scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A haplotype-resolved gap-free chromosome assembly of the threatened plant *Paradombeya sinensis*

Lidan Tao[1,2,3], Lingyun Tang[4], Yuying Chen[1,2,3] & Weibang Sun[1,2,4] ✉

Paradombeya sinensis **is an ecologically and phylogenetically valuable tree species endemic to southwestern China, yet its genomic resources remain underdeveloped. Here, we present a haplotype-resolved, chromosome-scale genome assembly to address this gap. The assembly spans about 890 Mb with high continuity (contig N50 of 44 Mb) and completeness (99.93% of sequences anchored to 20 pseudo-chromosomes). Gene annotation identified 68,639 genes. This high-quality genome assembly will serve as a fundamental genomic resource to facilitate comprehensive conservation strategies, potential utilization, and definitive phylogenetic placement of this species.**

Background & Summary

The genus *Paradombeya* (Malvaceae) comprises two threatened species: *P. sinensis*, endemic to southwest China, and *P. burmanica*, found in Thailand and Myanmar[1]. Both are small trees or shrubs with narrow distributions. *P. sinensis* inhabits dry-hot valleys (200–1500 m elevation) in fragmented populations along the upper Yangtze River in Yunnan and Sichuan provinces[2]. These ecologically fragile regions are characterized by arid climates, severe soil erosion, and degraded vegetation due to steep terrain, foehn winds, and prolonged overgrazing and cultivation[2,3]. As a result, they are highly vulnerable to mudflows, wildfires, and exhibit slow ecological recovery[3]. As one of the few native woody plants in these savanna-like habitats, *P. sinensis* displays exceptional drought tolerance[4]. However, intensive human disturbance and limited population size have led to its classification as Endangered and a Grade II nationally protected wild plant in China[5,6]. Successful *ex situ* cultivation at Kunming Botanical Garden confirms its adaptability, rapid growth, and ease of propagation, supporting its use in restoring degraded upper Yangtze ecosystems. The recent approval of an ecological restoration demonstration project further underscores its applied conservation value. Additionally, *P. sinensis* exhibits horticultural potential, with its slender, gracefully drooping branches and elongated white inflorescences (Fig. 1a) making it an attractive ornamental candidate, already utilized in hedges at Kunming Botanical Garden.

Recent phylogenetic research[7] has proposed merging *Paradombeya* into the genus *Corchoropsis*, which originally contained only *C. crenata*. However, the two genera remain morphologically distinct, and taxonomic evidence supporting this merger is still considered insufficient. As a result, the reclassification has not been widely accepted by the research community. Resolving this phylogenetic controversy is a key motivation for the present study. A high-quality genome assembly of *P. sinensis* is crucial to provide the genomic evidence needed to clarify its taxonomic placement and address this long-standing debate. Pending stronger evidence, we retain the established name *Paradombeya sinensis* rather than adopting *Corchoropsis sinensis* and emphasize that the genome presented here will serve as a vital resource for definitively addressing these taxonomic issues. A high-quality *P. sinensis* reference genome is pivotal for both conservation and utilization, enabling the delineation of management units, assessment of *ex-situ* conservation, analysis of deleterious mutations, evaluation of future adaptability, exploration of functional genes, and the development of improved varieties.

[1]Yunnan Key Laboratory for Integrative Conservation of Plant Species with Extremely Small Populations, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650201, China. [2]CAS Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650201, China. [3]University of Chinese Academy of Sciences, Beijing, 101408, China. [4]Kunming Botanic Garden, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650201, China. ✉e-mail: wbsun@mail.kib.ac.cn

