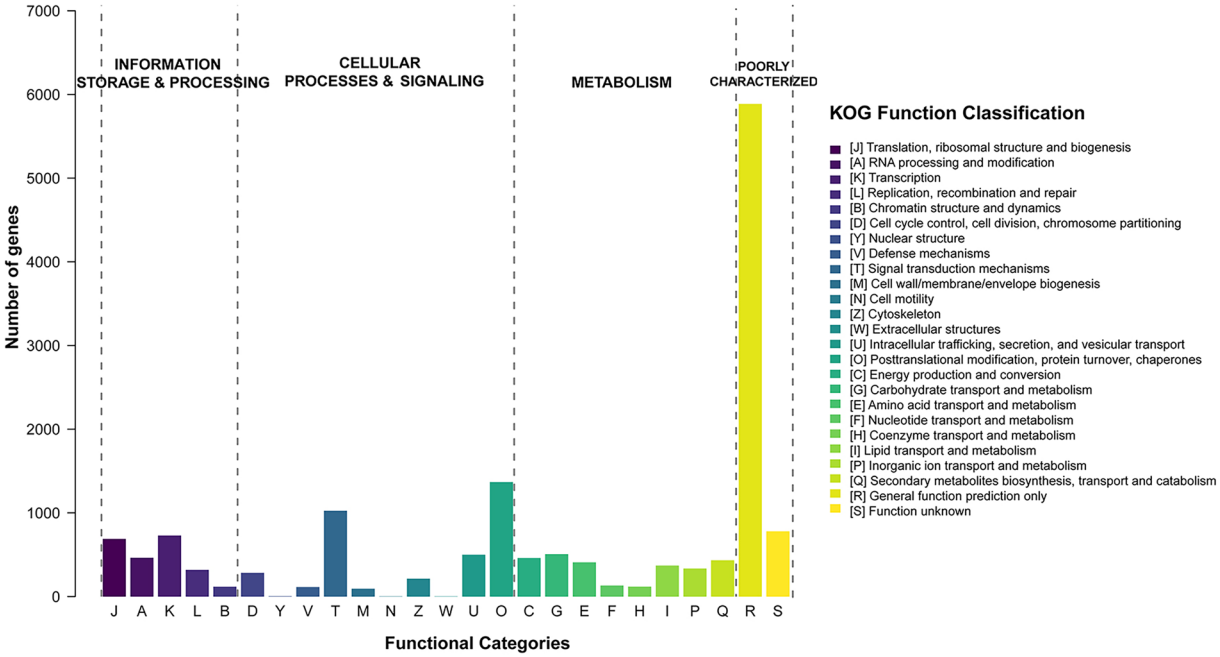

Fig. 5 Frequency distribution of each identified functional class in eukaryotic orthologous groups (KOG) annotation.

number of exons per gene, we found that the *P. kwangtungensis* genome exhibits distinct characteristics compared to the closely related species. Specifically, the average gene length in *P. kwangtungensis* is 3.93 Kb (Fig. 4a), the average CDS length is 1.10 Kb (Fig. 4b), and each gene contains an average of 4.69 exons. Additionally, the average exon length is approximately 0.23 Kb (Fig. 4c), while the average intron length is about 0.78 Kb (Fig. 4d). These findings provide valuable insights into the genomic architecture of *P. kwangtungensis* and its relatives.

Gene functional annotation. Functional annotation of protein-coding genes was performed using two methods. InterProScan v5.66-98[55], which integrates multiple public databases including Pfam, PRINTS, PANTHER, ProSite Profiles, and SMART, was used to identify conserved amino acid sequences, motifs, and domains. This resulted in the annotation of 27,596 genes, accounting for 87.01% of the predicted protein-coding genes. Additionally, eggNOG-mapper v2.0.0[56] was used to align predicted genes against the eggNOG database, which includes various databases such as GO, KEGG (KEGG_EC, KEGG_Pathway, KEGG_Knum), and Pfam_domain. This led to the functional annotation of 26,008 genes, representing 82.00% of the predicted protein-coding genes. We also established a local RefSeq database (updated 2025-05-05) and conducted MMseq2 searches to align *P. kwangtungensis* protein sequences against the RefSeq database. The top hits were retained after applying filtering thresholds of E-value < 0.001 and sequence similarity < 50%. In total, 30,912 genes were annotated across all databases, covering 97.46% of the predicted protein-coding genes (Table 2).

