Dong et al. *Horticulture Research* (2021)8:38
https://doi.org/10.1038/s41438-021-00471-9

Horticulture Research
www.nature.com/hortres

ARTICLE Open Access

The genome of *Magnolia biondii* Pamp. provides insights into the evolution of Magnoliales and biosynthesis of terpenoids

Shanshan Dong[1], Min Liu[2], Yang Liu[1,2], Fei Chen[3], Ting Yang[2], Lu Chen[1], Xingtan Zhang[4], Xing Guo[2], Dongming Fang[2], Linzhou Li[2], Tian Deng[1], Zhangxiu Yao[1], Xiaoan Lang[1], Yiqing Gong[1], Ernest Wu[5], Yaling Wang[6], Yamei Shen[7], Xun Gong[8], Huan Liu[2,9] and Shouzhou Zhang[1]

Abstract

Magnolia biondii Pamp. (Magnoliaceae, magnoliids) is a phylogenetically, economically, and medicinally important ornamental tree species widely grown and cultivated in the north-temperate regions of China. Determining the genome sequence of *M. biondii* would help resolve the phylogenetic uncertainty of magnoliids and improve the understanding of individual trait evolution within the *Magnolia* genus. We assembled a chromosome-level reference genome of *M. biondii* using ~67, ~175, and ~154 Gb of raw DNA sequences generated via Pacific Biosciences single-molecule real-time sequencing, 10X Genomics Chromium, and Hi-C scaffolding strategies, respectively. The final genome assembly was ~2.22 Gb, with a contig N50 value of 269.11 kb and a BUSCO complete gene percentage of 91.90%. Approximately 89.17% of the genome was organized into 19 chromosomes, resulting in a scaffold N50 of 92.86 Mb. The genome contained 47,547 protein-coding genes, accounting for 23.47% of the genome length, whereas 66.48% of the genome length consisted of repetitive elements. We confirmed a WGD event that occurred very close to the time of the split between the Magnoliales and Laurales. Functional enrichment of the *Magnolia*-specific and expanded gene families highlighted genes involved in the biosynthesis of secondary metabolites, plant–pathogen interactions, and responses to stimuli, which may improve the ecological fitness and biological adaptability of the lineage. Phylogenomic analyses revealed a sister relationship of magnoliids and Chloranthaceae, which are sister to a clade comprising monocots and eudicots. The genome sequence of *M. biondii* could lead to trait improvement, germplasm conservation, and evolutionary studies on the rapid radiation of early angiosperms.

Introduction

The family Magnoliaceae Juss., with more than 300 species[1] worldwide, comprises two genera, *Liriodendron* L., which includes only two species, and *Magnolia* L., which includes the others[2]. Approximately 80% of all extant Magnoliaceae species are distributed in the temperate and tropical regions of Southeast Asia, and the others are distributed in the Americas, from temperate southeast North America through Central America to Brazil[3], forming disjunct distribution patterns[4].

Magnolia is a member of the magnoliids, which constitutes one of the earliest assemblages of flowering plants (angiosperms) and occupies a pivotal position in the phylogeny of angiosperms[5]. After the early divergence of angiosperms (Amborellales, Austrobaileyales, and Nymphaeales), the rapid radiation of five lineages of mesangiosperms (magnoliids, Chloranthaceae, *Ceratophyllum*, monocots, and eudicots) occurred within a very short time frame of < 5MYA[6], leading to unresolved/

Correspondence: Huan Liu (liuhuan@genomics.cn) or Shouzhou Zhang (shouzhouz@126.com)
[1]Laboratory of Southern Subtropical Plant Diversity, Fairy Lake Botanical Garden, Shenzhen & Chinese Academy of Sciences, Shenzhen 518004, China
[2]State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China
Full list of author information is available at the end of the article
These authors contributed equally: Shanshan Dong, Min Liu

© The Author(s) 2021
Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, 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 license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit http://creativecommons.org/licenses/by/4.0/.

Downloaded from https://academic.oup.com/hr/article/doi/10.1038/s41438-021-00471-9/6446648 by guest on 05 August 2026

controversial phylogenetic relationships among some lineages of mesangiosperms[5]. To date, of the 323 genome sequences available for angiosperm species[7], mostly those of plants with agronomic value, genomes are available for only five magnoliids: black pepper[8], avocado[9], soursop[10], stout camphor tree[11], and *Liriodendron chinense*[12]. Phylogenomic analyses based on these genomic data have led to controversial taxonomic placements of magnoliids. Specifically, magnoliids are resolved as sister to eudicots, with relatively strong support[11], which is consistent with the results of a phylotranscriptomic analysis of 92 streptophytes[13] and 20 representative angiosperms[14]. Alternatively, magnoliids are resolved as sister to eudicots and monocots, with weak support[8–10,12], which is in agreement with the large-scale plastome phylogenomic analysis of land plants, Viridiplantae, and angiosperms[15–17]. As phylogenetic inferences rely heavily on the sampling of lineages and genes, as well as analytical methods[5], these controversial taxonomic placements of magnoliids relative to monocots and eudicots need to be further examined with more genomic data from magnoliids.

Magnolia species are usually cross-pollinated with precocious pistils, resulting in a very short pollination period. Many species of this genus have relatively low rates of pollen and seed germination[18], as well as low production of fruits and seeds, which leads to difficult regeneration of natural populations in nature[19–21]. Exacerbated by native habitat loss due to logging and agriculture, approximately 48% of all *Magnolia* species are threatened in the wild[1]. Conservation of the germplasm resources of *Magnolia* has many economic and ecological values. Most of the *Magnolia* species are excellent ornamental tree species[22] due to their attractive flowers with sweet fragrances and erect tree shape with graceful foliage, as is the case for *M. denudata*, *M. liliiflora*, and *M. grandiflora*. *Magnolia* species also contain a rich array of terpenoids in their flowers[23] and have considerable varieties of phenolic compounds in their bark[24]. Many *Magnolia* species, such as *M. officinalis*, *M. biondii*, *M. denudata*, and *M. sprengeri*, are cultivated for medicinal and cosmetic purposes[25]. However, the lack of a high-quality reference genome assembly for *Magnolia* hinders current conservation and utilization efforts. Genome sequences of *Magnolia* could greatly aid molecular breeding, germplasm conservation, and scientific research of the genus.

One *Magnolia* species that is cultivated for ornamental, pharmaceutical, and timber purposes is *Magnolia biondii* Pamp. (Magnoliaceae, magnoliids). *M. biondii* is a deciduous tree species widely grown and cultivated in the north-temperate regions of China. Its flowers are showy and fragrant, and essential oils can be extracted from them. Chemical extracts of the flower buds are used for local stimulation and anesthesia, anti-inflammatory, antimicrobial, analgesic, blood pressure-decreasing, and anti-allergic effects[25]. Modern phytochemical studies have characterized the chemical constituents of volatile oils[26], lignin[27], and alkaloids[28] from different parts of *M. biondii* plants. Volatile oils contain a rich array of terpenoids, among which the main ingredients include 1,8-cineole, β-pinene, α-terpineol, and camphor[25]. These terpenoids are synthesized by terpene synthase (TPS), which belongs to the TPS gene family. In this study, we sequenced and assembled the reference genome of *M. biondii* using PacBio long read, 10X Genomics Chromium, and Hi-C scaffolding strategies. The ~2.22 Gb genome sequence of *M. biondii* represents the largest genome assembled to date for early-diverging magnoliids. This genome will support future studies on floral evolution and the biosynthesis of the primary and secondary metabolites unique to the species and will be an essential resource for understanding rapid changes that took place throughout the phylogenetic backbone of angiosperms. Finally, this information could be used to further improve genome-assisted cultivation and conservation efforts of *Magnolia*.

