



OPEN

DATA DESCRIPTOR

Near telomere-to-telomere (T2T) level genome assembly of the critically endangered plant *Magnolia zenii* (Magnoliaceae)

Nai-wei Li¹, Zheng-nan Zhang¹, Shuai Yuan², Dong-rui Yao¹ & Bao-cheng Wu¹✉

Magnolia zenii (Magnoliaceae) is an extremely endangered tree species endemic to eastern China, has only 115 individuals remaining in the wild. By integrating PacBio HiFi (49×), Oxford Nanopore (ONT; 101×) long-read sequencing, and Hi-C chromatin interaction data, we achieved a high-quality chromosome-level genome with scaffold N50 of 86 Mb, covering 99.97% of the estimated genome size. The genome annotation predicted 74,782 protein-coding genes and revealed a repetitive sequence content of 72.87%, dominated by long terminal repeat (LTR) retrotransposons (53.59%). The mitochondrial (915,126 bp) and chloroplast (160,079 bp) genomes were fully characterized. BUSCO analysis confirmed 99.4% genome completeness and 98.4% gene annotation completeness. This genomic resource provides a foundation for studying the genetic mechanisms underlying its endangered status and future horticultural.

Background & Summary

Magnolia zenii W. C. Cheng, a Tertiary relict species, persists as a critically endangered (CR)¹ plant with only 115 mature individuals² remaining in fragmented distribution across Baohua Mountain of Jiangsu Province, eastern China, as documented in the China Species Red List¹. This species is valued in horticulture because of its elegant pale blossoms and stately upright trunk³ (Fig. 1a). Despite its ecological importance, while previous research have mainly concentrated on chloroplast genome sequencing for phylogenetic analyses and only a limited number of microsatellite markers developed for population genetic diversity analyses^{4,5}, the absence of whole-genome investigations hinders comprehensive understanding of its reproductive limitations⁶, genetic bottlenecks, and adaptive mechanisms. Magnoliaceae is one of the oldest extant angiosperm families, yet genomic studies on this family have been relatively limited. Globally and within China, nearly 50% of Magnoliaceae species are threatened^{7,8}. Despite recent advancements, only eight species (*Liriodendron chinense* (Hemsley) Sargent, *Magnolia biondii* Pamp., *M. kwangsiensis* Figlar & Noot., *M. hypoleuca* Sieb. & Zucc., *M. officinalis* Rehd., *M. sieboldii* K. Koch, *M. sinica* (Law) Noot., *Michelia alba* DC) have been subjected to in-depth genomic analysis in the past five years^{9–16}. This limited genomic resource significantly hinders both the understanding of the endangered mechanisms of Magnoliaceae species and their evolutionary relationships.

Whole-genome sequencing would enable systematic identification of key reproductive development genes and functional variations, while elucidating the molecular basis of pollen and embryo sac abortion⁶. Furthermore, it would facilitate assessment of population genetic diversity and genomic architecture, providing critical genetic insights for conservation strategies such as artificial propagation (e.g., seedling transplantation), population restoration optimization, and stress resistance improvement. Particularly crucial given current population decline and ecological competition pressures¹⁷, genomic data integration with population dynamics and niche analysis may help to inform precision conservation measures. This integrated approach would enhance the species' long-term survival prospects and support sustainable utilization efforts.

¹Jiangsu Key Laboratory for Conservation and Utilization of Plant Resources, Institute of Botany, Jiangsu Province and Chinese Academy of Sciences (Nanjing Botanical Garden Mem. Sun Yat-Sen), Nanjing, Jiangsu, China. ²BaoHua Mountain National Forest Park Administration Committee, Zhenjiang, Jiangsu, China. ✉e-mail: wubaocheng2015@163.com