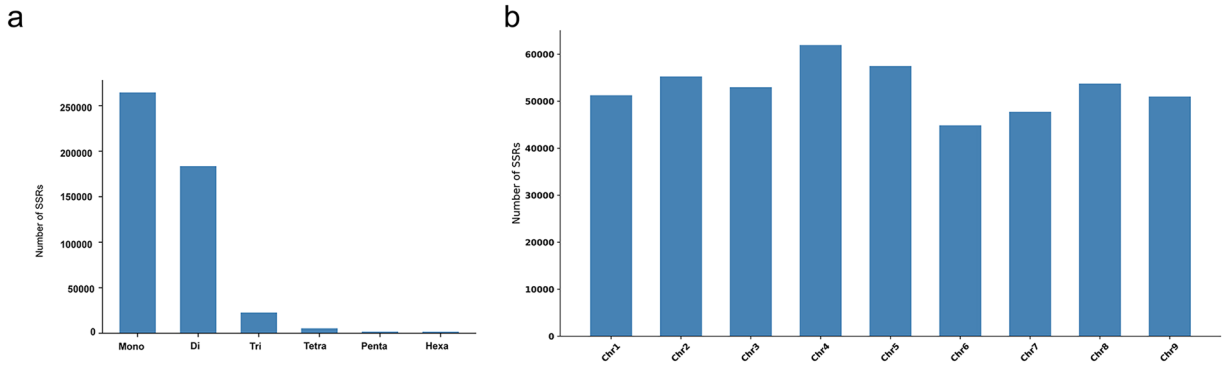

Fig. 6 Quantification and distribution of simple sequence repeats (SSRs). (**a**) Quantification of six types of SSRs. (**b**) Chromosomal distribution of SSRs across nine chromosomes.

Type	Description	Annotated Proteins Value	Ratio (%)
Genome assembly	Complete BUSCOs (C)	2228	95.8
	Complete and single-copy BUSCOs (S)	2127	91.4
	Complete and duplicated BUSCOs (D)	101	4.3
	Fragmented BUSCOs (F)	30	1.3
	Missing BUSCOs (M)	68	2.9
	Total BUSCO groups searched	2326	—
Genome annotation	Complete BUSCOs (C)	1556	96.4
	Complete and single-copy BUSCOs (S)	1504	93.2
	Complete and duplicated BUSCOs (D)	52	3.2
	Fragmented BUSCOs (F)	9	0.6
	Missing BUSCOs (M)	49	3.0
	Total BUSCO groups searched	1614	—

Table 3. BUSCO evaluation of genome assembly and annotated protein sequences of *Primula kwangtungensis*.

The protein-coding genes of *P. kwangtungensis* were compared with the KOG (EuKaryotic Orthologous Groups) and COG (Cluster of Orthologous Groups) databases[57]. KOG identifies protein sequences with similar functions through comparative analysis of complete genomes and forms superfamilies, each constituting a KOG category[58]. After removing redundant and low-similarity sequences (filtering parameter: 1e-5)[59], 27,119 genes (85.56%) were annotated in KOG and 9,580 genes (30.21%) in COG. Further functional classification of protein isoforms using the KOG database helps elucidate the functional characteristics of *P. kwangtungensis* genes (Fig. 5).