Materials and methods

Plant materials, DNA extractions, and sequencing

Fresh leaves and flower materials were collected from a 21-year-old *M. biondii* tree (a cultivated variety) at three developmental stages planted in the Xi'an Botanical Garden, Xi'an, China. The specimen (voucher number: Zhang 201801M) has been deposited in the Herbarium of Fairy Lake Botanical Garden, Shenzhen, China. Total genomic DNA was extracted from fresh young leaves of *M. biondii* using the modified cetyltrimethylammonium bromide (CTAB) method[29]. The quality and quantity of the DNA samples were evaluated using a NanoDrop™ One UV-Vis spectrophotometer (Thermo Fisher Scientific, USA) and a Qubit® 3.0 Fluorometer (Invitrogen Ltd, Paisley, UK), respectively. Three different approaches were subsequently used for genomic DNA sequencing at BGI-Shenzhen (BGI Co., Ltd., Shenzhen, Guangdong, China) (Supplementary Table S1). First, high-molecular-weight genomic DNA was prepared for 10X Genomics libraries with insert sizes of 350–500 bp according to the manufacturer's protocol (Chromium Genome Chip Kit v1, PN-120229, 10X Genomics, Pleasanton, USA). The resulting barcoded library was sequenced on a BGISEQ-500 platform to generate 150-bp reads. Duplicate reads, reads with ≥20% low-quality bases, or reads with ≥5% ambiguous ("N") bases were filtered using SOAPnuke v. 1.5.6[30] with the parameters "-l 10 -q 0.1 -n 0.01 -Q 2 -d –misMatch 1 –matchRatio 0.4 and -t 30,20,30,20". Second, single-molecule real-time (SMRT) Pacific Biosciences (PacBio) libraries were constructed using the PacBio 20-kb protocol (https://www.pacb.com/) and sequenced on a PacBio RS-II instrument. Third, a Hi-C

Downloaded from https://academic.oup.com/hr/article/doi/10.1038/s41438-021-00471-9/6446648 by guest on 05 August 2026

library was generated using DpnII restriction enzymes following in situ ligation protocols[31]. The DpnII-digested chromatin was end-labeled with biotin-14-dATP (Thermo Fisher Scientific, Waltham, MA, USA) and used for in situ DNA ligation. The DNA was extracted, purified, and then sheared using Covaris S2 (Covaris, Woburn, MA, USA). After A-tailing, pull down, and adapter ligation, the DNA library was sequenced on a BGISEQ-500 instrument to generate 100-bp reads.

RNA extraction and sequencing

Young leaves (LEAF), opening flowers (FLOWER), and flower buds (BUDA and BUDB) at two developmental stages (pre-meiosis and post-meiosis) were collected from the same individual tree planted in the Xi'an Botanical Garden. Total RNA was extracted using an E.Z.N.A.® Total RNA Kit I (Omega Bio-Tek), after which quality control was performed using a NanoDrop™ One UV-Vis spectrophotometer (Thermo Fisher Scientific, USA) and a Qubit® 3.0 Fluorometer (Thermo Fisher Scientific, USA). All RNA samples with integrity values close to 10 were selected for cDNA library construction and next-generation sequencing. A cDNA library was prepared using a TruSeq RNA Sample Preparation Kit v2 (Illumina, San Diego, CA, USA) followed by paired-end (150 bp) sequencing on the HiSeq 2000 platform (Illumina, Inc., CA, USA) at Majorbio (Majorbio Co., Ltd., Shanghai, China). The newly generated raw sequence reads were trimmed and filtered for adapters, low-quality reads, undersized inserts, and duplicate reads using Trimmomatic v. 0.38[32].

Genome size estimation

We used 17 k-mer counts[33] of high-quality reads from small-insert 10X Genomics libraries to evaluate the genome size and level of heterozygosity. First, k-mer frequency distribution analyses were performed according to the methods of Chang et al.[34] to determine the occurrence of k-mers based on clean paired-end 10X Genomics data. GCE[35] was then used to estimate the general characteristics of the genome, including the total genome size, repeat proportions, and level of heterozygosity (Supplementary Table S2 and Supplementary Fig. S1).

De novo genome assembly and chromosome construction

De novo assembly was performed with five different genome assemblers: Canu v. 0.1[36], Miniasm v. 0.3[37], Wtdbg v. 1.1.006 (https://github.com/ruanjue/wtdbg), Flye v. 2.3.3[38], and SMARTdenovo 1.0.0 (https://github.com/ruanjue/smartdenovo) with/without a priori Canu correction with default parameters. Based on the size of the assembled genome, the total number of assembled contigs, the N50 contig length, the maximum length of the contigs, and the completeness of the genome assembly as assessed via Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis[39] (1375 single-copy orthologs of the Embryophyta odb10 database) with a BLAST e-value cutoff of 1e−5, the genome assembly from the Miniasm assembler was selected for further polishing and scaffolding (Supplementary Table S3). The consensus sequences of the assembly were further improved using all the PacBio reads for three rounds of iterative error correction using Racon software v. 1.2.1[40] with the default parameters, and the resultant consensus sequences were further polished using Pilon v. 1.22[41] (parameters: –fix bases, amb –vcf –threads 32) with one round of error correction using all the clean paired-end 10X Genomics reads. The Hi-C reads were subjected to quality control measures (Supplementary Table S2) and then mapped to the contig assembly of *M. biondii* using Juicer[42], with default parameters. A candidate chromosome-length assembly was then generated automatically using the 3D-DNA pipeline[43] (parameters: -m haploid -s 4 -c 19 -j 15) to correct misjoins, order, and orientation and to organize the contigs from the draft chromosome assembly. Manual check and refinement of the draft assembly was carried out via Juicebox Assembly Tools[44] (Table 1, Supplementary Fig. S2, and https://doi.org/10.5061/dryad.s4mw6m947).

Table 1 Final genome assembly based on the assembled contigs from Miniasm

	PacBio Assembly (polished)	Hi-C Assembly
Total scaffold length (Gb)		2.232
Number of scaffolds		9510
Scaffold N50 (Mb)		92.86
Scaffold N90 (Mb)		19.29
Max scaffold length (Mb)		168.50
Total contig length (Gb)	2.22	
Number of contigs	15,615	
Contig N50 (kb)	269.114	
Contig N90 (kb)	60.09	
Max contig length (kb)	2,134.98	
Complete BUSCOs	91.90%	88.50%
Complete and single-copy BUSCOs	87.00%	85.20%
Complete and duplicate BUSCOs	4.90%	3.30%
Fragmented BUSCOs	3.00%	4.40%

Downloaded from https://academic.oup.com/hr/article/doi/10.1038/s41438-021-00471-9/6446648 by guest on 05 August 2026

Genome evaluation

The completeness of the genome assembly of *M. biondii* was evaluated by comparisons with the DNA and RNA mapping results, comparisons with the transcript unigene mapping results, and BUSCO analysis[39]. First, all the paired-end reads from the 10X Genomics and Hi-C data were mapped against the final assembly of *M. biondii* using BWA-MEM v. 0.7.10[45]. The RNA-seq reads from four different tissues were also mapped back to the genome assembly using TopHat v. 2.1.0[46]. Second, unigenes were generated from the transcript data of *M. biondii* using Bridger software[47] with the parameters "–kmer length 25 – min kmer coverage 2" and then aligned to the scaffold assembly using the Basic Local Alignment Search Tool (BLAST)-like alignment tool BLAT[48]. Third, BUSCO analysis[39] of the final scaffold assembly was also performed to evaluate the genome completeness of the reference genome of *M. biondii*.