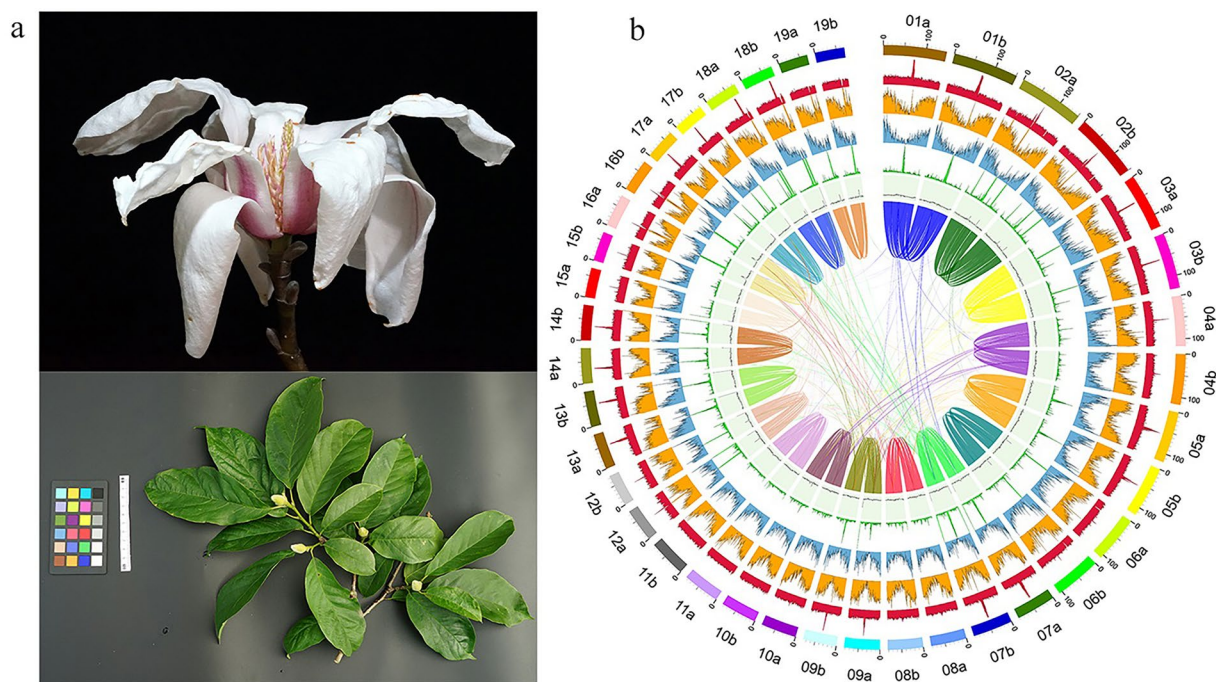


Fig. 1 The floral characteristics (a) and genomic features of *M. zenii* (b). From outer to inner layers (excluding the outermost chromosome coordinates): Class I TE density, Class II TE density, protein-coding gene density, tandem repeat proportion, GC content, and collinearity blocks (minimum length of 100 kb). Window size: 500 kb.

Here, we combined PacBio HiFi, ONT ultra-long reads, and Hi-C interaction maps to generate a haplotype-resolved chromosome-level assembly and performed robust gene annotation with ONT transcriptomic data. This approach addresses challenges in resolving complex regions and ensures high contiguity and annotation accuracy. The assembly spans 3.6 Gb with a scaffold N50 of 86 Mb and 7 gaps (Table 1), achieving near telomere-to-telomere resolution. Benchmarking Universal Single-Copy Orthologs (BUSCO¹⁸, v. 5.3.2) assessment confirmed 99.4% genome completeness and 98.4% gene annotation completeness, with 74,782 protein-coding genes identified. Repetitive elements (72.87%) were dominated by LTR retrotransposons (53.59%). The mitochondrial (915,126 bp) and chloroplast (160,079 bp) genomes were also fully characterized. The assembly provides a foundational resource for investigating its genomic basis of endangerment, identifying key genomic variants for breeding, and reconstructing evolutionary history. It also advances horticultural trait research by enabling precise genetic and genomic analysis of *M. zenii*.

Methods

Sample preparation and DNA/RNA extraction. Sampled fresh flowers and leaves from a mature *M. zenii* individual in Baohua Mountain, Jiangsu, China. Flowers and leaves were preserved in liquid nitrogen and stored at -80° . DNA was extracted from leaves using a modified CTAB methods¹⁹. Total RNA was extracted from flowers and leaves using the R6827 Plant RNA Kit (Omega Bio-Tek, Norcross, GA, USA) following the manufacturer's instructions. Subsequently, the concentration of the extracted DNA and RNA samples was determined using the NanoDrop One spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and the Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA), while their integrity was assessed via 1% agarose gel electrophoresis.

Library construction and sequencing. DNA integrity was assessed using the Femto Pulse system. Genomic DNA (8 μ g) was fragmented using the Megaruptor 3, followed by purification with AMPure PB magnetic beads. SMRT bell libraries were constructed using the Pacific Biosciences SMRT Bell Express Template Prep Kit 2.0. Libraries were size-selected on the BluePippinTM system to target 15 kb insert sizes, followed by primer annealing and binding of SMRT bell templates to polymerases using a DNA/polymerase binding kit. The sequencing was performed on the PacBio Sequel II/Ile platform at Benagen Genomics. This yielded 88 Gb ($49\times$) of HiFi data with N50 of 17 kb.

ONT library construction and sequencing. A 2.5 μ g aliquot of quality-checked genomic DNA was purified with $1\times$ magnetic beads. A 1 μ L sample was quantified via Qubit fluorometry. DNA underwent end repair and damage remediation, followed by $1\times$ magnetic bead purification. DNA was eluted with 61 μ L EB buffer, and a 1 μ L aliquot was re-quantified. Adapter ligation was performed, followed by $0.4\times$ magnetic bead purification. Elution was conducted with 25 μ L Elution Buffer (SQK-LSK110), and a final 1 μ L sample was quantified. Libraries

Statistics	<i>Magnolia zenii</i>
Total size (bp)	3,599,591,302
Total size without N (bp)	3,599,590,602
GC content (%)	40.43%
(a) Contig	
Number	47
Max. (bp)	151,657,134
Min. (bp)	160,079
Median (bp)	77,708,874
Mean (bp)	76,587,034
N10 (bp)	150,127,340
N50 (bp)	85,095,647
N90 (bp)	50,682,438
L10	3
L50	16
L90	35
(b) Scaffold	
Number	40
Max. (bp)	151,657,134
Min. (bp)	160,079
Median (bp)	82,556,826
Mean (bp)	89,989,782
N10 (bp)	150,127,340
N50 (bp)	86,565,855
N90 (bp)	71,954,542
L10	3
L50	15
L90	33
(c) Gap	
Number	7
Max. (bp)	100
Min. (bp)	100
Median (bp)	100
Mean (bp)	100

Table 1. Summary of the *M. zenii* genome assembly data.

were loaded onto R9.4 sequencing chips and sequenced on the PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK) for 48–72 hours. This generated 182 Gb (101 ×) of data with an N50 of 60 kb.

