scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A high-quality chromosome-level genome assembly of the endangered species *Magnolia amoena*

Yu Liu[1,2], Xing-Jian Liu[1,2], Ke Hu[1], Ming-Han Li[1], Zi-Jie Shen[1], Xiao-Qin Sun[1] & Rui-Sen Lu[1] ✉

Magnolia amoena **(Magnoliaceae), a deciduous tree endemic to eastern China, is valued for its striking floral diversity and traditional medicinal uses, yet remains understudied. To support its conservation and evolutionary research, we generated a high-quality chromosome-scale assembly using DNBSEQ-T7, PacBio HiFi, and Hi-C data. The final assembly spanned 1.87 Gb with a contig N50 of 36.92 Mb, of which 95.73% (1.79 Gb) was anchored onto 19 pseudochromosomes. Repetitive elements occupied 79.55% of the genome, dominated by long terminal repeats (57.72%) and DNA transposons (14.25%). A total of 39,739 protein-coding genes (mean length: 10.21 kb) were predicted, with 86.34% functionally annotated, alongside 270 miRNAs, 631 tRNAs, 668 rRNAs, and 4,133 snRNAs. Comparative genomic analysis across 11 magnoliid species identified 23,244 gene families in** ***M. amoena*****, of which 1,905 were unique. Phylogenomic reconstruction strongly supported** ***M. amoena*** **as sister to** ***M. biondii*****, with an estimated divergence time of ~18.5 Mya. Overall, this genome assembly lays a robust foundation for future research into the evolution, adaptation, and conservation of** ***M. amoena*****.**

Background & Summary

The family Magnoliaceae Juss., an early-diverging lineage within the Magnoliids (APG IV, 2016), comprises over 300 species of evergreen and deciduous trees and shrubs[1]. Distributed across temperate and tropical regions of Southeast and East Asia, North America, and parts of Central and South America[2], Magnoliaceae is now often taxonomically divided into two subfamilies: Liriodendroideae, which includes *Liriodendron*, and Magnolioideae, which, under the broad generic concept, includes all remaining genera merged into *Magnolia*[1,3]. This family is evolutionarily significant for its ancient floral characteristics, particularly spirally arranged stamens and pistils on a conical receptacle, providing valuable insights into the early evolution of angiosperms[4]. Many species are cultivated worldwide as ornamental trees, valued for their attractive flowers[5]. Additionally, several species have long been used in traditional medicine in East Asia, and bioactive compounds like magnolol and honokiol are now actively studied for their pharmacological effects[6–8].

Magnolia amoena W.C.Cheng (2n = 2x = 38), a deciduous tree endemic to eastern China[9], inhabits fragmented mountain forests at an altitude of 200–1,200 m across Zhejiang, Anhui, Jiangsu, and Jiangxi provinces[10,11]. This species is renowned for its striking floral diversity, with flowers displaying a wide range of colors, including shades of red, pink, and white (Fig. 1). Additionally, the flower buds have long been used in traditional Chinese medicine[12]. However, overharvesting of these buds is believed to impair the species' natural regeneration capacity[10], posing a serious threat to its survival. Consequently, *M. amoena* is currently categorized as Vulnerable in both the IUCN Red List[13] and China Biodiversity Red List—Higher Plants Volume (2020)[14]. Despite its ecological and traditional medical significance, research on this species remains limited, largely focusing on specific aspects such as volatile compound profiling[15], domestication trials[16], distribution mapping[17] and genetic diversity assessment[18]. Moreover, the limited availability of genomic resources further constrains deeper investigations into its evolutionary dynamics and habitat-specific adaptation.

[1]Jiangsu Key Laboratory for Conservation and Utilization of Plant Resources, Institute of Botany, Jiangsu Province and Chinese Academy of Sciences, Nanjing, 210014, China. [2]These authors contributed equally: Yu Liu, Xing-Jian Liu. ✉e-mail: lurs@cnbg.net

Fig. 1 Flowers of *M. amoena*.

To address these gaps and advance effective conservation, we generated a high-quality chromosome-level genome assembly of *M. amoena*, using a hybrid approach that combined DNBSEQ, PacBio HiFi, and Hi-C technologies (Fig. 2). The genome assembly was 1.87 Gb, with a contig N50 of 36.92 Mb (Table 1), and 95.73% (1.79 Gb) was successfully anchored to 19 pseudochromosomes (Table S1). Repetitive sequences constituted 79.55% of the genome, with long terminal repeats (LTR) and DNA transposons as the predominant types, accounting for 57.72% and 14.25%, respectively (Table 2). Gene prediction identified 39,739 protein-coding genes (Table 3), of which 86.34% (34,312) were functionally annotated (Table 4). In addition, the genome harbors 270 miRNAs, 631 tRNAs, 668 rRNAs and 4,133 snRNAs (Table 5). Comparative genomic analysis across 11 magnoliid species identified 23,244 gene families in *M. amoena*, of which 1,905 were species-specific (Table S2). Further comparison with its close relatives revealed that *M. amoena* has a relatively compact genome, a high repeat content, and an intermediate number of predicted protein-coding genes (Table S3). Phylogenetic reconstruction placed *M. amoena* in a strongly supported clade with *M. biondii* (bootstrap value = 100%), and divergence time estimation indicated that the two species diverged approximately 18.5 million years ago (Mya) (Fig. 2B). Collectively, this study provides an important genomic resource for *M. amoena*, supporting its conservation and sustainable utilization, as well as future research on magnoliid evolution.