Repeat annotations

Transposable elements (TEs) were identified by a combination of homology-based and de novo approaches. Briefly, the genome assembly was aligned to the known repeat database Repbase v. 21.01[49] using RepeatMasker v. 4.0.5[50] and Repeat-ProteinMask[50] at both the DNA and protein levels for homology-based TE characterization. For the de novo approach, RepeatModeler 2.0[51] and LTR Finder v. 1.0.6[52] were used to construct a de novo repeat library using the *M. biondii* assembly. TEs in the genome were then identified and annotated by RepeatMasker v. 4.0.5[50], and tandem repeats were annotated using TRF v. 4.04[53] (Supplementary Table S4, and https://doi.org/10.5061/dryad.s4mw6m947).

Gene predictions

Protein-coding genes were predicted by using the MAKER-P pipeline v. 2.31[54] based on de novo predictions, homology search results, and transcriptome evidence. For de novo gene prediction, GeneMark-ES v. 4.32[55] was first used for self-training, with the default parameters. Second, the alternative spliced transcripts obtained by a genome-guided approach by using Trinity with the parameters "–full_cleanup –jaccard_clip –no_version_check –genome_guided_max_intron 100000 –min_contig_length 200" were mapped to the genome by using PASA v. 2.3.3 with default parameters. The complete gene models (https://doi.org/10.5061/dryad.s4mw6m947) were then selected and used for training Augustus[56] and SNAP[57]. The models were used to predict coding genes on the repeat-masked *M. biondii* genome. For homologous comparisons, protein sequences from *Arabidopsis thaliana*, *Oryza sativa*, *Amborella trichopoda*, and two related species (*Cinnamomum kanehirae* and *Liriodendron chinense*) were provided as protein evidence.

For RNA evidence, a completely de novo approach was chosen. The clean RNA-seq reads were assembled into inchworm contigs using Trinity v. 2.0.6[58] with the parameters "–min_contig_length 100 –min_kmer_cov 2 –inchworm_cpu 10 –bfly_opts "-V 5 –edge-thr = 0.05 –stderr" –group_pairs_distance 200 –no_run_chrysalis" and then provided to MAKER-P as expressed sequence tags (Supplementary Fig. S3, and https://doi.org/10.5061/dryad.s4mw6m947). After two rounds of MAKER-P, a consensus gene set was obtained. tRNAs were identified using tRNAscan-SE v. 1.3.1[59]. snRNAs and miRNAs were detected by searching the reference sequence against the content of the Rfam database[60] using BLAST[61] and rRNAs were detected by alignment with BLASTN[61] against known plant rRNA sequences[62] (Supplementary Table S5). We also mapped the gene density, GC content, *Gypsy* density, and *Copia* density onto the individual chromosomes using the Circos tool (http://www.circos.ca) (Fig. 1).

Functional annotation of protein-coding genes

Functional annotation of protein-coding genes was performed by searching the predicted amino acid sequences of *M. biondii* against the contents of public databases based on sequence identity and domain conservation. The sequences of protein-coding genes were previously searched within several protein sequence databases, including the Kyoto Encyclopedia of Genes and Genomes (KEGG)[63], National Center for Biotechnology Information (NCBI) nonredundant (NR), Clusters of Orthologous Groups (COG)[64], SwissProt[65], and TrEMBL[65] databases, for best matches using BLASTP with an e-value cutoff of $1e-5$. InterProScan 5.0[66] was then used to characterize protein domains and motifs based on data acquired from Pfam[67], SMART[68], PANTHER[69], PRINTS[70], and ProDom[71] (Supplementary Table S6).

Gene family construction

Protein and nucleotide sequences from *M. biondii* and five other angiosperms (*A. trichopoda*, *A. thaliana*, *C. kanehirae*, *L. chinense*, *Vitis vinifera*) were used to construct gene families using OrthoFinder[72] (https://github.com/davidemms/OrthoFinder) based on an all-versus-all BLASTP alignment with an e-value cutoff of $1e-5$. Potential gene pathways were obtained via gene mapping against the KEGG database, and Gene Ontology (GO) terms were extracted from the corresponding InterProScan or Pfam results (Fig. 2).

Phylogenomic reconstruction and gene family evolution

To understand the relationships between the *M. biondii* gene families and those of other plant species and the phylogenetic placements of magnoliids among angiosperms, we performed a phylogenetic comparison of genes

Downloaded from https://academic.oup.com/hr/article/doi/10.1038/s41438-021-00471-9/6446648 by guest on 05 August 2026