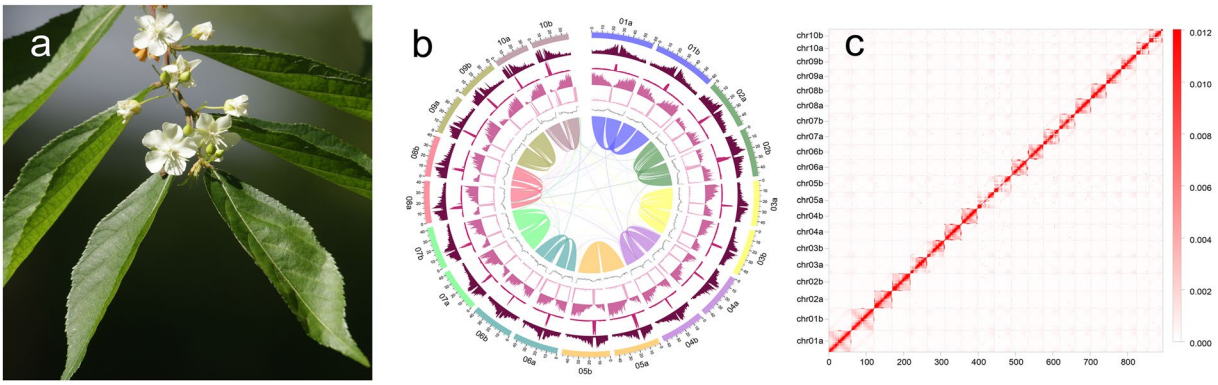

Fig. 1 (**a**) Flowers and leaves of *Paradombeya sinensis*. (**b**) The genome assembly of *P. sinensis* (window size: 500 kb). From outer to inner: chromosome coordinates, Class I TE density, Class II TE density, coding gene density, tandem repeat proportion, GC content, collinear blocks (minimum 100 kb). (**c**) Hi-C interactive heatmap (bin size = 100 kb).

Using integrated data from PacBio HiFi (38×), Oxford Nanopore (74×), Illumina (135×), Hi-C, and RNA-seq, we produced a high-quality genome assembly for *P. sinensis*. The final assembly of 890.11 Mb comprises two complete haplotypes and was successfully anchored into 20 pseudochromosomes (Supplementary Table S1). We further determined the chloroplast and mitochondrial genomes to be 426,438 bp and 160,291 bp, respectively. Repeat sequences accounted for 54.44% of the genome, predominantly LTR retrotransposons. A total of 68,639 genes were annotated, including 61,439 protein-coding genes. This gap-free, chromosomal-level reference genome provides a crucial foundation for elucidating the phylogenetics of *Paradombeya–Corchoropsis* and advancing conservation efforts for this species.

Methods

Sampling. Fresh leaves for DNA extraction, along with young shoots, leaves, axillary buds, and flowers for RNA extraction, were collected from a single *P. sinensis* individual located in the Kunming Botanical Garden (25°08′32.32″N, 102°44′30.43″E). This plant was transplanted from a wild population in Yunhe Village, Yunnan (27°58′12.39″N, 103°53′31.47″E) in September 2018. All samples for RNA analysis were immediately flash-frozen in liquid nitrogen and stored at −80 °C. Nucleic acid extraction and sequencing were performed by Wuhan Benagen Tech Co., Ltd.

Short-read (Illumina) sequencing. Total DNA was extracted from young *P. sinensis* leaves using modified CTAB method[8]. DNA concentration was quantified using both NanoDrop spectrophotometer and Qubit 3.0 fluorometer, while purity and integrity were confirmed by 1% agarose gel electrophoresis. A short-read library with an insert size of 200–400 bp was constructed from 1 μg of genomic DNA following the manufacturer's instructions (BGI). Paired-end sequencing (PE150) was subsequently performed on a DNBSEQ-T7 platform, generating approximately 59.74 Gb of raw data (~398 million reads), which provided ~150 × coverage (Supplementary Table S2).

Long-read (PacBio HiFi) sequencing. For HiFi sequencing, genomic DNA was purified with a DNeasy Plant Mini Kit, and its integrity was verified using a Femto Pulse instrument. Subsequently, 8 μg of DNA was sheared using a Megaruptor 3 system, and fragments shorter than 5 kb were removed with AMPure PB magnetic beads. A SMRTbell library was prepared using the SMRTbell Prep Kit 3.0, and fragments of ~30 kb were size-selected via the BluePippin system. Sequencing was conducted on a PacBio Revio system in Circular Consensus Sequencing mode. Raw data were processed into high-fidelity reads using the CCS workflow (v7.0.0)[9] with a minimum read quality of 0.9, yielding ~17.49 Gb of HiFi data (~38 × coverage) (Supplementary Table S3). The HiFi reads had an average length and N50 of 19 kb.

Long-read (Oxford Nanopore) sequencing. A Nanopore sequencing library was prepared using the SQK-LSK109 kit and sequenced on a PromethION platform. This generated ~33.30 Gb of whole-genome sequencing data (~74 × coverage), with an average read length of 14 kb and an N50 of 72 kb (Supplementary Table S4).

Hi-C library construction and sequencing. Hi-C libraries were prepared from fresh leaf tissue. Briefly, tissues were cross-linked with formaldehyde and digested with DpnII restriction enzyme. The resulting ends were repaired using biotin-labeled nucleotides, and the cross-linked fragments were ligated. After reverse cross-linking and DNA purification, the DNA was sheared into 200–400 bp fragments using a fragmentase enzyme. These fragments subsequently underwent 5′-end phosphorylation and 3′-end dA-tailing, followed by adapter ligation. Biotin-labeled fragments were enriched with streptavidin-coated magnetic beads and amplified via PCR. Library quality was assessed by measuring DNA concentration (Qubit 3.0) and insert size distribution (Agilent 2100 Bioanalyzer). The effective library concentration was accurately quantified by qPCR. The final qualified libraries were sequenced on the DNBSEQ-T7 platform with paired-end 150 bp (PE150) chemistry, generating 69.54 Gb of high-quality data (approximately 463.60 million reads; Supplementary Table S5).