Hi-C library construction and sequencing. Input Hi-C DNA was fragmented, followed by end blunt-ing. 5' ends were phosphorylated, and 3' dA tails were added. Adapters were ligated to fragmented DNA ends. Streptavidin magnetic beads were used to capture biotin-labeled regions. Captured Hi-C DNA underwent PCR amplification, followed by quantification and quality control. Libraries were circularized and validated. Qualified libraries were sequenced on the DNBSEQ-T7 platform using paired-end sequencing protocols.

ONT transcriptome sequencing. A 500 ng aliquot of quality-checked total RNA was diluted to 9 µL with nuclease-free water. Reverse transcription primers were ligated, followed by cDNA synthesis and amplification. Libraries were purified using AMPure beads and adapter ligated. Final libraries were loaded onto R9.4 sequencing chips and sequenced on the PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK) for 48–72 hours. This produced 18 Gb of transcriptome data. This produced 18 Gb of transcriptome data.

Chromosome-level genome assembly. Quality filtering of second-generation Hi-C data was performed using the fastp²⁰ (v. 0.21.0) software, including adapter trimming, removal of low-quality regions, and discarding of short reads. PacBio HiFi reads were initially assembled using Hifiasm²¹ (v. 0.19.8-r602) to generate a draft genome. Hi-C reads were then aligned to this draft genome using Juicer²² (v. 0.6.14), followed by preliminary Hi-C-assisted chromosome scaffolding with 3D-DNA²³ (v. 180922). Subsequent manual validation and iterative refinement were performed in Juicebox²⁴ (v. 3.1.4), focusing on optimizing chromosomal partition boundaries, correcting structural misassemblies, and resolving switch errors. Each chromosome was then individually re-scaffolded using 3D-DNA to enhance contiguity, with further manual adjustments in Juicebox to remove spurious insertions, correct orientation discrepancies, and address other detectable errors. Following rigorous manual curation, the final assembly comprised a chromosome-scale framework and dispersed sequences, with gaps fixed to 100 bp. Statistical details are provided in Table S1-4.

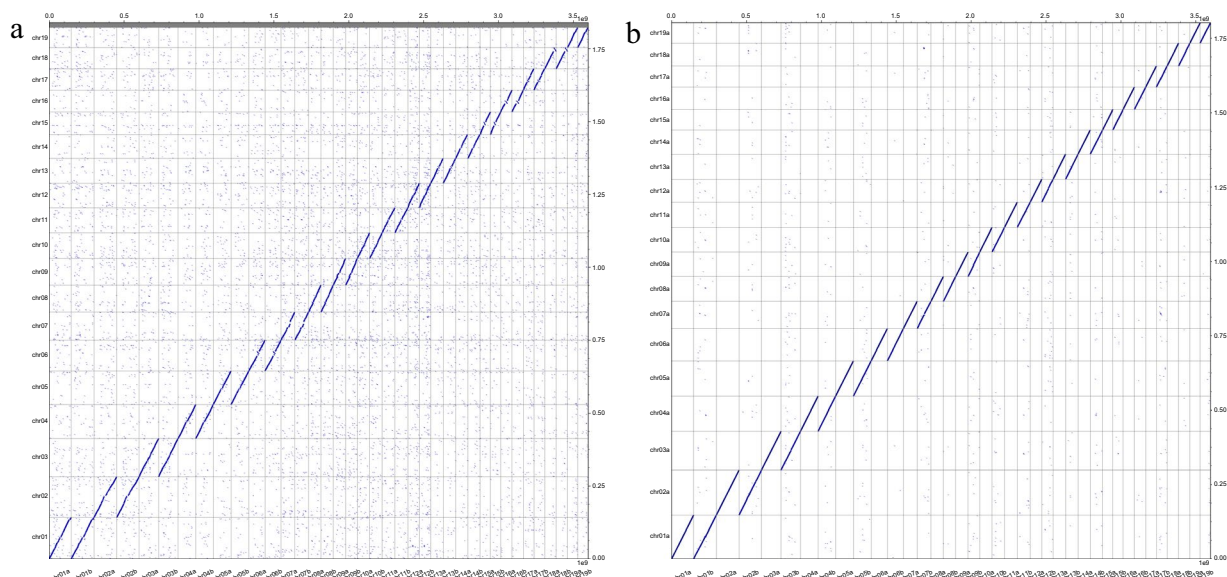


Fig. 2 Chromosome-level collinearity between the current assembly and the reference (*P. sinica*) genome and haplotype assembly. **(a)** Collinearity with the reference genome (*P. sinica*); **(b)** Collinearity with the haplotype assembly. The x-axis represents the current assembly.

To further refine the assembly, gaps were closed using HiFi reads with the QuarTeT²⁵ (v. 1.1.3) software, and telomeric extensions were implemented. Chromosome termini with incomplete (TTTAGGG)_n repeats were extended by re-aligning HiFi reads to the assembly using Minimap2²⁶ (v. 2.24-r1122). Telomeric regions were reassembled into contigs and integrated into the final scaffolds to achieve near-telomere-to-telomere continuity. Given the incomplete assembly of chloroplast (160 Kbp) and mitochondrial (915 Kbp) genomes generated by Hifiasm, we employed OATK²⁷ (v. 0.9) to independently assemble these organelle genomes using HiFi reads.