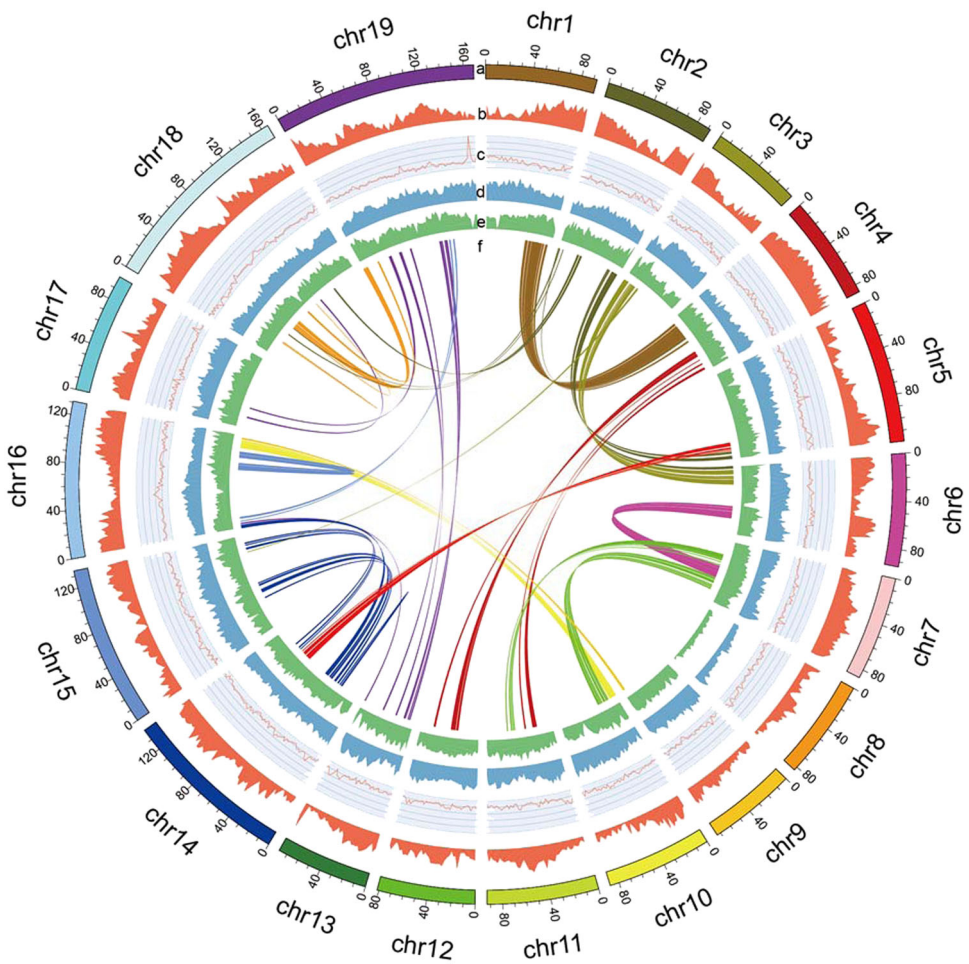

Fig. 1 Reference genome assembly of 19 pseudochromosomes. a Assembled pseudochromosomes, **b** gene density, **c**. GC content, **d** *Gypsy* density, **e** *Copia* density, and **f** chromosome synteny (from outside to inside)

among different species along a 20-seed plant phylogeny reconstructed with a concatenated amino acid dataset derived from 109 single-copy nuclear genes. Putative orthologous genes were constructed from 18 angiosperms (two eudicots, two monocots, two Chloranthaceae species, eight magnoliid species, two *Illicium* species, *A. trichopoda*, *Nymphaea* sp.) and the gymnosperm outgroup *Picea abies* (Supplementary Table S7) using OrthoFinder[72] and compared with protein-coding genes from the genome assembly of *M. biondii*. The total one-to-one orthologous gene sets were identified and extracted for alignment using MAFFT v. 5.0[73], further trimmed using GBlocks 0.91b[74], and concatenated by Geneious 10.0.2 (www.geneious.com). The concatenated amino acid dataset from 109 single-copy nuclear genes (each with >85% taxon occurrences) was analyzed using PartitionFinder[75], with an initial partitioning strategy for each gene for the optimal data partitioning scheme and associated substitution models, resulting in 18 partitions. The concatenated amino acid dataset was then analyzed using the maximum likelihood (ML) method with RAxML-VI-HPC v. 2.2.0[76] to determine the most reasonable tree. Nonparametric bootstrap analyses were performed by PROTGAMMAAUTO approximation for 500 pseudoreplicates (Fig. 3).

The best ML tree was used as a starting tree to estimate species divergence time using MCMC Tree, implemented in PAML v. 4[77]. Two node calibrations were defined by the TimeTree web service (http://www.timetree.org/), including the split between *Liriodendron* and *Magnolia* (34–77 MYA) and the split between angiosperms and gymnosperms (168–194 MYA). The orthologous gene clusters inferred from the OrthoFinder[72] analysis and phylogenetic tree topology constructed using RAxML-VI-HPC v. 2.2.0[76] were inputted into CAFE v. 4.2[78] to determine whether significant expansion or contraction occurred in each gene family across species.

Downloaded from https://academic.oup.com/hr/article/doi/10.1038/s41438-021-00471-9/6446648 by guest on 05 August 2026

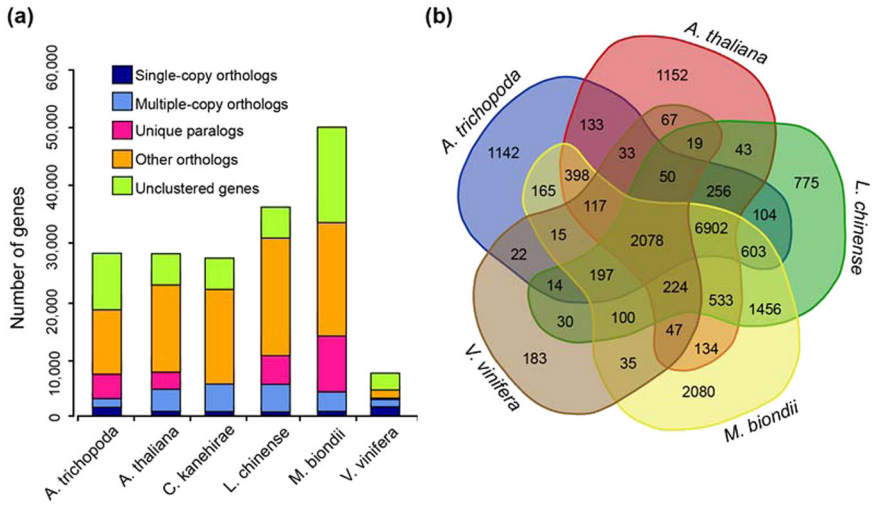

Fig. 2 Comparative analysis of the *M. biondii* genome. a Number of genes in various plant species, showing a high gene number for *M. biondii* compared to a model species (*Arabidopsis thaliana*) and other species (*Amborella trichopoda*, *Cinnamomum kanehirae*, *Liriodendron chinense*, and *Vitis vinifera*). **b** Venn diagram showing overlaps of gene families between *M. biondii*, *L. chinense*, *A. trichopoda*, *A. thaliana*, and *V. vinifera*

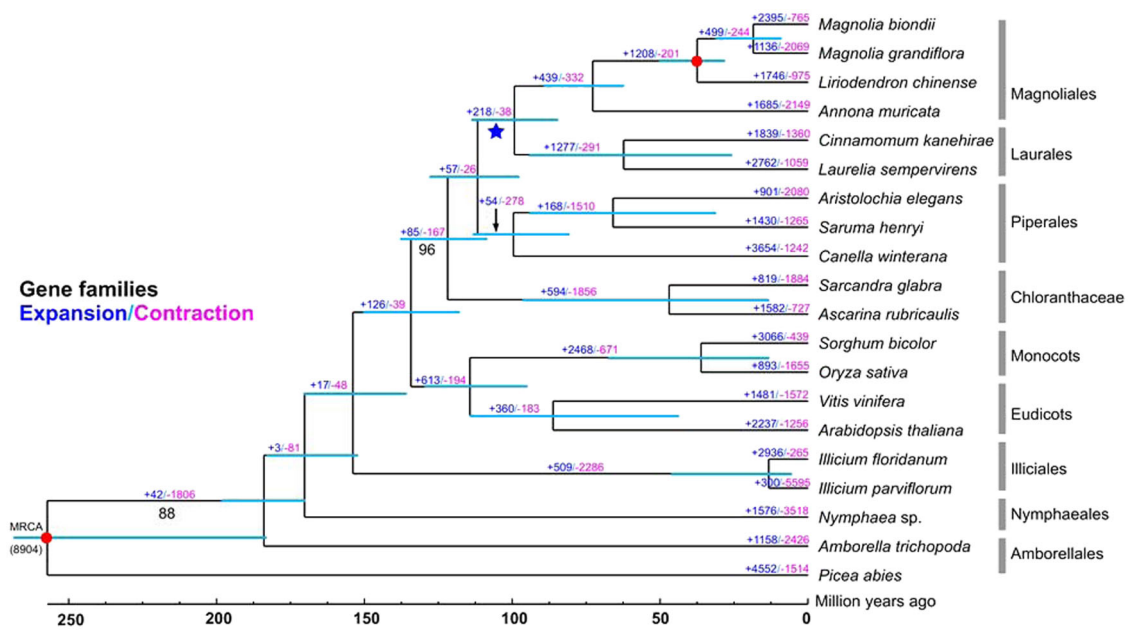

Fig. 3 Phylogenetic tree and number of gene families that have expanded or contracted among 20 plant species. Estimated divergence time confidence intervals are shown at each internal node as teal bars. Calibrated nodes are indicated by red dots. The WGD shared between the Magnoliales and Laurales is indicated with a blue star. All the branches are maximally supported by maximum likelihood analysis unless otherwise indicated below the branches