Parameter	Genome
Genome size	890,692,568 bp
GC content	33.77%
Contig number	22
Contig N10	60,072,766 bp
Contig N50	43,928,138 bp
Contig N90	39,511,115 bp
Gap number	0
Chromosome number	20
Chromosome length	890,105,839 bp (99.93%)
Mitochondria length	426,438 bp (0.05%)
Chloroplast length	160,291 bp (0.02%)

Table 1. Summary of *Paradombeya sinensis* genome assembly.

Transcriptome sequencing. Total RNA was extracted from young leaves, stems, buds, and flowers of a single mature plant. Poly(A) + mRNA was enriched using oligo(dT) beads and fragmented. Strand-specific cDNA libraries were constructed through first-strand and second-strand synthesis (incorporating dUTP), followed by end repair, A-tailing, adapter ligation, and size selection with magnetic beads. The PCR-amplified libraries were quality-controlled, circularized, and proportionally pooled for sequencing. Sequencing was performed on the DNBSEQ-T7 platform, yielding 26.68 Gb of raw data (approximately 177.85 million reads; Supplementary Table S6).

Genome size estimation. The genome size of *P. sinensis* was estimated using both flow cytometry (FCM) and k-mer frequency analysis[10]. For FCM, nuclear DNA content was measured on a BD FACScalibur flow cytometer (BD Biosciences, USA) with tomato (*Solanum lycopersicum*) as an internal reference, yielding an estimated genome size of 0.53 Gb (1 C). For k-mer analysis, 19-mer frequencies were counted using Jellyfish v2.2.10[11]. Genome size, heterozygosity, and repeat content were then estimated with GenomeScope v2.0[12]. This analysis produced a final estimated genome size of 421.24 Mb (1 C), with a heterozygosity rate of 0.91% and a repeat content of 42.76% (Supplementary Table S7).

Chromosome-level genome assembly. We generated a chromosome-level genome assembly for *P. sinensis* using a multi-platform sequencing approach. PacBio HiFi, Oxford Nanopore Technologies (ONT) whole-genome shotgun, and Hi-C reads were initially assembled into contigs with Hifiasm (v0.19.8-r602)[13]. The most representative haplotype assembly was selected for subsequent scaffolding. Hi-C reads were aligned to this draft assembly using BWA (v0.7.17-r1188)[14], and chromosome-scale scaffolding was performed with 3D-DNA (v180922)[15] (parameters:–early-exit -m haploid -r 0). The scaffolded assembly was manually inspected and adjusted using Juicebox (v1.11.08)[16] (pre -n -q 0 or 1) to refine chromosome boundaries, correct misassemblies, and resolve switch errors. This process yielded a final assembly comprising 20 gap-free chromosomes (2n = 20) and some unanchored sequences.

Although most chromosomal telomeres contained the canonical (TTTAGGG)n repeat[17], some were short or absent, indicating potential incomplete assembly. To address this, we mapped HiFi reads back to the chromosomes. Reads from telomeric regions were extracted, de novo assembled into contigs using Hifiasm and then mapped back to the chromosome ends. The chromosomes were subsequently extended using these telomere-enriched contigs to maximize sequence completeness. Separately, the chloroplast and mitochondrial genomes were assembled from HiFi reads using OATK[18].

The entire nuclear assembly was then polished with Nextpolish2 (v0.1.0)[19] using both HiFi and Illumina short reads. Potential redundancies, such as rDNA fragments and haplotigs, were removed using Redundans (v0.13c)[20] (parameters: -identity 0.98 -overlap 0.8) combined with manual curation. Ultimately, we produced a high-quality genome assembly where approximately 99.93% of the sequence was anchored into 20 pseudochromosomes (designated chr01-10ab), with the mitochondrial and chloroplast genomes accounting for 0.05% and 0.02%, respectively (Table 1; Fig. 1b,c; Supplementary Table S1).

Identification of repetitive elements. Repetitive elements were identified through a combination of de novo and homology-based approaches. First, a custom transposable element (TE) library was constructed de novo using EDTA (v1.9.9)[21] with sensitive parameters (–sensitive 1–anno 1). This library was then used alongside RepeatMasker (v4.0.7)[22] to annotate repetitive regions across the genome (-no_is -xsmall). In total, we identified 1,196,272 repetitive elements, spanning 484.91 Mb and accounting for 54.44% of the genome. The most abundant class was long terminal repeat (LTR) retrotransposons, with 253,957 elements covering 240.32 Mb (26.90% of the genome), as detailed in Supplementary Table S8.