Methods

Sample collection. Fresh young leaves of *M. amoena* were collected from a single adult plant at the Institute of Botany, Jiangsu Province and Chinese Academy of Sciences (Nanjing Botanical Garden Mem. Sun Yat-sen) for genomic DNA extraction. Additionally, leaves, shoots, flower buds, and fruits from the same individual were sampled for transcriptome sequencing (RNA-Seq) and full-length transcript sequencing (Iso-Seq). All fresh plant materials were immediately frozen in liquid nitrogen and stored at −80 °C until further processing.

Library construction and genome sequencing. Genomic DNA was extracted from young leaves using the cetyltrimethylammonium bromide (CTAB) method[19]. DNA concentration was quantified using a NanoDrop spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA) and Qubit 3.0 fluorometer (Life Technologies, Carlsbad, California, USA), while purity and integrity were assessed via 1% agarose gel electrophoresis.

For short-read sequencing, high-quality genomic DNA was used to construct a paired-end library with an insert size of 200–400 bp, according to the manufacturer's protocol. Paired-end sequencing (2 × 150 bp) was performed on the DNBSEQ-T7 platform (BGI, Shenzhen, China), generating ~104.10 Gb raw data (~61.24 × coverage).

For PacBio HiFi sequencing, approximately 8 μg of genomic DNA was sheared using the Megaruptor 3 system (Diagenode., Seraing, Belgium), followed by size selection and concentration with AMPure PB magnetic beads (Pacific Biosciences, Menlo Park, CA, USA). SMRTbell libraries were prepared using the SMRTbell Express Template Prep Kit 3.0 (Pacific Biosciences, Menlo Park, CA, USA), with ~15 kb inserts selected via the BluePippin system (Sage Science, Beverly, MA, USA). PacBio HiFi sequencing was conducted on a PacBio Revio platform in circular consensus sequencing (CCS) mode. Raw data were processed using SMRTLink v13.0 with the default parameters, generating ~82.96 Gb (~48.8 × coverage) of HiFi reads with an average read length of ~15.72 kb and an N50 of ~16 kb.

For Hi-C sequencing, fresh leaf tissue was fixed with 2% formaldehyde. The cross-linked DNA was digested with DpnII restriction enzyme, and the resulting sticky ends were labeled by incorporating biotinylated nucleotides. Proximity ligation was performed under highly diluted conditions to favor ligation between spatially proximate fragments. After reversing cross-links, the DNA was purified and sheared to 300–500 bp. Biotin-labeled fragments were enriched using streptavidin magnetic beads, followed by PCR amplification. The Hi-C library was sequenced on the Illumina NovaSeq X Plus platform (Illumina, San Diego, CA, USA) using a PE150 strategy, generating ~213.19 Gb of raw data (~125.41 × coverage).