Redundancy and contamination filtering. Isolated contigs were aligned to chromosome, mitochondrial, and chloroplast sequences using Redundans²⁸ (v. 0.13c). This workflow identified redundant fragments (e.g., haplotigs) and high-copy regions (e.g., rDNA arrays) prone to fragmentation in Hifiasm assemblies. Contigs with a typical low coverage or rDNA signatures were identified and flagged, followed by manual validation and systematic removal of suspected redundancies and contaminants in a consolidated workflow. In addition, the chromosome number ($2n = 38$) was determined based on the results of karyotype analysis. Finally, we generated a high-quality, haplotype-phased gapless genome of *M. zenii* (Fig. 1b).

Collinearity and synteny. Collinear analysis between the *M. zenii* assembly and reference genomes (*Pachylarnax sinica* (M. W. Law) N. H. Xia & C. M. Wu) using Minimap2 confirmed consistent chromosome ordering (Fig. 2).

Identification of chromosomal landmarks. Key genomic features were mapped onto chromosomes. Telomere repeats: The telomeric motif (TTTAGGG) was identified at termini of most chromosomes. Centromeric tandem repeat: A highly repetitive centromeric-like sequence (CCAACGGAGATGCTCATAACTCCCTCGTTTGATGTCGGATTGGGTCGATTTTGGTCATCTGAAAGGTACTTTAATGCTCTTTCATTCTGATTGGAATCGTGGTATTCCGAGACCATTGGGCATCCGTAAGTGGGATCGCAGTCGTCGGTCTGAATTTTCCTTCATCTCCTTGTCCTCAATAGTAACCTTCATAACTCTCTTTCAGTGGTTCGAATTTAGACTATTTTGAGCTCAT-TGGAATGGTAACGACAAAGCTTCCAAATGGGACCAGAAATGTCTAATTCTGAGACCGTTTCGATATCCC AAAAGCCCGTTGGAACTCTTCGATCGTCATTTTGCAATTTGGATTGTAAATCTTTCCAAATCGCGCGA) was detected across multiple chromosomes, consistent with centromeric regions. rDNA arrays: The 18-5.8-28S rDNA cluster was distributed across multiple chromosomes, as were 5S rDNA arrays (Supplementary Figure S1).

Annotation. Repetitive element characterization. Repetitive elements were identified using the EDTA²⁹ (v. 2.2.0; parameters: --sensitive 1--anno 1) pipeline to generate a transposable element (TE) library and annotate repetitive regions in the genome. In total, 5,100,109 repetitive sequences were identified, spanning 2.62 Gb and accounting for 72.87% of the genome. Among these, LTR retrotransposons were the most abundant class, representing 2,588,781 elements and 1.93 Gb, or 53.59% of the repetitive content (Table S5).

Gene annotation. A multi-evidence strategy was employed for gene annotation. Homology-based evidence was generated using non-redundant protein sequences from 14 related species (*M. biondii* Pampan., *M. sinica* (Y. W. Law) Noot., *Liriodendron chinense* (Hemsl.) Sargent., *Chimonanthus salicifolius* H. H. Hu, *Aristolochia fimbriata*

Feature	All genes					Coding genes				
	Number	Min.	Max.	Median	Mean	Number	Min.	Max.	Median	Mean
Gene	80,062	61	349,722	2,048	9,526	74,782	153	349,722	2,602	10,186
Transcript	92,505	61	21,715	1,083	1,495	87,225	153	21,715	1,187	1,575
CDS	87,225	153	20,727	795	1,061	87,225	153	20,727	795	1,061
Exon	421,311	3	20,758	153	328	415,982	3	20,758	155	330
Intron	328,806	4	197,629	613	2,773	328,757	22	197,629	613	2,774
Exons Transcript	92,505	1	79	3	5	87,225	1	79	3	5

Table 2. Gene annotation statistics of *M. zenii*.