Gene identification and functional annotation. We performed gene identification and functional annotation through an integrated evidence-based approach. First, we compiled homologous protein evidence by merging 309,313 non-redundant sequences from 11 plant species, including *Bombax ceiba*[23], *Theobroma cacao*[24], *Durio zibethinus*[25], *Microcos paniculata*[26], *Firmiana kwangsiensis*[27], *Ochroma pyramidale*[28], *Gossypium raimondii*[29], *Dipterocarpus gracilis*[30], *Aquilaria sinensis*[31], *Arabidopsis thaliana*[32], and *Vitis vinifera*[33].

For transcriptome evidence, we processed RNA-seq reads via two strategies: (1) de novo assembly with Trinity v2.0.6[34], and (2) mapping to the genome using HISAT2 v2.1.0[35] followed by assembly with StringTie

Feature	Total Number	Coding Genes Number
gene	68,639	61,439
transcript	94,877	87,677
CDS	87,677	87,677
exon	548,268	540,993
intron	453,391	453,316

Table 2. Summary of *Paradombeya sinensis* genome annotations.

Program	Database	Number	Percent (%)
eggNOG-mapper	GO	29,883	48.64%
	KEGG_KO	27,859	45.34%
	EC	12,315	20.04%
	KEGG_Pathway	17,552	28.57%
	eggNOG	54,363	88.48%
	COG	58,881	95.84%
DIAMOND	Swiss_Prot	45,758	74.48%
	TrEMBL	60,035	97.71%
	NR	59,837	97.39%
	TAIR10	54,982	89.49%
InterProScan	CDD	21,319	34.70%
	Pfam	49,780	81.02%
	SUPERFAMILY	39,230	63.85%
	Interpro	51,904	84.48%
	Coils	9,679	15.75%
	Gene3D	41,627	67.75%
	Phobius	21,929	35.69%
	PRINTS	8,836	14.38%
	TIGRFAM	6,207	10.10%
	SMART	18,966	30.87%

Table 3. Statistics of functional annotation results of *Paradombeya sinensis* genome.

v2.1.5[36] (Supplementary Table S9). The PASA pipeline v2.4.1[37] then annotated gene structures using these assembled transcripts, and we identified full-length genes by comparing them with reference proteins. To improve ab initio prediction, we trained both AUGUSTUS v3.4.0[38] and SNAP[39] using the set of full-length genes, subjecting AUGUSTUS to five rounds of optimization.

We executed structural gene annotations with MAKER2 v2.31.9[40], which integrated evidence from three sources: ab initio predictions (AUGUSTUS, SNAP), transcript assemblies (aligned via BLASTN), and homologous proteins (aligned via BLASTX and refined with Exonerate v2.2.0[41]). RepeatMasker-masked regions were excluded, and alignment-derived hints files guided the AUGUSTUS and SNAP predictors.

The annotations from MAKER and PASA were subsequently integrated into a consensus gene set using EVidenceModeler (EVM) v1.1.1[42]. Prior to EVM, we masked transposable element-related coding regions identified by TEsorter v1.4.1[43]. The resulting EVM gene set was then refined with PASA v2.4.1 to add untranslated regions (UTRs) and alternative splicing isoforms, and to remove genes containing abnormal coding frames or short sequences (<50 amino acids). We annotated non-coding RNAs using Barrnap v0.9 (rRNAs), tRNAscan-SE v1.3.1[44] (tRNAs), and RfamScan v14.2[45] (other ncRNAs).

For functional annotation of protein-coding genes, we employed three complementary approaches: (1) eggNOG-mapper v2.0.0[46] for assigning Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms through alignment to the eggNOG v5.0 database (targeting Viridiplantae, using DIAMOND); (2) DIAMOND v0.9.24[47] for sequence matching (e-value < 1e-5, max-target-seqs = 5, identity >30%) against the Swiss-Prot, TrEMBL, NR, and Arabidopsis databases to identify best hits; and (3) InterProScan v5.27-66.0[48] for predicting protein domains and motifs using the PRINTS, Pfam, SMART, PANTHER, and CDD sub-databases. Ultimately, 60,635 genes (98.69% of all predicted protein-coding genes) were functionally annotated in at least one database (see Tables 2, 3 and Supplementary Table S11).

Data Records

The datasets supporting this study have been deposited in the following repositories. Sequencing data, including BGI short-reads, PacBio HiFi long-reads, Hi-C, WGS ONT, Iso-Seq, and RNA-Seq data, are available under BioProject accession number PRJNA1266756[49] in the Genome Sequence Archive (GSA). The final chromosome assembly is available in the GenBank Genome Warehouse (GWH) under accession number GCA_050660705.1[50]. Genome annotation data are hosted on Figshare[51].