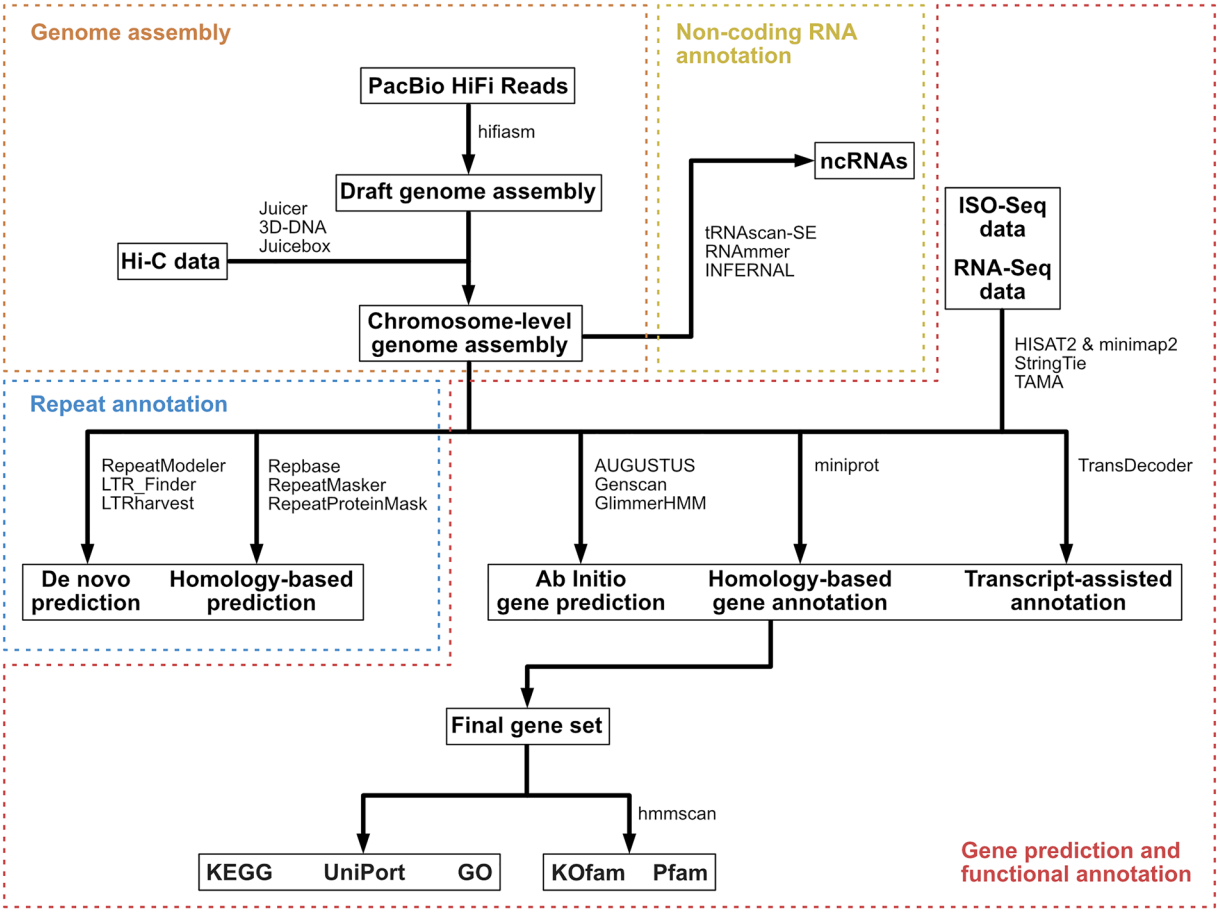

Fig. 2 Flowchart of chromosome-level genome assembly and annotation of *M. amoena*.

Assembly feature	Value
Genome size (bp)	1,872,648,096
GC content (%)	40.62%
Chromosome	19
Contig number	1,132
Contig N50 (bp)	36,921,386
Average contig length (bp)	1,654,276.67
Maximum contig length (bp)	120,037,700
Scaffold number	1,069
Scaffold N50 (bp)	91,492,038
Average scaffold length (bp)	1,751,775.58
Maximum scaffold length (bp)	159,354,873

Table 1. Assembly statistics of the *M. amoena* genome.

Additionally, for RNA sequencing, total RNA was extracted from leaves, shoots, flower buds, and fruits of the same plant, using the RNAprep Pure Plant Kit (Tiangen Biotech, Beijing, China). Stranded RNA-Seq libraries were constructed using the MGIEasy RNA Sample Prep Kit (BGI, Shenzhen, China) and subjected to sequencing on the DNBSEQ-T7 platform. For Full-length isoform sequencing (Iso-Seq), RNA aliquots from each of the four tissues were pooled into a single sample and sequenced on the PromethION platform (Oxford Nanopore Technologies, U.K.), generating approximately 10.95 Gb of full-length Iso-Seq data. All library preparations and sequencing procedures were carried out by Benagen Technology Co., Ltd. (Wuhan, China).

Genome survey and assembly. Raw short-read sequencing data were processed using fastp v0.23.2[20] to remove adapter sequences, excessively short reads, and low-quality reads (with parameters:–qualified_quality_phred 15,–length_required 30,–detect_adapter_for_pe). Clean reads were then analyzed with Jellyfish v2.2.10[21] to

Type	Length (bp)	Percentage of genome
DNA	255,361,602	14.25
LINE	80,472,348	4.49
SINE	6,634,040	0.37
LTR	1,034,689,186	57.72
LTR-Gypsy	315,529,785	17.60
LTR-Copia	229,165,428	12.78
Other	3,568	0.00
Unknown	153,398,846	8.56
Total	1,426,033,823	79.55

Table 2. Summary of transposable elements in the *M. amoena* genome.

Category	Value
Number of gene	39,739
Average mRNA length (bp)	10,207.74
Average CDS length per gene (bp)	997.06
Average exon number per gene	4.3
Average exon length (bp)	311.53
Average intron length (bp)	2,690.29
Total exon count	170,748
Total intron count	131,009
Total intron length (bp)	352,452,283

Table 3. Summary of protein-coding genes in the *M. amoena* genome.

Database	Annotated genes	Percentage (%)
General annotation	34,312	86.34
KEGG	20,866	52.51
Pathway	10,884	27.39
Nr	34,207	86.08
Uniprot	31,427	79.08
GO	24,225	60.96
KOG	1,786	4.49
Pfam	22,766	57.29

Table 4. Functional annotation of protein-coding genes in the *M. amoena* genome.