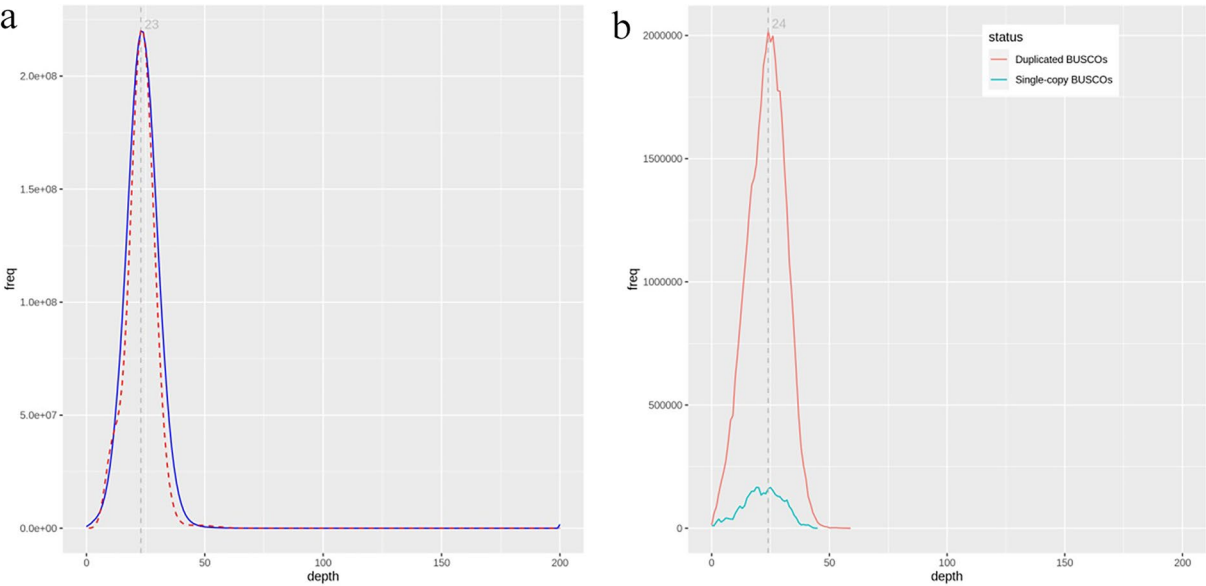


Fig. 3 Distribution of coverage depth assessed using HiFi data for the whole genome (a) and BUSCO core gene regions (b).

Cham., *Annona glabra* Linn., *Piper nigrum* Linn., *Saururus chinensis* (Lour.) Baill., *Warburgia ugandensis* Sprague, *Cinnamomum kanehirae* (Hayata) K.F.Chung & C.L.Hsieh, *Acorus tatarinowii* Schott, *Amborella trichopoda* Baill., *Tetracentron sinense* Oliv., *Ceratophyllum demersum* L.) in an aggregated database (398,887 sequences). For transcriptome evidence, ONT long-reads were mapped with Minimap2 and assembled into transcripts via StringTie³⁰ (v. 1.3.5), assembly statistics are presented in Table 2.

Transcriptome assembly results were first annotated using the PASApipeline³¹ (v. 2.5.3) to define gene structures, followed by alignment to the reference protein dataset of 398,887 non-redundant sequences. Full-length genes were identified through this alignment, and their sequences were used to train both AUGUSTUS³² (v. 3.5.0) and SNAP³³, with five rounds of iterative optimization.

The MAKER2³⁴ (v. 2.31.9) annotation workflow was then applied, integrating *ab initio* predictions (via AUGUSTUS and SNAP), transcript evidence, and homologous protein evidence. Key steps included: (1) masking repetitive regions with RepeatMasker; (2) *de novo* gene prediction using AUGUSTUS and SNAP; (3) aligning transcript sequences to the genome via BLASTN and protein sequences via BLASTX; (4) refining alignments with Exonerate³⁵ (v. 2.2.0); and (5) generating hints files to guide the integration of gene models by AUGUSTUS and SNAP.

To address the lower precision of MAKER process annotations³⁶, EvidenceModeler³⁷ (EVM, v. 2.0.0) was employed to consolidate MAKER and PASA results into a unified gene set. TE protein domains identified by TESorter³⁸ (v. 1.4.6) were masked during EVM integration to exclude transposable element (TE) coding regions. PASA was further applied to refine EVM annotations by incorporating untranslated regions (UTRs) and alternative splicing events. Final annotations were filtered to exclude genes with disrupted coding frames (e.g., internal stop codons, ambiguous bases, or missing start/stop codons) or lengths less than 50 amino acids. ncRNA annotation was performed with tRNAScan-SE³⁹ (v. 2.0.12) for tRNAs, barrnap (<https://github.com/tseemann/barrnap>) for rRNAs, and RfamScan for other ncRNAs, ensuring comprehensive functional characterization.