Data set	Reads mapped	Bases mapped	>=1×	>=5×	>=10×	>=20×
HiFi	99.93%	99.93%	100.00%	99.92%	98.47%	43.78%
RNA-Seq	98.22%	98.22%	21.57%	15.46%	12.86%	10.49%
Short reads	99.85%	98.85%	99.99%	99.93%	99.76%	99.14%

Table 4. Mapping ratio and coverage percentage of different data sets.

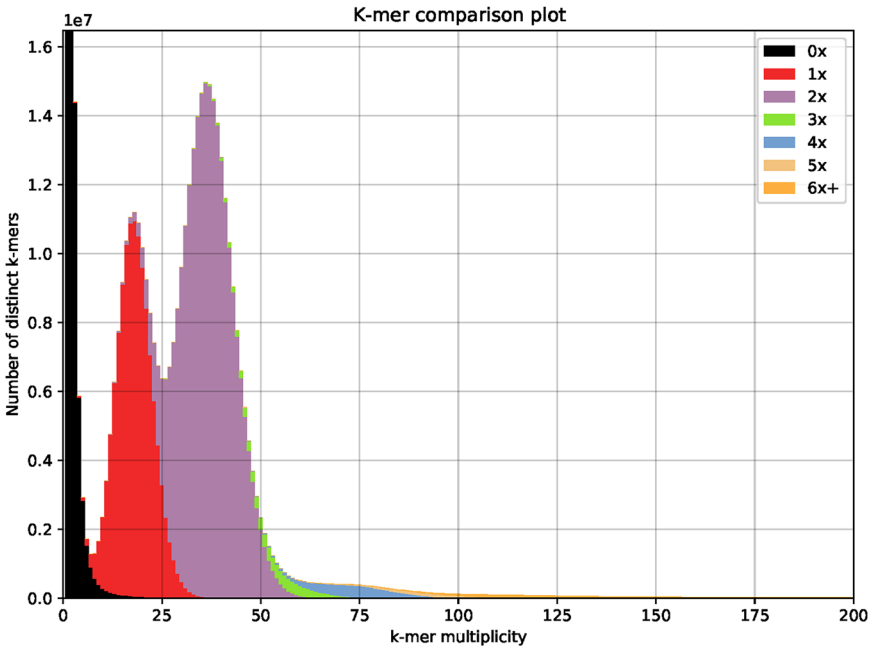

Fig. 2 Graphical representation of copy number variation spectra for genome assembly validation against HiFi sequencing data.

Technical Validation

Genome assembly quality assessment. The final gap-free assembly spans approximately 890 Mb and consists of two complete haplotypes, which is consistent with K-mer analysis estimates (Supplementary Table S8). The assembly exhibited high continuity, with a contig N50 of 44 Mb. Short reads and third-generation reads were mapped to the genome using BWA-MEM (v0.7.17-r1188) and Minimap2 (v2.24)[52], respectively. After filtering out non-primary alignments, the mapping ratios and coverage (Table 4) confirmed high sequence coverage. BUSCO (v5.3.2)[53] analysis revealed 99.0% complete core genes, with only 0.4% missing, indicating high gene completeness. Coverage depth distribution analysis against HiFi data confirmed the absence of assembly redundancy or homologous collapse (Fig. 2).

Assessment of the sequencing data revealed no significant GC bias in either short or long reads (Supplementary Fig. S1). Furthermore, Hi-C data mapped to the assembly using BWA demonstrated effective chromosome clustering without apparent misassemblies (Supplementary Fig. S1). We identified a putative centromere tandem repeat motif from HiFi reads using Centromics (https://github.com/zhangrengang/Centromics), and large tandem repeat regions, corresponding to putative centromeres, were detected on all chromosomes (Fig. 3). Additionally, 18S-5.8S-28S rDNA arrays were identified on Chr5a, Chr5b, Chr7a, Chr10a, and Chr10b, while 5S rDNA arrays were found on Chr5a, Chr5b, Chr8a, Chr8b, Chr10a, and Chr10b (Fig. 3). To further evaluate the assembly quality, we calculated the LTR Assembly Index (LAI) using LTR_Finder (v1.0.2)[54]. The LAI values for most chromosomes ranged from 15 to 23.27 (Supplementary Table S1), with lower values predominantly concentrated at chromosome ends and centromeric regions (Supplementary Fig. S2). The telomere integrity of the assembled chromosomes was assessed using quarTeT (v1.2.5)[55]. The results showed that all chromosomes contained telomeres in both ends, with the number of repeats ranging from 106 to 589 (Fig. 3; Supplementary Table S10). Collectively, these results support the high quality of this gap-free, telomere-to-telomere assembly.