Analyses of genome synteny and whole-genome duplication (WGD)

To investigate the source of the large number of predicted protein-coding genes (47,547) in *M. biondii*, WGD events were analyzed by making use of the genome sequences of *M. biondii*. Given that the grape genome has one well-established whole-genome triplication and the cofamilial *L. chinense* has one reported WGD event[12], the protein-coding genes (of CDSs and their translated protein sequences) of *M. biondii*, *L. chinense*, and grape were used to perform synteny searches with MCscanX[79] (python version), with at least five gene pairs required per syntenic block. The resultant dot plots were examined to predict the

Downloaded from https://academic.oup.com/hr/article/doi/10.1038/s41438-021-00471-9/6446648 by guest on 05 August 2026

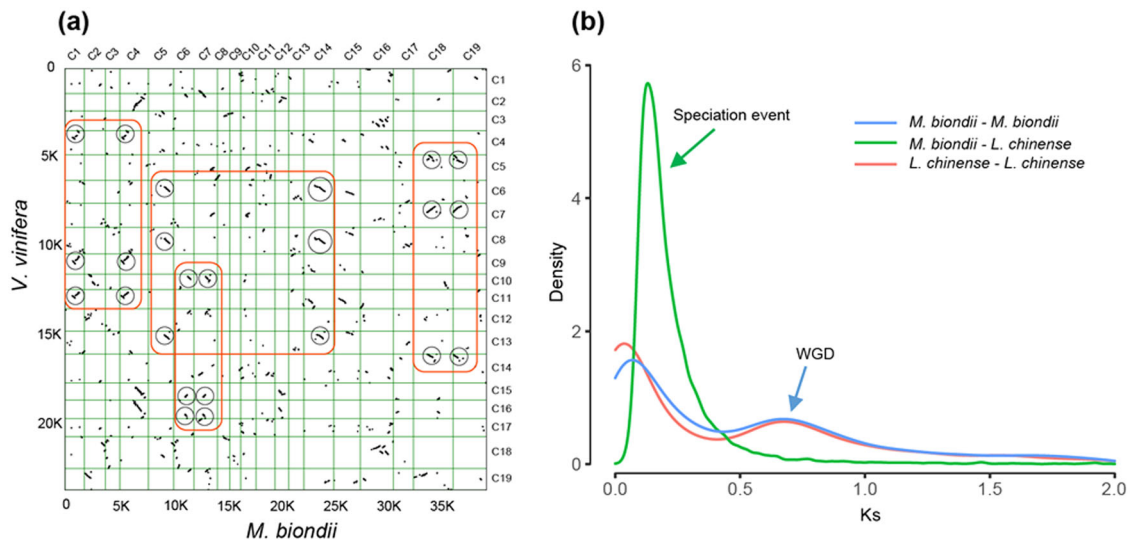

Fig. 4 Evidence for WGD events in *M. biondii*. a Comparison of *M. biondii* and grape genomes. Dot plots of orthologs showing a 2–3 chromosomal relationship between the *M. biondii* genome and grape genome. **b** Synonymous substitution rate (Ks) distributions of paralogs and orthologs retrieved from gene family clustering results from OrthoMCL for *M. biondii* and *Liriodendron chinense*

paleoploidy level of *M. biondii* compared with that of the other angiosperms by determining the syntenic depth in each genomic region (Supplementary Figs. S4 and S5). For synonymous substitution rate (Ks) distribution analysis, the gene family clustering results of OrthoMCL[80,81] were sorted, and gene families with only one member of both *L. chinense* and *M. biondii* and gene families with two members of either species were extracted (Supplementary Fig. S4). PAML[77] software was then used to calculate the Ks values for the gene pairs (Fig. 4b).

Identification of TPS genes and expression analysis

We selected two species (*A. trichopoda* and *A. thaliana*) to perform a comparative TPS gene family analysis together with *M. biondii*. Previously annotated TPS genes of the two species were retrieved from the data deposited by Chaw et al.[11]. Two Pfam domains, PF03936 and PF01397, were used as queries for searching against the *M. biondii* proteome (of the contig version) using HMMER v. 3.0, with an e-value cutoff of 1e−5[82]. The putative protein sequences of 102 TPS genes were aligned using MAFFT v. 5[73] and manually adjusted using MEGA v. 4[83]. A phylogenetic tree was constructed using IQ-TREE[84] with 1000 bootstrap replicates (Fig. 5).

Bowtie2[85] was used to map the RNA-seq reads to the protein-coding sequences of the gene set. eXpress[86] software was used to calculate the expression results of different tissues, and edgeR[87] was used for analysis of differentially expressed genes. Parameter thresholds including an FDR of <0.001 and a $\log_2FC > 2$ or a $\log_2FC < -2$ were applied to identify differentially expressed TPS genes in *M. biondii* (https://doi.org/10.5061/dryad.s4mw6m947).

Data access

The genome assembly, annotations, and other supporting data are available in the Dryad database under the https://doi.org/10.5061/dryad.s4mw6m947. The raw sequence data have been deposited in the China National GeneBank DataBase (CNGBdb) under Accession No. CNP0000884.

Results

Sequencing summary

DNA sequencing generated 33-fold PacBio single-molecule long reads (a total of 66.78 Gb, with an average length of 10.32 kb), 80-fold 10X Genomics paired-end short reads (175.45 Gb), and Hi-C data (~153.78 Gb). Transcriptome sequencing generated 4.62, 4.60, 4.67, and 4.73 Gb of raw data for young leaves, opening flowers, and flower buds at two developmental stages (pre-meiosis and post-meiosis), respectively (Supplementary Table S1).

Determination of genome size and heterozygosity

K-mer frequency distribution analyses suggested a k-mer peak with a depth of 48 and an estimated genome size of 2.17 Gb (Supplementary Fig. S1a and Table S2). GCE[35] analysis resulted in a k-mer peak with a depth of 29 and a calculated genome size of 2.24 Gb, an estimated heterozygosity of 0.73%, and a repetitive content of 61.83% (Supplementary Fig. S1b and Table S2). The estimated genome size of *M. biondii* is the largest among all the sequenced genomes of magnoliids.