Type	Number	Average length (bp)	Total length (bp)	Percentage of genome (%)
tRNA	631	75	47,424	0.0026
rRNA	668	814	544,039	0.0303
miRNA	270	126	33,984	0.0019
snRNA	4,133	109	451,420	0.0252

Table 5. Annotation statistics of non-coding RNAs in the *M. amoena* genome.

generate a *19*-mer frequency distribution. The resulting *k*-mer profile was used as input for GenomeScope v2.0[22] to estimate genomic characteristics. The *M. amoena* genome was estimated to be approximately 1.70 Gb in size (Fig. 3A), consistent with the estimation of 1.70 Gb using flow cytometry. Additionally, the genome exhibited a heterozygosity rate of 1.15% and a duplication rate of 60.34%, indicating that the *M. amoena* genome is highly heterozygous and extensively duplicated.

De novo genome assembly was performed using hifiasm v0.19.8-r603[23], utilizing PacBio HiFi long reads (Fig. 2). The draft genome assembly consisted of 1,132 contigs, with a total length of ~1.87 Gb, a GC content of 40.62%, and a contig N50 of ~36.92 Mb (Table 1). For chromosome-level scaffolding, Hi-C reads were preprocessed using fastp v0.23.2[20], retaining ~210.79 Gb of clean data with Q20 and Q30 scores of 98.34% and 94.89%, respectively, and then aligned to the draft assembly using Juicer v1.6[24]. Pseudochromosome construction was performed with 3D-DNA v201008[25], and Hi-C interaction heatmaps were generated at 200 kb resolution using

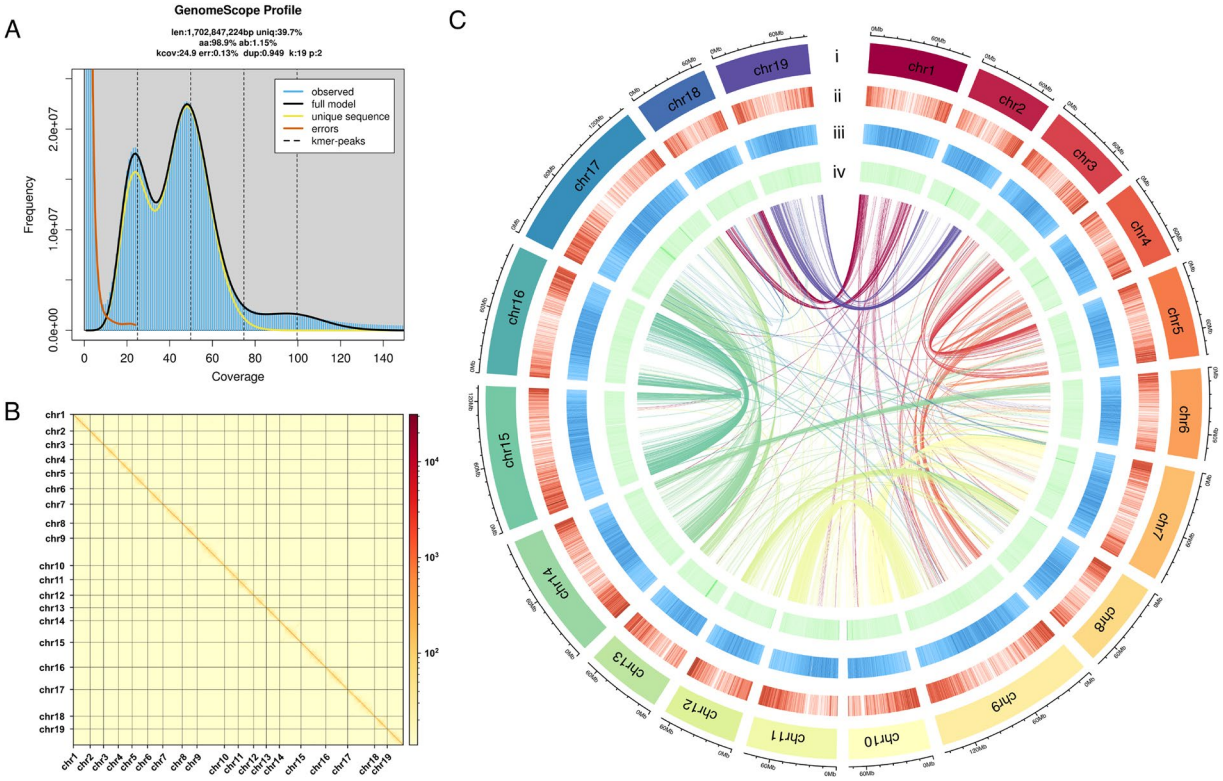

Fig. 3 Genomic features of the *M. amoena* genome. (**A**) *K*-mer analysis based on the *19*-mer frequency distribution. (**B**) Hi-C interaction heatmap at 200 kb bin resolution, showing pairwise correlations among the 19 pseudomolecules of *M. amoena*. Higher interaction strength is indicated by increased red intensity. (**C**) An overview of the chromosome-level genome of *M. amoena*. The concentric circles, from outermost to innermost, display: (i) gene density, (ii) repeat elements density, (iii) GC content and (iv) intragenomic synteny links within 500Kbp sliding windows.

the visualization module in 3D-DNA[25] and Juicebox v1.11.08[26] (Fig. 3B). The final chromosome-level assembly successfully anchored 95.73% (~1.79 Gb) of the genome sequence onto 19 chromosomes (Fig. 3C; Table S1).