Evaluation of the gene annotation. The annotated proteins were assessed using BUSCO analysis (embryophyta_odb10 lineage). Of the 1,614 BUSCO groups, 99.1% were complete (comprising 2.9% single-copy and 96.2% duplicated), 0.2% were fragmented, and 0.7% were missing (Table 5). These results confirm the high quality of the gene annotation.

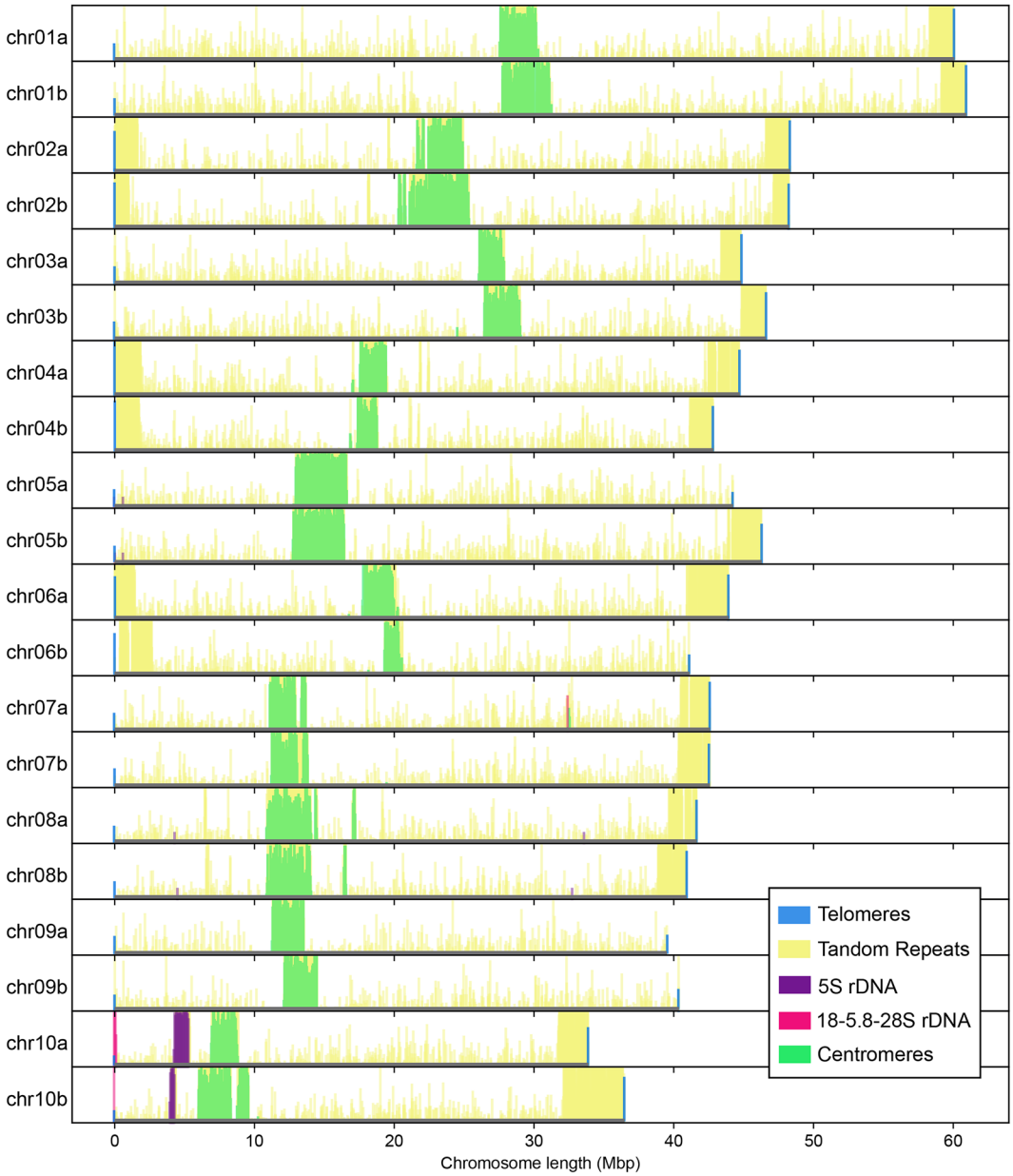

Fig. 3 The distribution of repeated elements on the chromosomes: telomeric ACTAT, tandem repeat, 5S rDNA, 18-5.8-28S rDNA, and putative centromeres. The vertical axis represents the count of repeated elements within 20k intervals.

Type	Number
Complete BUSCOs (C)	1,598 (99.1%)
Complete and single-copy BUSCOs (S)	46 (2.9%)
Complete and duplicated BUSCOs (D)	1,552 (96.2%)
Fragmented BUSCOs (F)	4 (0.2%)
Missing BUSCOs (M)	12 (0.7%)
Total BUSCO groups searched	1,614

Table 5. BUSCO assessment result.