Microsatellite or simple sequence repeat (SSR) identification. Using the MISA tool[60], we performed a genome-wide search for simple sequence repeats (SSRs) across nine assembled chromosomes, with detection thresholds set to a minimum of six repeats for dinucleotides and five repeats for tri-, tetra-, penta-, and hexanucleotides. A total of 479,733 SSRs were identified, with the following distribution by motif type: mononucleotide repeats were the most abundant (264,524, 55.14%), followed by dinucleotide repeats (183,534, 38.25%), trinucleotide (22,765, 4.74%), tetranucleotide (5,382, 1.12%), pentanucleotide (1,804, 0.38%), and hexanucleotide repeats (1,724, 0.36%) (Fig. 6a). Spatial distribution analysis revealed 89,171 compound SSRs (18.58% of total SSRs), with all nine chromosomes harboring SSRs. Chromosome 4 contained the highest number of SSRs (61,942), while chromosome 6 had the lowest (44,871) (Fig. 6b).

Data Records

The raw sequencing data, encompassing Nanopore data (accession number: SRX26340527), BGISEQ short-reads (accession number: SRX26340528), Hi-C data generated using the Illumina NovaSeq platform (accession number: SRX26340529), and RNA-sequencing data (accession number: SRX26340530 - SRX26340535) which includes the long-styled and short-styled flowers, fruits, leaves (with two replicates), and roots of *P. kwangtungensis*, have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive[61]. The genome assembly and annotation files are available in the GenBank with the accession number GCA_054094925.1[62] as well as in Figshare database[63].

Technical Validation

Evaluation of the genome assembly. The quality and completeness of the *P. kwangtungensis* genome assembly were assessed through four distinct approaches. Initially, the genome assembly was evaluated using BUSCO v5.2.2[64] with the eudicots_odb10 dataset of single-copy orthologs. The assessment revealed that the genome encompassed 2,228 BUSCOs, achieving a completeness score of 95.8% (Table 3). Subsequently,

BWA-MEM v0.7.17[65] was employed to align BGI-DIPSEQ sequencing data to the assembled reference genome, achieving a mapping rate of 99.82%. The LTR Assembly Index (LAI), computed by LTR_retriever v3.0.1[66], was also used to assess the genome quality. The LAI score of 12.62 demonstrated that our assembled *P. kwangtungensis* genome exhibits high quality and completeness. Lastly, the Merqury[67] software was utilized to estimate the consensus QV of the assembly. Optimal k-mer length was determined as 19 via analysis with the "best_k.sh" script of the Merqury software. A k-mer spectrum of the whole genome was then constructed using BGI-DIPSEQ data with Meryl under default settings. Merqury evaluation, integrating the k-mer database with assembled sequences, showed a k-mer completeness of 83.24% and a QV score of 32.15 based on k-mer consistency.

Evaluation of the genome annotation. The genome annotation results were evaluated using BUSCO with the eudicots_odb10 database. The analysis showed that 96.4% of the predicted protein-coding genes are complete, with 93.2% being single-copy genes and 3.2% being duplicated genes (Table 3). Additionally, the length distributions of genes, CDS, exons, and introns in *P. kwangtungensis* were compared with those of other species in the Ericales order. This comparison indicated that the length distributions of these genomic features in *P. kwangtungensis* are similar to those of other species in the same order, indicating the high accuracy of the annotation results (Fig. 4).

Data availability

The raw sequencing data, including Nanopore, BGISEQ, Hi-C and RNA-sequencing data, generated in this study are available at https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA1156912. The genome assembly and annotation files are available at https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_054094925.1/ and https://doi.org/10.6084/m9.figshare.29479610.

Code availability

In this study, all commands and scripts used were executed in accordance with the manuals or protocols of the tools employed. The software and tools used are publicly accessible, and the versions and parameters are detailed in the "Methods" section. Default parameters were used in case specific arguments were not mentioned. No custom code was used in this research.