Repeat annotation. A comprehensive repeat annotation strategy was employed, combining *de novo* prediction and homology-based methods (Fig. 2). First, a *de novo* repeat library was constructed using RepeatModeler v2.0.6[27] (with parameters -database mydb -threads 16), supplemented with LTR retrotransposon predictions from LTR_Finder v1.0.7[28] (with parameters -threads 16 -harvest_out -size 1000000 -time 300) and LTRharvest v1.6.5[29] (with parameters -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1 -similar 85 -vic 10 -seed 20 -seqids yes). The resulting sequences were classified using TEclass v2.1.3[30] with default parameters, to reannotate unknown elements. This *de novo* library was then merged with Repbase (https://www.girinst.org/repbase/) and analyzed through RepeatMasker v4.1.7 (with parameters -noLowSimple -pvalue 0.0001) (http://repeatmasker.org/) to generate combined *de novo* + RepBase predictions. Protein-coding transposable elements were identified using RepeatProteinMask v4.1.7 (with parameters -noLowSimple -pvalue 0.0001) (https://www.repeatmasker.org/RepeatProteinMask.html). Finally, all repeat predictions were merged and deduplicated to produce a comprehensive set of transposable elements.

A total of ~1.43 Gb of repetitive elements were identified in the *M. amoena* genome, which constituted 79.55% of the genome. LTR retrotransposons dominated these repeat elements, accounting for 57.72% of the genome, of which Gypsy (17.60%) and Copia (12.78%) superfamily elements were the most abundant (Table 2).

Gene prediction and functional annotation. Gene annotation was achieved through three complementary approaches: transcriptome-guided prediction, homology-based inference, and *ab initio* gene modeling (Fig. 2). For transcript-assistant annotation, DNBSEQ-T7 short-read RNA-Seq and PacBio Iso-Seq data were quality-filtered and mapped to the *M. amoena* genome using HISAT v2.2.1[31] and minimap2 v2.26-r1175[32], respectively. Transcripts were assembled using StringTie v2.2.1[33] with default parameters. Subsequently, transcripts from all samples were merged by the Transcriptome Annotation by Modular Algorithms (TAMA) v1.0[34], and were analyzed for protein-coding sequences using TransDecoder v5.7.0 (https://github.com/TransDecoder/TransDecoder). Homology-based prediction was performed by aligning protein sequences from five published *Magnolia* species, including *M. alba*[35], *M. biondii*[36], *M. obovate*[37], *M. officinalis*[38], and *M. sinica*[39] to the genome using miniprot v0.13[40]. *Ab initio* prediction was conducted using AUGUSTUS v3.5.0[41], Genscan v1.0[42], and GlimmerHMM v3.0.4[43], each trained on high-quality transcripts generated in the previous steps. The

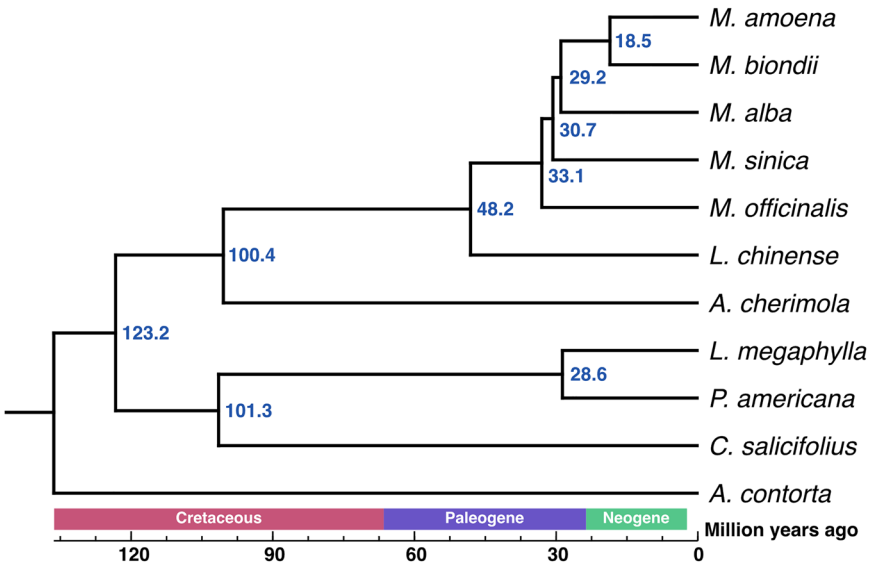

Fig. 4 Phylogenetic relationships and estimated divergence time of *M. amoena* and ten other magnoliid species. Blue numbers at each node indicate the inferred divergence times million years ago (Mya).

outputs of all methods were integrated with MAKER v3.01.03[44] and EVidenceModeler v2.1.0[45]. In total, 39,739 protein-coding genes were predicted in the *M. amoena* genome, with an average length of 10,207.74 bp (Table 3).