Downloaded from https://academic.oup.com/hr/article/doi/10.1038/s41438-021-00471-9/6446648 by guest on 05 August 2026

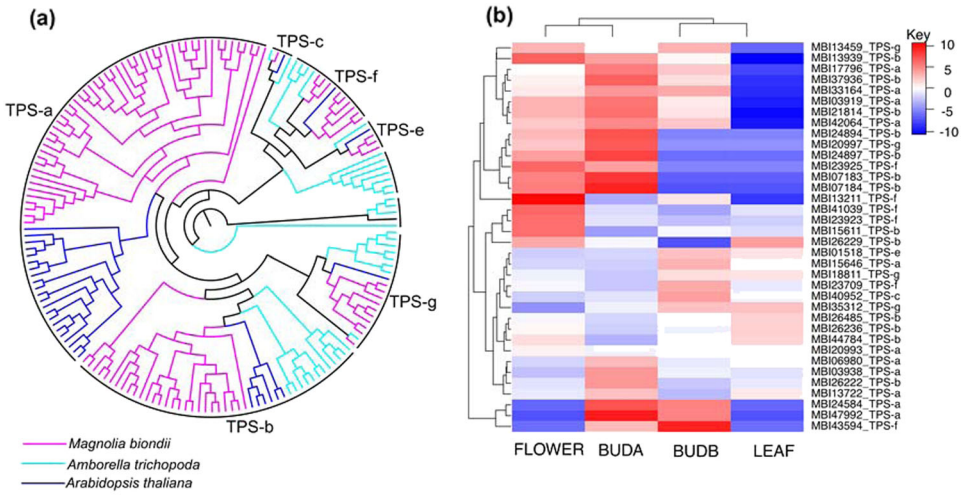

Fig. 5 TPS gene family in *M. biondii*. a Phylogenetic tree of TPS genes from *Amborella trichopoda* (38 genes), *Arabidopsis thaliana* (32 genes), and *M. biondii* (102 genes). **b** Heatmap showing the differential expression of TPS genes according to the transcriptome data from young leaves (LEAF), opening flowers (FLOWER), flower buds pre-meiosis (BUDA), and flower buds post-meiosis (BUDB)

Genome assembly and quality assessment

The selected primary assembly from Miniasm v. 0.3[37] had a genome size of 2.20 Gb across 15,713 contigs, with a contig N50 of 267.11 kb. After three rounds of error correction with PacBio long reads and one round of correction with 10X Genomics reads, we arrived at a draft contig assembly size of 2.22 Gb spanning 15,615 contigs, with a contig N50 of 269.11 kb (Table 1 and https://doi.org/10.5061/dryad.s4mw6m947). Approximately 89.17% of the contig bases were organized onto the 19 chromosomes (1.98 Gb) with ambiguous Ns representing 7,365,981 bp (accounting for 0.33% of the genome length). About 9455 contigs (0.24 Gb) were not placed (Supplementary Fig. S2). The raw scaffold assembly was further improved with PacBio long reads and 10X Genomics short reads, resulting in an assembled genome size of 2.23 Gb represented by 9510 scaffolds, with a scaffold N50 of 92.86 Mb (Table 1). Our assembled genome size of *M. biondii* is very similar to the genome size estimated according to the k-mer analysis (Supplementary Table S2).

For genome quality assessment, first, all the paired-end reads from 10X Genomics and Hi-C were mapped against the final assembly of *M. biondii*, resulting in 98.40% and 92.50% of the total mapped reads, respectively. Sequencing coverage of the 10X Genomics reads and Hi-C reads showed that more than 98.04% and 86.00% of the genome bases had a sequencing depth >10X, respectively. The RNA-seq reads from four different tissues were also mapped back to the genome assembly using TopHat v. 2.1.0[46], resulting in 93.3%, 94.4%, 92.9%, and 93.7% of the total mapped RNA-seq reads for leaves, opening flowers, and flower buds pre-meiosis and post-meiosis, respectively. Second, unigenes generated from the transcriptomic data of *M. biondii* were aligned to the scaffold assembly. The results indicated that the assemblies covered approximately 86.88% of the expressed unigenes. Third, BUSCO analysis[39] of the final scaffold assembly showed that 88.50% (85.20% complete and single-copy genes and 3.30% complete and duplicated genes) and 4.40% of the expected 1375 conserved embryophytic genes were identified as complete genes and fragmented genes, respectively. The results of the DNA/RNA read and transcriptome unigene mapping studies and BUSCO analysis suggested that the completeness of the reference genome of *M. biondii* was acceptable.

Repeat annotations

We identified 1,478,819,185 bp (66.48% of the genome length) bases of repetitive sequences in the genome assemblies of *M. biondii*. LTR elements constituted the predominant repeat type, accounting for 58.06% of the genome length (Supplementary Table S4). With respect to the two LTR superfamily elements, *Copia* and *Gypsy* elements constituted 659,463,750 and 727,531,048 bp, corresponding to 45.26% and 50.66% of the total LTR repeat length, respectively. The density of *Gypsy* elements decreased with increasing density of genes, whereas the *Copia* elements were distributed more evenly across the genome and showed no obvious patterns or relationships with the distribution of genes (Fig. 1). DNA transposons, satellites, simple repeats, and other repeats constituted 130,503,028, 5,540,573, 17,626,796, and 7,240,517 bp accounting for 5.86%, 0.24%, 0.79%, and 0.32%, respectively, of the genome length.

Gene annotation and functional annotation

The assembled genome of *M. biondii* contained 47,547 protein-coding genes, 109 miRNAs, 904 tRNAs, 1918

Downloaded from https://academic.oup.com/hr/article/doi/10.1038/s41438-021-00471-9/6446648 by guest on 05 August 2026

rRNAs, and 7426 snRNAs (Supplementary Table S5). The protein-coding genes in *M. biondii* had an average gene length of 10,980 bp, an average coding DNA sequence (CDS) length of 957 bp, and an average exon number per gene of 4.4. The various gene structure parameters were compared to those of the five selected angiosperm species: *A. trichopoda*, *A. thaliana*, *C. kanehirae*, *L. chinense*, and *Oryza sativa*. *M. biondii* had the highest predicted gene numbers and the largest average intron length (~2774 bp) among these species (Supplementary Table S5), which appears to be in agreement with the relatively large genome size of *M. biondii*. However, the relatively small median gene length (3701 bp) and intron length (525 bp) in *M. biondii* suggested that some genes with exceptionally long introns significantly increased the average gene length.

Functional annotation of protein-coding genes assigned potential functions to 39,111 protein-coding genes out of the total of 47,547 genes in the *M. biondii* genome (82.26%) (Supplementary Table S6). Among ~17.74% of the predicted genes without predicted functional annotations, some may stem from errors in the genome assembly and annotations, while others might be potential candidates for novel functions.

Gene family construction

Among a total of 15,089 gene families identified in the genome of *M. biondii*, 10,280 genes and 1928 gene families were found to be specific to *M. biondii* (Fig. 2a). The Venn diagram in Fig. 2b shows that 2078 gene families were shared among the five species *M. biondii*, *L. chinense*, *A. trichopoda*, *A. thaliana*, and *V. vinifera*.