Received: 12 February 2025; Accepted: 24 December 2025;
Published online: 03 January 2026

References

1. Tang, Z., Wang, Z., Zheng, C. & Fang, J. Biodiversity in China's mountains. *Front. Ecol. Environ.* **4**, 347–352 (2006).
2. Darwin, C. On the Two Forms, or Dimorphic Condition, in the Species of *Primula*, and on their remarkable Sexual Relations. *Bot. J. Linn. Soc.* **6**, 77–96 (1862).
3. Pm, G. On the origins of observations of heterostyly in *Primula*. *New Phytol.* **208** (2015).
4. Hu, C.-M. in *Flora Reipublicae Popularis Sinicae* Vol. 59 (eds. Chen, F.-H. & H. C.-M.) *Primula* (Science Press, 1990).
5. Zhang, C.-Y. *et al*. Characterization of the whole chloroplast genome of an endangered species *Primula kwangtungensis* (Primulaceae). *Conserv. Genet. Resour.* **9**, 87–89 (2017).
6. Jamieson, I. G. & Allendorf, F. W. How does the 50/500 rule apply to MVPs? *Trends Ecol. Evol.* **27**, 578–584 (2012).
7. IUCN Red List categories and criteria, version 3.1, second edition. https://iucn.org/resources/publication/iucn-red-list-categories-and-criteria-version-31-second-edition (2012).
8. Xu, Y., Yu, X.-L., Hu, C.-M. & Hao, G. Morphological and molecular phylogenetic data reveal a new species of *Primula* (Primulaceae) from Hunan, China. *PLOS ONE* **11**, e0161172 (2016).
9. Anderberg, A. A. & El-Ghazaly, G. Pollen morphology in *Primula* sect. *Carolinella* (Primulaceae) and its taxonomic implications. *Nord. J. Bot.* **20**, 5–14 (2000).
10. Potente, G. *et al*. Comparative genomics elucidates the origin of a supergene controlling floral heteromorphism. *Mol. Biol. Evol.* **39**, msac035 (2022).
11. Darwin Tree of Life Project Consortium. Sequence locally, think globally: the Darwin Tree of Life Project. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2115642118 (2022).
12. Feng, X. *et al*. Genomic insights into molecular adaptation to intertidal environments in the mangrove *Aegiceras corniculatum*. *New Phytol.* **231**, 2346–2358 (2021).
13. Potente, G. *et al*. The *Primula edelbergii S* -locus is an example of a jumping supergene. *Mol. Ecol. Resour.* **24**, e13988 (2024).
14. Mora-Carrera, E. *et al*. Genomic Patterns of Loss of Distyly and Polyploidization in Primroses. *Mol. Biol. Evol.* **42**, msaf162 (2025).
15. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34**, 518–524 (2016).
16. Li, J., Wang, S., Yu, J., Wang, L. & Zhou, S. A Modified CTAB Protocol for Plant DNA Extraction. *Chin. Bull. Bot.* **48**, 72 (2013).
17. Huang, J. *et al*. A reference human genome dataset of the BGISEQ-500 sequencer. *GigaScience* **6**, gix024 (2017).
18. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
19. Brown, J., Pirrung, M. & McCue, L. A. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* **33**, 3137–3139 (2017).
20. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
21. Belton, J.-M. *et al*. Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
22. Adan, A., Alizada, G., Kiraz, Y., Baran, Y. & Nalbant, A. Flow cytometry: basic principles and applications. *Crit. Rev. Biotechnol.* **37**, 163–176 (2017).
23. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
24. Su, X. *et al*. A high-continuity and annotated tomato reference genome. *BMC Genomics* **22**, 898 (2021).
25. Chen, Y. *et al*. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience* **7**, gix120 (2018).
26. Hesse, U. K-Mer-Based Genome Size Estimation in Theory and Practice. *Methods Mol. Biol.* **2672**, 79–113 (2023).
27. 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).