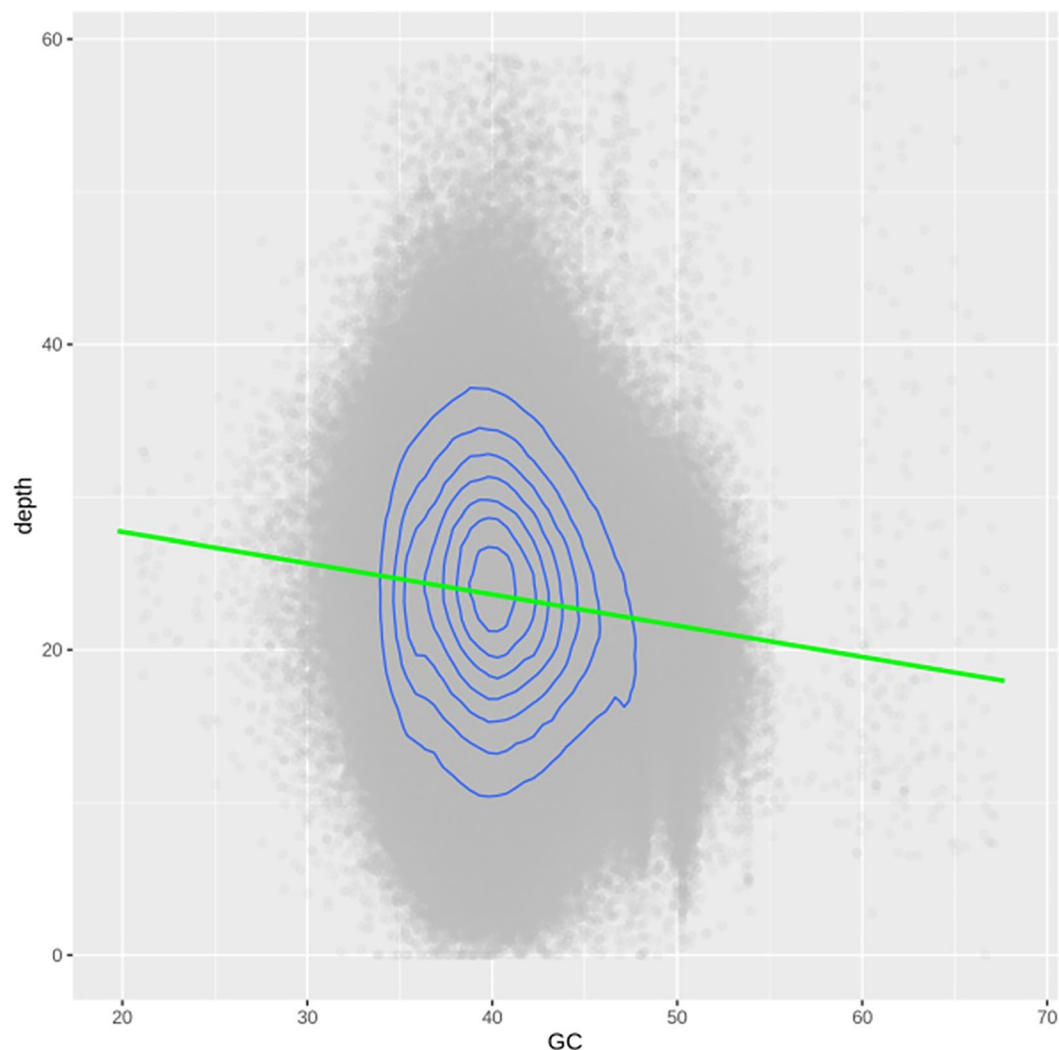


Fig. 4 Coverage depth distribution of HiFi sequencing data across varying GC content.

Integrated annotation and functional characterization. Protein-coding genes were named systematically using a species-specific nomenclature (e.g., Yuzen01aG0000100), while non-coding RNAs were annotated separately.

Functional annotation combined homology-based approaches. The functional annotation of protein-coding genes was conducted using three strategies. Gene functions were annotated by aligning against the EggNOG-Mapper⁴⁰ homologous gene database, incorporating functional information such as Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG). Protein sequences were aligned against protein databases (Swiss-Prot, TrEMBL, NCBI NR, and *Arabidopsis*) using DIAMOND⁴¹ (v. 2.4.0; parameters: Identity $\geq 30\%$, E-value $< 1e-5$) to identify the best hit for each gene. Finally, the sequence was aligned with the sub databases in InterPro (PRIMTS, Pfam, SMART, PANTHER, and CDD) using InterProScan⁴² (v. 5.14-53.0) to identify the conserved domains and motifs. Ultimately, 15,883 genes (21.24%) remained unannotated. Full annotation statistics are provided in Table S6.

Data Record

All data associated with this study have been deposited in the National Genomics Data Center (NGDC)^{43,44}, Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation, under the BioProject (accession number PRJCA037928) and the BioSample (accession number SAMC5173072), accessible via <https://ngdc.cncb.ac.cn/gwh/>. The raw sequencing data (Hi-C reads, ONT reads, PacBio HiFi reads and Iso-Seq ONT reads) have been deposited in the Genome Sequence Archive (GSA) in NGDC under the accession number CRR1877518⁴⁵, CRR1877520⁴⁶, CRR1877521⁴⁷ and CRR1877519⁴⁸, respectively. The complete genome assembly and annotation have been submitted in the Genome Warehouse (GWH) in NGDC under the accession number GWHGEBD00000000.1⁴⁹. The sequencing data have been also deposited in the NCBI Sequence Read Archive (SRA) with accession number SRR33871901⁵⁰, SRR33871902⁵¹, SRR33871903⁵² and SRR33871904⁵³ under BioProject accession number PRJNA1273192. The two haplotypes of the final assembled