Functional annotation of protein-coding genes was performed through both sequence similarity and domain-based analyses (Fig. 2). Protein sequences were aligned against the Universal Protein Knowledgebase (UniProt) database[46], the Kyoto Encyclopedia of Genes and Genomes (KEGG) database[47], and the Gene Ontology (GO) database[48] using DIAMOND v2.1.8[49]. Domain-based annotation was performed using hmmscan v3.3.2[50] against two major hidden Markov model (HMM) domain databases, Pfam v37.0[51] and KOfam v2024-05-27[52]. In total, 34,312 protein-coding genes were functionally annotated, accounting for 86.34% of all predicted protein-coding genes (Table 4; Fig. S1).

For the annotation of non-coding RNAs, tRNAscan-SE v.2.0.12[53], RNAmmer v1.2[54], and Infernal v1.1.4[55] were employed to predict the tRNAs, rRNAs, and other various non-coding RNAs (e.g., miRNA and snRNA), respectively. In total, 631 tRNA, 668 rRNA, 270 miRNA, and 4,133 snRNA were identified in the *M. amoena* genome (Table 5).

Phylogenetic analysis and divergence time estimation. Orthogroups were identified from *M. amoena* and nine closely related species, including *M. alba*, *M. biondii*, *M. officinalis*, *M. sinica*, *Annona cherimola*, *Chimonanthus salicifolius*, *Liriodendron chinense*, *Lindera megaphylla*, and *Persea americana*, using OrthoFinder v 2.5.5[56], with *Aristolochia contorta* as the outgroup. A total of 82,671 orthologous families (425,438 genes), including 573 single-copy nuclear genes (SCNGs) were yielded (Table S2). These SCNGs were aligned using MUSCLE v3.8.31[57], and poorly aligned regions were trimmed with trimAl v1.5[58]. The concatenated alignments were used to construct a maximum-likelihood (ML) phylogeny with RAxML v8.2.13[59], with 1000 bootstrap replicates. Divergence times were estimated via the program MCMCtree, implemented in PAML v4.10.7[60]. Four calibration points were derived from TimeTree (http://www.timetree.org/), including the split between *A. contorta* and *L. chinense* (130.0–142.6 Mya), *L. megaphylla* and *M. biondii* (114.0–130.3 Mya), *A. cherimola* and *M. biondii* (95.0–113.6 Mya), and *L. chinense* and *M. biondii* (27.8–50.0 Mya).

The resulting phylogenetic tree revealed that *M. amoena* formed a well-supported monophyletic clade with *M. biondii*, *M. officinalis*, *M. sinica*, and *M. alba*, which is sister to *L. chinense*. This topology suggests a close evolutionary relationship between *Magnolia* and *Liriodendron*, with their divergence occurring around 48.2 Mya during the Middle Eocene. Within *Magnolia*, *M. amoena* is most closely related to *M. biondii*, with the divergence time estimated at approximately 18.5 Mya (Fig. 4).

Data Records

The chromosome assembly data has been uploaded and deposited in NCBI GenBank/ENA under accession number JBNYVC000000000[61]. The genome assembly and annotation files have been submitted to figshare database[62]. All sequencing data generated for *Magnolia amoena* have been deposited in the NCBI Sequence Read Archive under BioProject accession number PRJNA1264003. DNBSEQ-T7 short-read, PacBio long-read and Hi-C sequencing data are available under SRR33610905[63], SRR33610904[64] and SRR33610903[65], respectively. Full-length transcript sequencing (Iso-Seq) data are available under SRR33610898[66]. RNA-Seq datasets from leaf, shoot, flower, and fruit tissues are available under SRR33610902[67], SRR33610901[68], SRR33610900[69], and SRR33610899[70], respectively.

Technical Validation

The completeness and quality of the genome assembly and gene annotation were assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.4.7[71] with the embryophyta_odb10 lineage dataset. The genome assembly showed 99.32% completeness, with 94.67% (1,528) single-copy and 4.65% (75) duplicated BUSCOs. Only 0.37% (6) were fragmented, and 0.31% (5) were missing (Table S4). Gene annotation similarly demonstrated high completeness, including 95.11% (1,535) single-copy and 4.21% (68) duplicated. Fragmented and missing BUSCOs accounted for only 0.37% (6) and 0.31% (5), respectively. These metrics indicated that both the assembly and annotation are highly complete and accurate. Additionally, ~99.48% of DNBSEQ-T7 reads were mapped to the assembly, achieving 99.85% genome coverage with an average depth of 54.58×, confirming uniform sequencing and high assembly quality for downstream analyses.