A KEGG pathway analysis of the *M. biondii*-specific gene families revealed marked enrichment in genes involved in nucleotide metabolism, plant–pathogen interactions, and the biosynthesis of alkaloids, ubiquinone, terpenoid-quinones, phenylpropanoids, and other secondary metabolites (Supplementary Table S8). These results are consistent with the biological features of *M. biondii*, which has rich arrays of terpenoids, phenolics, and alkaloids. According to GO analysis, the *M. biondii*-specific gene families were enriched in binding, nucleic acid binding, organic cyclic compound binding, heterocyclic compound binding, and hydrolase activity (Supplementary Table S9). These specific genes associated with the biosynthesis of secondary metabolites and plant–pathogen interactions in the *M. biondii* genome assembly may play important roles in plant–pathogen resistance mechanisms[8] by stimulating beneficial interactions with other organisms[11].

Phylogenomic reconstruction

Our phylogenetic analyses based on 109 orthologous nuclear single-copy genes and 19 angiosperms plus one gymnosperm outgroup revealed a robust topology and supported the sister relationship between magnoliids and the Chloranthaceae (BPP = 96), which together formed a sister group relationship (BPP = 100) with a clade comprising monocots and eudicots. The phylogenetic tree (Fig. 3 and Supplementary Fig. S5) indicated that the Magnoliales and Laurales orders have a close genetic relationship and that they diverged ~99 MYA (84–116 MYA). The estimated divergence of the Magnoliaceae and Annonaceae in the Magnoliales clade occurred ~73 MYA (57–92 MYA), while the split of *Liriodendron* and *Magnolia* is estimated to have occurred ~38 MYA (31–50 MYA).

Gene family evolution

The orthologous gene clusters inferred from the OrthoFinder[72] analysis and phylogenetic tree topology constructed using RAxML-VI-HPC v. 2.2.0[76] were inputted into CAFE v. 4.2[78] to investigate whether significant expansion or contraction occurred in each gene family across species (Fig. 3). Among the total 15,683 gene families detected in the *M. biondii* genome, 2395 had significantly expanded ($P < 0.05$), and 765 had contracted ($P < 0.005$). KEGG pathway analysis of these expanded gene families revealed marked enrichment in genes involved in metabolic pathways, biosynthesis of secondary metabolites, plant hormone signal transduction, ABC transporters, etc. (Supplementary Table S10). By the use of GO analysis, the *M. biondii* expanded gene families were enriched in ion binding, transferase activity, metabolic processes, cellular processes, oxidoreductase activity, localization, responses to stimuli, etc. (Supplementary Table S11). The expansion of these genes, especially those associated with biosynthesis of secondary metabolites, plant hormone signal transduction and responses to stimuli, could possibly contribute to the ecological fitness and biological adaptability of the species.

Analyses of genome synteny and WGD

A total of 1738 colinear gene pairs on 147 colinear blocks were inferred within the *M. biondii* genome (Supplementary Fig. S6a). There were 13,674 colinear gene pairs from 393 colinear blocks detected between *M. biondii* and *L. chinense* (Supplementary Fig. S6b) and 10,042 colinear gene pairs from 928 colinear blocks detected between *M. biondii* and *V. vinifera* (Fig. 4a). Dot plots of longer syntenic blocks between *M. biondii* and *L. chinense* revealed a nearly 1:1 orthology ratio, indicating a similar evolutionary history of *M. biondii* to *L. chinense*. Like *Liriodendron*, *Magnolia* may have probably also experienced a WGD event[12] after the most recent common ancestor (MRCA) of angiosperms. The nearly 2:3 orthology ratio between *M. biondii* and grape confirmed this WGD event in the lineage leading to *Magnolia* (Supplementary Fig. S6b).

Downloaded from https://academic.oup.com/hr/article/doi/10.1038/s41438-021-00471-9/6446648 by guest on 05 August 2026

The Ks distribution for *M. biondii* paralogs revealed a main peak at approximately 0.70 Ks (~116 Ma) units, which appears to coincide with the Ks peak of *L. chinense* in our observations (Fig. 4b), indicating that these two lineages might have experienced a shared WGD in their common ancestor or two independent WGDs at a similar time. The results of one-vs-one ortholog comparisons between *Liriodendron* and *Magnolia* suggested the divergence of the two lineages occurred at approximately 0.15 Ks (~24.80 Ma) units, which largely postdates the potential WGD peak of 0.70 Ks units observed in either species, indicating that this WGD event should be shared by at least the two genera of Magnoliaceae.

TPS genes

Volatile oils isolated from the flower buds of *M. biondii* comprise primarily terpenoid compounds that are produced by the catalytic activity of TPS enzymes. We identified a total of 102 putative TPS genes in the genome assembly of *M. biondii*, which is comparable to that of *C. kanehirae*, with 101 genes[11]. To determine the classification of TPS proteins in *M. biondii*, we constructed a phylogenetic tree using all the TPS protein sequences from *M. biondii*, *A. thaliana*, and *A. trichopoda*. These TPS genes found in *M. biondii* could be assigned to six subfamilies: TPS-a (52 genes), TPS-b (27 genes), TPS-c (1 gene), TPS-e (3 genes), TPS-g (10 genes), and TPS-f (9 genes) (Fig. 5a). We compared the expression profiles of TPS genes in the young leaves and flowers at three different developmental stages (Fig. 5b) and identified a total of 36 TPS genes (11, 13, 1, 1, 6, and 4 genes for the subfamilies of TPS-a, TPS-b, TPS-c, TPS-e, TPS-f, and TPS-g, respectively) that were substantially expressed, among which 33 TPS genes (including 10 genes for both the TPS-a and TPS-b subfamilies) exhibited higher transcript abundance in the flowers than in the leaves (Fig. 5b), suggesting that these genes may be involved in a variety of terpenoid metabolic processes during *M. biondii* flower growth and development.

Discussion

The genome of *M. biondii* is relatively large and complex, as k-mer frequency analysis suggested an estimated genome size of 2.24 Gb, with an estimated heterozygosity of 0.73% and a repeat content of 61.83%. Compared with an estimated heterozygosity of 0.087% for the genome of, for example, *Oropetium thomaeum*[88], the heterozygosity of the *M. biondii* genome is approximately ten times higher, which probably contributed to the low contiguity of the assembly. Our DNA sequencing generated approximately 33-fold PacBio long-read data, which resulted in an assembly of 2.23 Gb spanning 15,615 contigs, with a contig N50 of 269.11 kb. The small contig N50 length might imply a fragmentary and incomplete genome assembly, which might affect the quality and precision of the Hi-C assembly. Indeed, when these contigs were organized into chromosomes using Hi-C data, approximately 6,899 contigs adding up to 1.00 Gb were disrupted by the Hi-C scaffolding processes, contributing to 0.18 Gb of genome sequence being discarded. After manual correction of the Hi-C map in Juicebox, the final scaffold assembly still showed 6911 contigs disrupted, 2358 genes disturbed, and 0.24 Gb of genome sequences not placed. BUSCO assessments showed decreased percentages of complete BUSCOs and increased percentages of fragmented BUSCOs for the scaffold assembly compared with the contig assembly (Table 1). Therefore, we used the Hi-C assembly for chromosome collinearity analysis and the contig assembly for the rest of the comparative analyses. The exceptionally large protein gene set predicted for the *M. biondii* genome might be attributed to gene fragmentation problems induced by poor genome assembly and a high content of TEs, as evidenced by the dramatically short average/median CDS length of *M. biondii* compared with that of the cofamilial *L. chinense* (Supplementary Table S5).