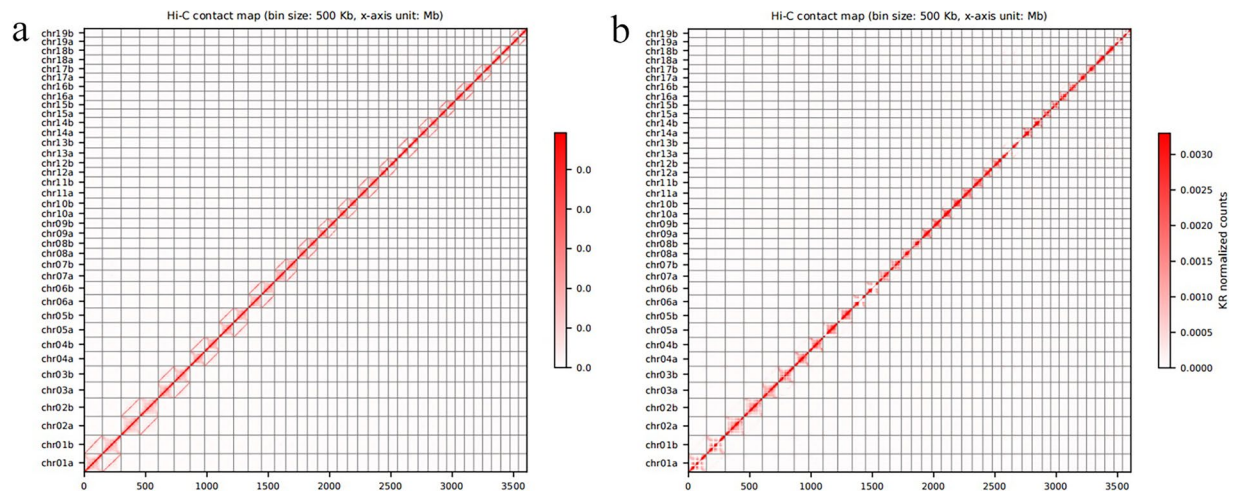


Fig. 5 Hi-C contact maps of the final assembled genome. **(a)** Minimum mapping quality of 1; **(b)** Minimum mapping quality of 0.

genome have also been submitted to GenBank, with assembly accession numbers GCA_052055185.1⁵⁴ and GCA_052055145.1⁵⁵.

Technical Validation

Assessment of the genome assembly. The final genome assembly spans 3.6 Gb, containing two complete haplotypes consistent with the k-mer based size estimation. The assembly achieved 85 Mb contig N50 and 7 gaps, demonstrating high continuity (Table 1). To evaluate coverage, HiFi reads were mapped using BWA⁵⁶ (v. 0.7.17-r1188), while transcriptomic ONT reads were mapped to the genome using Minimap2. Non-primary alignments were filtered out, both HiFi reads and Iso-seq demonstrated very high mapping rates of 99.14% and 95.29%, respectively. This indicates that our sequencing data provides relatively high coverage of the genome. Redundancy was assessed by remapping sequencing data to the assembled genome (filtered for non-primary alignments). In redundant regions, coverage depths were approximately half of the expected values due to allelic phasing. The absence of prominent peaks preceding the main coverage peak (determined by genome size estimation) indicated no significant redundancy (Fig. 3a). Similarly, BUSCO analysis of single-copy and multi-copy core genes showed uniform coverage distributions with no overlapping peaks (Fig. 3b). GC content distribution across HiFi sequencing data at varying coverage depths showed no detectable bias (Fig. 4), ensuring even representation of genomic regions. Remapping Hi-C data to the final assembly using BWA demonstrated strong chromosome clustering (Fig. 5), with no evident misassembly artifacts.

Assessment of the gene annotation. The genome was evaluated using BUSCO, revealing that 99.4% of core genes were complete (including both single-copy and duplicated genes), with only 0.3% missing, indicating a high level of genome completeness. For the transcriptome data integrated with annotated proteins, BUSCO analysis showed that 98.4% of core genes were complete. Additionally, 0.4% were fragmented (F) and 1.2% were missing (M), with a total of 1,614 BUSCOs assessed.

Data availability

All raw data from this study are publicly accessible via NGDC (<https://download.cncb.ac.cn/gsa5/CRA026213>) and NCBI (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1273192>). The complete genome assembly and annotation results are also available through both platforms (NGDC: <https://ngdc.cncb.ac.cn/gwh/Assembly/98046/show>; NCBI: http://identifiers.org/ncbi/insdc.gca:GCA_052055185.1 and http://identifiers.org/ncbi/insdc.gca:GCA_052055145.1).

Code availability

All data processing procedures were performed following the protocols and manuals of the bioinformatics software, without employing any custom codes or scripts in this study. Detailed parameters are thoroughly documented in the Methods section.

Received: 23 June 2025; Accepted: 24 October 2025;

Published online: 08 December 2025

References

1. Wang, S., Xie, Y. China species red list (Beijing: Higher Education Press, 2004).
2. Li, N. *et al.* Analyses on population structure and dynamic characteristics of *Yulania zenii* in Baohua Mountain of Jiangsu Province. *Journal of Plant Resources and Environment* **33**, 98–105 (2024).
3. Wang, P. *et al.* Classification and ordination of a *Magnolia zenii* community based on a self-organizing feature map. *Journal of Zhejiang A&F University* **36**, 255–263 (2019).