Data availability

All data generated for *Magnolia amoena* in this study have been made publicly available. The chromosome-level genome assembly has been deposited in NCBI GenBank under accession number JBNYVC000000000. Raw sequencing reads (DNBSEQ-T7 short-read, PacBio long-read, Hi-C, Iso-Seq, and RNA-Seq) are available in the NCBI Sequence Read Archive under BioProject accession PRJNA1264003. Genome annotation files are accessible via figshare (https://doi.org/10.6084/m9.figshare.29095601).

Code availability

Software and analysis pipelines were implemented according to the official manuals and published protocols of established bioinformatics tools. Detailed software versions and parameters are provided in the Methods section.

Received: 26 May 2025; Accepted: 25 February 2026;
Published online: 05 March 2026

References

1. Wang, Y. B. *et al*. Major clades and a revised classification of *Magnolia* and Magnoliaceae based on whole plastid genome sequences via genome skimming. *J. Syst. Evol.* **58**, 673–695, https://doi.org/10.1111/jse.12588 (2020).
2. Shankar, U. Primitive angiosperms in the Indian Subcontinent: Taxonomic diversity and geographical distribution of Magnoliaceae Juss. (APG IV). *Pleione* **14**, 137–151, https://doi.org/10.3767/000651904X486214 (2020).
3. Figlar, R. B. & Nooteboom, H. P. Notes on Magnoliaceae IV. *Blumea-Biodiversity, Evolution and Biogeography of Plants* **49**, 87–100, https://doi.org/10.3767/000651904X486214 (2004).
4. Xie, H. H. *et al*. Diversity patterns and conservation gaps of Magnoliaceae species in China. *Sci. Total Environ.* **813**, 152665, https://doi.org/10.1016/j.scitotenv.2021.152665 (2022).
5. Dong, S. S. *et al*. Plastid and nuclear phylogenomic incongruences and biogeographic implications of *Magnolia* s.l. (Magnoliaceae). *J. Syst. Evol.* **60**, 1–15, https://doi.org/10.1111/jse.12727 (2022).
6. Lee, Y. J. *et al*. Therapeutic applications of compounds in the *Magnolia* family. *Pharm. Thera.* **130**, 157–176, https://doi.org/10.1016/j.pharmthera.2011.01.010 (2011).
7. Xu, J. W. & Xu, H. Magnolol: Chemistry and biology. *Ind. Crop. Prod.* **205**, 117493, https://doi.org/10.1016/j.indcrop.2023.117493 (2023).
8. Rauf, A. *et al*. Honokiol: A review of its pharmacological potential and therapeutic insights. *Phytomedicine* **90**, 153647, https://doi.org/10.1016/j.phymed.2021.153647 (2021).
9. Wang, Y. L. *et al*. *Magnolia sinostellata* and relatives (Magnoliaceae). *Phytotaxa* **154**, 47–58, https://doi.org/10.11646/phytotaxa.154.1.3 (2013).
10. Rivers, M., Beech, E., Murphy, L. & Oldfield, S. *The Red List of Magnoliaceae- revised and extended*. (Botanic Gardens Conservation International, 2016).
11. Liu, D., Chu, L. & Yang, Y. Genetic diversity of rare and endangered plant *Magnolia amoena*. *Chin. J. Appl. Ecol.* **15**, 1139–1142, https://www.cjae.net/CN/Y2004/V/I7/1139 (2004).
12. Nanjing University of Chinese Medicine. *Zhong Yao Da Ci Dian* 2nd edn Vol. 1 (Shanghai Scientific & Technical Publishers, 2006).
13. China Expert Workshop. *Magnolia amoena*. *The IUCN Red List of Threatened Species 2014*: e.T32423A2818554. https://doi.org/10.2305/IUCN.UK.2014-1.RLTS.T32423A2818554.en (2014)
14. Ministry of Ecology and Environment of the People's Republic of China & Chinese Academy of Sciences. *China Biodiversity Red List: Higher Plants Volume* (2020). 577 (2023).
15. Ma, H. F., Sima, Y. K. & Xiang, W. Composition analysis of the volatile chemicals in *Magnolia amoena*. *Yunnan For. Sci. Technol.* **4**, 65–67, https://doi.org/10.16473/j.cnki.xblykx1972.2001.04.014 (2001).
16. Yu, Z. J., Yi, G. M., Shan, W. & Du, J. Introduction and domestication of *Magnolia amoena*. *Pract. For. Technol.* **3**, 8–9, https://doi.org/10.13456/j.cnki.lykt.2008.01.014 (2008).
17. Sun, Q. M., Dou, J., Liu, X. J. & Liu, X. W. Investigation on the Magnoliaceae Plant Resources in East China. *J. Anhui Agri. Sci.* **36**, 14956–14957+15018, https://doi.org/10.13989/j.cnki.0517-6611.2008.34.103 (2008).
18. Wang, C. Y., Liu, X. L. & Yu, C. The genetic diversity analysis of subgenus *Yulania* and its related species. *Mol. Plant Breed.* **18**, 3786–3796, https://doi.org/10.13271/j.mpb.018.003786 (2020).
19. Doyle, J. 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. F., Zhou, Y. Q., Chen, Y. R. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
21. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
22. 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).
23. Cheng, H. Y., Concepcion, G. T., Feng, X. W., Zhang, H. W. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
24. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
25. 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).
26. 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).

27. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *P. Nat. A. Sci. USA* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
28. Xu, Z. & 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).
29. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 1–14, https://doi.org/10.1186/1471-2105-9-18 (2008).
30. Abrusán, G., Grundmann, N., DeMester, L. & Makalowski, W. TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329–1330, https://doi.org/10.1093/bioinformatics/btp084 (2009).
31. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915, https://doi.org/10.1038/s41587-019-0201-4 (2019).
32. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110, https://doi.org/10.1093/bioinformatics/btw152 (2016).
33. 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).
34. Kuo, R. I. *et al*. Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics* **21**, 1–22, https://doi.org/10.1186/s12864-020-07123-7 (2020).
35. Jiang, S. R. *et al*. A high-quality haplotype genome of *Michelia alba* DC reveals differences in methylation patterns and flower characteristics. *Mol. Hort.* **4**, 23, https://doi.org/10.1186/s43897-024-00098-z (2024).
36. Dong, S. S. *et al*. The genome of *Magnolia biondii* Pamp. provides insights into the evolution of Magnoliales and biosynthesis of terpenoids. *Hortic. Res.* **8**, https://doi.org/10.1038/s41438-021-00471-9 (2021).
37. Zhou, L. J. *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, https://doi.org/10.3389/fpls.2023.1108701 (2023).
38. Yin, Y. P. *et al*. The chromosome-scale genome of *Magnolia officinalis* provides insight into the evolutionary position of magnoliids. *Iscience* **24**, https://doi.org/10.1016/j.isci.2021.102997 (2021).
39. Cai, L. *et al*. The chromosome-scale genome of *Magnolia sinica* (Magnoliaceae) provides insights into the conservation of plant species with extremely small populations (PSESP). *GigaScience* **13**, giad110, https://doi.org/10.1093/gigascience/giad110 (2024).
40. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014, https://doi.org/10.1093/bioinformatics/btad014 (2023).
41. 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).
42. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94, https://doi.org/10.1006/jmbi.1997.0951 (1997).
43. Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679, https://doi.org/10.1093/bioinformatics/btm009 (2007).
44. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 1–14, https://doi.org/10.1186/1471-2105-12-491 (2011).
45. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, 1–22, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
46. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699–2699, https://doi.org/10.1093/nar/gky092 (2018).
47. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30, https://doi.org/10.1093/nar/28.1.27 (2000).
48. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29, https://doi.org/10.1038/75556 (2000).
49. 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).
50. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Informatics 2009: Genome Informatics Series Vol.* **23**, 205–211, https://doi.org/10.1142/9781848165632_0019 (World Scientific, 2009).
51. Finn, R. D. *et al*. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230, https://doi.org/10.1093/nar/gkt1223 (2014).
52. Aramaki, T. *et al*. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252, https://doi.org/10.1093/bioinformatics/btz859 (2020).
53. 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).
54. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108, https://doi.org/10.1093/nar/gkm160 (2007).
55. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935, https://doi.org/10.1093/bioinformatics/btt509 (2013).
56. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14, https://doi.org/10.1186/s13059-019-1832-y (2019).
57. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797, https://doi.org/10.1093/nar/gkh340 (2004).
58. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973, https://doi.org/10.1093/bioinformatics/btp348 (2009).
59. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313, https://doi.org/10.1093/bioinformatics/btu033 (2014).
60. Yang, Z. H. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591, https://doi.org/10.1093/molbev/msm088 (2007).
61. Liu, Y. *Genbank* http://identifiers.org/insdc:JBNYVC000000000 (2025).
62. Liu, Y. Genome assembly and annotation of *Magnolia amoena*. *figshare. Dataset.* https://doi.org/10.6084/m9.figshare.29095601.v2 (2025).
63. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33610905 (2025).
64. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33610904 (2025).
65. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33610903 (2025).
66. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33610898 (2025).
67. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33610902 (2025).
68. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33610901 (2025).
69. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33610900 (2025).
70. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33610899 (2025).
71. Manni, M. *et al*. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).

Acknowledgements
This work was supported by the Jiangsu Forestry Science and Technology Innovation and Promotion Project (LYKJ[2025]06), Special Fund Project for Forestry Development in Jiangsu Province, and the Jiangsu Key Laboratory for Conservation and Utilization of Plant Resources (JSPKLB202401).

Author contributions
R.S.L. conceived and designed the research; Y.L. and X.J.L. collected the samples; Y.L., K.H., M.H.L. and Z.J.S. conducted the data analysis; X.Q.S. provided valuable suggestions throughout the study; Y.L., X.J.L. and R.S.L. wrote the manuscript. 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/s41597-026-06973-2.

Correspondence and requests for materials should be addressed to R.-S.L.