28. Hu, J. *et al*. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol.* **25**, 107 (2024).
29. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
30. Guiglielmoni, N., Houtain, A., Derzelle, A., Van Doninck, K. & Flot, J.-F. Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms. *BMC Bioinformatics* **22**, 303 (2021).
31. Joshi, N. A. & Fass, J. N. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (2011).
32. Langdon, W. B. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Min.* **8**, 1 (2015).
33. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
34. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
35. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* **3**, 99–101 (2016).
36. Akdemir, K. C. & Chin, L. HiCPlotter integrates genomic data with interaction matrices. *Genome Biol.* **16**, 198 (2015).
37. Krzywinski, M. *et al*. Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
38. Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e49 (2012).
39. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
40. Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinforma.* **5**, 4.10.1–4.10.14 (2004).
41. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9451–9457 (2020).
42. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
43. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
44. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
45. Cossu, R. M. *et al*. LTR Retrotransposons Show Low Levels of Unequal Recombination and High Rates of Intraelement Gene Conversion in Large Plant Genomes. *Genome Biol. Evol.* **9**, 3449–3462 (2017).
46. Haas, B. G. *et al*. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
47. Haas, B. J. *et al*. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
48. Gabriel, L. *et al*. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777 (2024).
49. Zhou, G.-L. *et al*. Chromosome-scale genome assembly of *Rhododendron molle* provides insights into its evolution and terpenoid biosynthesis. *BMC Plant Biol.* **22**, 342 (2022).
50. Yu, J., Hulse-Kemp, A. M., Babiker, E. & Staton, M. High-quality reference genome and annotation aids understanding of berry development for evergreen blueberry (*Vaccinium darrowii*). *Hortic. Res.* **8**, 228 (2021).
51. Tang, W. *et al*. Chromosome-scale genome assembly of kiwifruit *Actinidia eriantha* with single-molecule sequencing and chromatin interaction mapping. *GigaScience* **8**, giz027 (2019).
52. Zhang, R.-G. *et al*. TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic. Res.* **9**, uhac017 (2022).
53. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
54. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
55. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
56. Huerta-Cepas, J. *et al*. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
57. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).
58. Tatusov, R. L. *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
59. Karlin, S. & Ladunga, I. Comparisons of eukaryotic genomic sequences. *Proc. Natl. Acad. Sci.* **91**, 12832–12836 (1994).
60. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585 (2017).
61. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP537605 (2025).
62. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_054094925.1 (2025).
63. Liu, T.-J. The genome annotation file and chromosome-level genome assembly for *Primula kwangtungensis*. *Figshare* https://doi.org/10.6084/m9.figshare.29479610 (2025).
64. 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).
65. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).
66. Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol.* **176**(2), 1410–1422 (2018).
67. 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 (2020).

Acknowledgements

We thank Xing Guo and Xuanming Guang from BGI Research (Shenzhen) for their invaluable support in the bioinformatics analysis. This work was supported by the Guangdong Flagship Project of Basic and Applied Basic Research (2023B0303050001), National Natural Science Foundation of China (32370223), Guangdong Provincial Key R&D Programme (2022B1111040003), Guangdong Forestry Bureau and Guizhou Provincial Basic Research Program (Qiankehe Fundamentals-MS [2025] 327).

Author contributions

Y.X., X.W., G.H. and R.J.W. conceived and designed the research. Y.X., X.W. and S.C. collected the samples. S.C. completed the chromosome ploidy estimation. T.J.L. completed the flow cytometry. Y.L.X., S.C. and S.S.C. performed DNA and RNA extraction, library construction, and genome sequencing. T.J.L., Y.L.X., Y.X. and S.S.C. conducted the bioinformatics analysis and visualized the results. Y.L.X. and T.J.L. drafted the manuscript. Y.X., T.J.L., R.J.W., G.H., M.I. and X.W. revised the manuscript. R.J.W., G.H. and M.I. supervised the study. All authors read and approved the final version of the 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/s41597-025-06529-w.

Correspondence and requests for materials should be addressed to G.H., R.W. or Y.X.