Data availability

All data from this study have been deposited in public repositories: the sequencing data (BGI, PacBio HiFi, Hi-C, WGS ONT, Iso-Seq, RNA-Seq) are available in the Genome Sequence Archive under BioProject PRJNA1266756[49];

the final chromosome assembly is in the GenBank GWH under accession GCA_050660705.1[50]; and the annotation is on Figshare (https://doi.org/10.6084/m9.figshare.29206643)[51].

Code availability

All analyses were performed using publicly available software and tools, following the standard protocols provided by the respective developers. Detailed versions, parameters, and computational pipelines are described in the Methods section. Any unspecified parameters were left at their default settings. No custom code was developed for this study.

Received: 3 June 2025; Accepted: 15 December 2025;
Published online: 20 December 2025

References

1. Tang, Y., Gilber, M. G. & Dorr, L. J. Sterculiaceae. in *Flora of China* (eds. Wu, Z., Raven, P. H. & Hong, D.) **Vol. 12** (Science Press, Beijing, and Missouri Botanical Garden Press, St. Louis, 2007).
2. Zhang, J. P., Yang, Z., Wang, B. J. & Zhang, X. B. Climate change and causes in the Yuanmou dry-hot valley of Yunnan, China. *J. Arid. Environ.* **51**, 153–162, https://doi.org/10.1006/jare.2001.0851 (2002).
3. Peng, S. L., Chen, A. Q., Fang, H. D., Wu, J. L. & Liu, G. C. Effects of vegetation restoration types on soil quality in Yuanmou dry-hot valley, China. *Soil Sci. Plant Nutr.* **59**, 347–360, https://doi.org/10.1080/00380768.2013.785918 (2013).
4. Zhu, H., Tan, Y. H., Yan, L. C. & Liu, F. Y. Flora of the Savanna-like vegetation in hot dry valleys, southwestern China with implications to their origin and evolution. *Bot. Rev.* **86**, 281–297, https://doi.org/10.1007/s12229-020-09227-x (2020).
5. Yang, J. & Sun, W. B. A new programme for conservation of Plant Species with Extremely Small Populations in south-west China. *Oryx* **51**, 396–397, https://doi.org/10.1017/S0030605317000710 (2017).
6. Sun, W. B. List of Yunan protected plant species with extremely small populations. (Yunnan Science and Technology Press, 2021) (2021).
7. Dorr, L. J. & Wurdack, K. J. Indo-Asian *Eriolaena* expanded to include two Malagasy genera, and other generic realignments based on molecular phylogenetics of Dombeyoideae (Malvaceae). *Taxon* **70**, 99–126, https://doi.org/10.1002/tax.12370 (2021).
8. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Lett.* **19**, 11–15 (1987).
9. Wenger, A. M. *et al*. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162, https://doi.org/10.1038/s41587-019-0217-9 (2019).
10. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. Preprint at https://doi.org/10.48550/arXiv.1308.2012 (2020).
11. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **7**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
12. 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).
13. Cheng, H. *et al*. 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).
14. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv*, 1303.3997 https://doi.org/10.48550/arXiv.1303.3997 (2013).
15. Dudchenko, O. *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95, https://doi.org/10.1126/science.aal3327 (2017).
16. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
17. Gao, D. *et al*. TAR30, a homolog of the canonical plant TTTAGGG telomeric repeat, is enriched in the proximal chromosome regions of peanut (*Arachis hypogaea* L.). *Chromosome Res.* **30**, 77–90, https://doi.org/10.1007/s10577-022-09684-7 (2022).
18. Zhou, C. X. *et al*. Oatk: a de novo assembly tool for complex plant organelle genomes. *bioRxiv*, https://doi.org/10.1101/2024.10.23.619857 (2024).
19. Hu, J. *et al*. NextPolish2: a repeat-aware polishing tool for genomes assembled using HiFi long reads. *Genom. Proteom. & Bioinform.* **22**, qzad9, https://doi.org/10.1093/gpbjnl/qzad009 (2024).
20. Pryszcz, L. P. & Gabaldón, T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* **44**, e113, https://doi.org/10.1093/nar/gkw294 (2016).
21. Ou, S. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275, https://doi.org/10.1186/s13059-019-1905-y (2019).
22. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **25**, 4.10.11–4.10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
23. Gao, Y. *et al*. De novo genome assembly of the red silk cotton tree (*Bombax ceiba*). *GigaScience* **7**, https://doi.org/10.1093/gigascience/giy051 (2018).
24. Argout, X. *et al*. The genome of *Theobroma cacao*. *Nat. Genet.* **43**, 101–108, https://doi.org/10.1038/ng.736 (2011).
25. Teh, B. T. *et al*. The draft genome of tropical fruit durian (*Durio zibethinus*). *Nat. Genet.* **49**, 1633, https://doi.org/10.1038/ng.3972 (2017).
26. Liu, D. T., Tian, X. L., Shao, S. C., Ma, Y. P. & Zhang, R. G. Haplotype-resolved chromosomal-level genome assembly of Buzhaye (*Microcos paniculata*). *Sci. Data* **10**, https://doi.org/10.1038/s41597-023-02821-9 (2023).
27. Wang, B. Q., Zhang, R. G., Sun, W. B. & Yang, J. A nearly telomere-to-telomere diploid genome assembly of *Firmiana kwangsiensis*, a threatened species in China. *Sci. Data* **11**, https://doi.org/10.1038/s41597-024-04250-8 (2024).
28. Sahu, S. K. *et al*. Chromosome-scale genomes of commercial timber trees (*Ochroma pyramidale*, *Mesua ferrea*, and *Tectona grandis*). *Sci. Data* **10**, https://doi.org/10.1038/s41597-023-02420-8 (2023).
29. Wang, M. J. *et al*. Genomic innovation and regulatory rewiring during evolution of the cotton genus *Gossypium*. *Nat. Genet.* **54**, https://doi.org/10.1038/s41588-022-01237-2 (2022).
30. Tian, Z. Z. *et al*. Thirteen Dipterocarpoideae genomes provide insights into their evolution and borneol biosynthesis. *Plant Commun.* **3**, https://doi.org/10.1016/j.xplc.2022.100464 (2022).
31. Ding, X. P. *et al*. Genome sequence of the agarwood tree *Aquilaria sinensis* (Lour.) Spreng: the first chromosome-level draft genome in the Thymelaeceae family. *GigaScience* **9**, https://doi.org/10.1093/gigascience/giaa013 (2020).
32. Cheng, C. Y. *et al*. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant. J.* **89**, 789–804, https://doi.org/10.1111/tpj.13415 (2017).
33. Shi, X. Y. *et al*. The complete reference genome for grapevine (*Vitis vinifera* L.) genetics and breeding. *Hortic. Res.-England* **10**, https://doi.org/10.1093/hr/uhad061 (2023).
34. 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).
35. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360, https://doi.org/10.1038/NMETH.3317 (2015).

36. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
37. Haas, B. J. *et al*. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666, https://doi.org/10.1093/nar/gkg770 (2003).
38. 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).
39. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59, https://doi.org/10.1186/1471-2105-5-59 (2004).
40. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491, https://doi.org/10.1186/1471-2105-12-491 (2011).
41. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31, https://doi.org/10.1186/1471-2105-6-31 (2005).
42. 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).
43. Zhang, R. G. *et al*. TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic. Res.* **9**, uhac017, https://doi.org/10.1093/hr/uhac017 (2022).
44. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964, https://doi.org/10.1093/nar/25.5.955 (1997).
45. Nawrocki, E. P. *et al*. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* **43**, D130–D137, https://doi.org/10.1093/nar/gku1063 (2014).
46. Huerta-Cepas, J. *et al*. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122, https://doi.org/10.1093/molbev/msx148 (2017).
47. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60, https://doi.org/10.1038/nmeth.3176 (2015).
48. Jones, P. *et al*. InterProScan5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
49. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/bioproject:PRJNA1266756 (2025).
50. *NCBI Assembly* https://identifiers.org/insdc.gca:GCA_050660705.1 (2025).
51. Tao, L. D., Tang, L. Y., Chen, Y. Y. & Sun, W. B. A haplotype-resolved gap-free chromosome assembly of the threatened plant *Paradombeya sinensis*. *Figshare* https://doi.org/10.6084/m9.figshare.29206643 (2025).
52. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
53. 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, https://doi.org/10.1093/bioinformatics/btv351 (2015).
54. Zhao, X. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268, https://doi.org/10.1093/nar/gkm286 (2007).
55. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic. Res.-England* **10**, https://doi.org/10.1093/hr/uhad127 (2023).

Acknowledgements

This work was supported by grants from the Key R&D Program of Yunnan Province Science and Technology Department (Grant No. 202403AC100028 to W.S.) and the Key R&D Program of China (Grant No. 2024YFF1307402 to W.S.).

Author contributions

W.S. conceived the project and designed the experiments. L.T. and L.T. prepared the samples. L.T., L.T. and Y.C. drafted the manuscript. L.T., L.T., Y.C. and W.S. revised the manuscript. All authors contributed to the research and approved the submitted version.

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/s41597-025-06470-y.

Correspondence and requests for materials should be addressed to W.S.

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.