4. Li, Y. *et al.* The complete plastid genome of *Magnolid zenii* and genetic comparison to Magnoliaceae species. *Molecules* **24**, 261 (2019).
5. Chen, Y. & Nan, C. ISSR analysis on genetic diversity of endangered plant *Magnolia zenii*. *Journal of Sichuan Agricultural University* **34**, 445–449 (2016).
6. Wang, S., Shen, Y., Bao, H. & He, H. Change of anatomical structure in the process of mega- and microsporogenesis and female and male gametophyte development of *Yulania zenii*. *Journal of Plant Resources and Environment* **30**, 46–53 (2021).
7. Rivers, M. *et al.* The Red List of Magnoliaceae—Revised and Extended (Richmond, VA: Botanic Gardens Conservation International, 2016).
8. Qin, H. *et al.* Threatened species list of China's higher plants. *Biodiv Sci* **2017** **25**, 696–744 (2017).
9. Chen, J. *et al.* *Liriodendron* genome sheds light on angiosperm phylogeny and species-pair differentiation. *Nat. Plants* **5**, 18–25 (2019).
10. Doing, S. *et al.* The genome of *Magnolia biondii* Pamp. provides insights into the evolution of Magnoliales and biosynthesis of terpenoids. *Hortic. Res.* **8**, 38 (2021).
11. Shi, C. *et al.* A high-quality chromosome-level genome assembly of the endangered tree *Kmeria septentrionalis*. *Sci. Data* **11**, 775 (2024).
12. Zhou, L. *et al.* The genome of *Magnolia hypoleuca* provides a new insight into cold tolerance and the evolutionary position of magnoliids. *Front. Plant Sci.* **14**, 1108701 (2023).
13. Yin, M. *et al.* The chromosome-scale genome of *Magnolia officinalis* provides insight into the evolutionary position of magnoliids. *iScience* **24**, 102997 (2021).
14. Lu, X. *et al.* The Chromosome-Scale Genome of *Magnolia sieboldii* K. Koch Provides Insight Into the Evolutionary Position of Magnoliids and Seed Germination. *Mol. Ecol. Resour.* **25**, e14030 (2025).
15. Cai, L. *et al.* The chromosome-scale genome of *Magnolia sinica* (Magnoliaceae) provides insights into the conservation of plant species with extremely small populations (PESP). *GigaScience* **13**, giad110 (2024).
16. Jiang, S. *et al.* A high-quality haplotype genome of *Michelia alba* DC reveals differences in methylation patterns and flower characteristics. *Mol. Hortic.* **4**, 23 (2024).
17. Jiang, G., Zhang, G., Zhang, G. & Wang, J. Intra- and interspecific competition of endangered plant *Magnolia zenii*. *Chinese Journal of Ecology* **29**, 201–206 (2010).
18. Simão, F. A. *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212. *ytocchem Bull.* **19**, 11–15 (1987) (2015).
19. Doyle, J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull.* **19**, 11–15 (1987).
20. Chen, S., Zhou, Y., Chen, Y. & Gu, J. *fastp*: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
21. Cheng, H. *et al.* Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
22. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
23. Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Sci.* **356**, 92–95 (2017).
24. Durand, N. C. *et al.* Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
25. Lin, Y. *et al.* quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic. Res.* **10**, uhad127 (2023).
26. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
27. Zhou, C. *et al.* OatK: a de novo assembly tool for complex plant organelle genomes. *Genome Biol.* **26**, 235 (2025).
28. Pryszcz, L. P. & Gabaldón, T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* **44**, e113 (2016).
29. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
30. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
31. Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
32. 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 (2008).
33. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
34. Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
35. Ewan, B. & Guy, S. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
36. Cook, D. E. *et al.* Long Read Annotation: automated eukaryotic genome annotation based on long-read cDNA sequencing. *Plant Physiol.* **179**, 38–54 (2019).
37. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVIDENCEModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
38. Zhang, R. *et al.* TESorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic. Res.* **9**, uhac17 (2022).
39. 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 (1997).
40. Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
41. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
42. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
43. Chen, M. *et al.* Genome Warehouse: a public repository housing genome-scale data. *Genom. Proteom. Bioinform.* **19**, 584–589 (2021).
44. Database Resources of the National Genomics Data Center. China National Center for Bioinformation in 2022. *Nucleic Acids Res.* **50**, D27–D38 (2022).
45. NGDC Genome Sequence Archive <https://ngdc.cnbc.ac.cn/gsa/browse/CRA026213/CRR1877518> (2025).
46. NGDC Genome Sequence Archive <https://ngdc.cnbc.ac.cn/gsa/browse/CRA026213/CRR1877520> (2025).
47. NGDC Genome Sequence Archive <https://ngdc.cnbc.ac.cn/gsa/browse/CRA026213/CRR1877521> (2025).
48. NGDC Genome Sequence Archive <https://ngdc.cnbc.ac.cn/gsa/browse/CRA026213/CRR1877519> (2025).
49. NGDC Genome Warehouse <https://ngdc.cnbc.ac.cn/gwh/Assembly/98046/show> (2025).
50. NCBI Sequence Read Archive <http://identifiers.org/ncbi/insdc.sra:SRX29084017> (2025).
51. NCBI Sequence Read Archive <http://identifiers.org/ncbi/insdc.sra:SRX29084016> (2025).
52. NCBI Sequence Read Archive <http://identifiers.org/ncbi/insdc.sra:SRX29084015> (2025).
53. NCBI Sequence Read Archive <http://identifiers.org/ncbi/insdc.sra:SRX29084014> (2025).
54. NCBI Assembly http://identifiers.org/ncbi/insdc.gca:GCA_052055185.1 (2025).
55. NCBI Assembly http://identifiers.org/ncbi/insdc.gca:GCA_052055145.1 (2025).
56. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* (2013).

Acknowledgements

This work was supported by the Scientific Fund of Nanjing Botanical Garden Men. Sun Yat-Sen (Grant No. JSPKLB202506).

Author contributions

B.W. conceived and designed the study. N.L. was responsible for sample collection and preparation and performed bioinformatics analysis. S.Y. collected the samples. B.W. drafted the manuscript, while N.L., Z.Z., S.Y. and D.Y. revised the manuscript. All authors contributed to the manuscript and approved the version for submission.

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/s41597-025-06208-w>.

Correspondence and requests for materials should be addressed to B.-c.W.

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.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025