The chromosome-scale reference genome of *M. biondii* provided information on the gene content, repetitive elements, and genome structure of the DNA on the 19 chromosomes. Our genomic data offered valuable genetic resources for both molecular and applied research on *M. biondii* and paved the way for studies on the evolution and comparative genomics of *Magnolia* and related species. Phylogenomic analyses of 109 single-copy orthologs from 20 representative seed plant genomes with a good representation of magnoliids (three out of four orders) strongly support the sister relationship of magnoliids and Chloranthaceae, which together form a sister group relationship with a clade comprising monocots and eudicots. This placement is in agreement with the plastid topology[15,16] and the results of multilocus phylogenetic studies of angiosperms[6] but in contrast to the placement of the sister group relationship of magnoliids with eudicots revealed by the phylogenomic analysis of angiosperms (with *Cinnamomum kanehirae* as the only representative for magnoliids)[11] and phylotranscriptomic analysis of 92 streptophytes[13] and of 20 representative angiosperms[14]. Multiple factors underlie the robust angiosperm phylogeny recovered in our study: (a) we used less homoplasious amino acid data rather than nucleotide sequences (especially those of the 3rd codon positions) that are more prone to substitution saturation; (b) we used an optimal partitioning strategy with carefully selected substitution models, which is usually neglected for large concatenated datasets in phylogenomic analyses; and (c) we included a relatively good taxon sampling that included representatives from all eight major angiosperm lineages, with the exception of the Ceratophyllales, for

Downloaded from https://academic.oup.com/hr/article/doi/10.1038/s41438-021-00471-9/6446648 by guest on 05 August 2026

which no genomic resources are available. Future phylogenomic studies with improved and more balanced lineage sampling and thorough gene sampling as well as comprehensive analytical methods would provide more convincing evidence concerning the divergence order of early mesangiosperms.

The current assembly of the *M. biondii* genome improved our understanding of the timing of the WGD event in magnoliids. Our genome syntenic and Ks distribution analyses confirmed the WGD event in the Magnoliales. This WGD occurred ~116 MYA, as estimated by Chen et al.[12] as well as by our study, which is close to the split time of the Magnoliales and Laurales, as the two lineages diverged approximately 113–128 MYA (mean, 120 MYA) according to the TimeTree web service (www.timetree.org) and 84–116 MYA (mean, ~99 MYA) according to our dating analysis. This WGD event might have occurred shortly before the split of the Magnoliales and Laurales, as was indicated in a recent study on the genome evolution of *Litsea*[89]. However, this hypothesis needs to be further examined in light of other results, such as the absence of a WGD event in Magnoliales *Annona muricata*[10].

The major effective component of the flower buds of *M. biondii*, a medicinal plant species, is the volatile oils comprising a rich array of terpenoids, mainly sesquiterpenoids and monoterpenoids[90]. TPS genes of the TPS-a and TPS-b subfamilies are mainly responsible for the biosynthesis of sesquiterpenoids and monoterpenoids, respectively, in mesangiosperms. Gene tree topologies of three angiosperm TPS proteins and comparisons of TPS subfamily members with those of the other angiosperms[11] revealed expansion of TPS genes in *M. biondii*, especially for members of TPS-a and TPS-b subfamilies. Expression profiles of TPS genes in different tissues revealed 33 TPS genes, primarily of the TPS-a and TPS-b subfamilies, that were substantially expressed in flowers compared to leaves. The expansion and significant expression of these TPS genes in the TPS-a and TPS-b subfamilies in *M. biondii* are in agreement with the high accumulation of sesquiterpenoids and monoterpenoids in the volatile oils extracted from the flower buds of *M. biondii*[90].

Conclusion

We constructed a reference genome of *M. biondii* by combining 10X Genomics Chromium, SMRT sequencing, and Hi-C scaffolding strategies. The ~2.22 Gb genome assembly of *M. biondii*, with a heterozygosity of 0.73% and a repeat ratio of 66.48%, represented the largest genome among six sequenced genomes of magnoliids. We predicted a total of 47,547 protein-coding genes from the genome assembly of *M. biondii*, 82.26% of which were functionally annotated. Phylogenomic reconstruction strongly supported the sister relationship of magnoliids and the Chloranthaceae, which together formed a sister relationship with a clade comprising monocots and eudicots. Our new genome information should enhance the understanding of the molecular basis of genetic diversity and individual traits in *Magnolia* as well as the molecular breeding and early radiation of angiosperms.

Acknowledgements
This work was supported by the National Key R&D Program of China (No. 2019YFC1711000), the National Natural Science Foundation (No. 31600171), the Shenzhen Urban Management Bureau Fund (No. 201520), and the Shenzhen Municipal Government of China (No. JCYJ20170817145512467). This work is part of the 10KP project. We sincerely thank the support provided by China National GeneBank.

Author details
[1]Laboratory of Southern Subtropical Plant Diversity, Fairy Lake Botanical Garden, Shenzhen & Chinese Academy of Sciences, Shenzhen 518004, China. [2]State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China. [3]Nanjing Forestry University, Nanjing 210037, China. [4]Fujian Agriculture and Forestry University, Fuzhou 350000, China. [5]University of British Columbia, Vancouver BC, Canada. [6]Xi'an Botanical Garden, Xi'an 710061, China. [7]Zhejiang Agriculture and Forestry University, Hangzhou 311300, China. [8]Kunming Botanical Garden, Chinese Academy of Sciences, Kunming 650201, China. [9]Department of Biology, University of Copenhagen, DK-2100 Copenhagen, Denmark

Author contributions
S.Z. and H.L. designed and coordinated the whole project. M.L., S.D., S.Z., and H.L. together led and performed all of the experiments. M.L, S.D., and F.C. performed the genome evolution and gene family analyses. S.D., M.L., H.L., S.Z., Y.L., X.G., and E.W. participated in the manuscript writing and revisions. All the authors have read and approved the final manuscript.

Conflict of interest
The authors declare that they have no conflict of interest.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41438-021-00471-9.

Received: 11 January 2020 Revised: 28 October 2020 Accepted: 12 December 2020
Published online: 01 March 2021

References

1. Rivers, M., Beech, E., Murphy, L. & Oldfield, S. The red list of Magnoliaceae-revised and extended. https://globaltrees.org/resources/the-red-list-of-magnoliaceae-revised-and-extended/ (2016).
2. Figlar, R. B. & Nooteboom, H. P. Notes on Magnoliaceae IV. *Blumea* **49**, 87–100 (2004).
3. Kim, S. & Suh, Y. Phylogeny of Magnoliaceae based on ten chloroplast DNA regions. *J. Plant Biol.* **56**, 290–305 (2013).
4. Azuma, H., García-Franco, J. G., Rico-Gray, V. & Thien, L. B. Molecular phylogeny of the Magnoliaceae: the biogeography of tropical and temperate disjunctions. *Am. J. Bot.* **88**, 2275–2285 (2001).
5. Soltis, D. E. & Soltis, P. S. Nuclear genomes of two magnoliids. *Nat. Plants* **5**, 6–7 (2019).
6. Soltis, D., Bell, C., Kim, S. & Soltis, P. S. Origin and early evolution of angiosperms. *Ann. N. Y. Acad. Sci.* **1133**, 3 (2008).
7. Kersey, P. J. Plant genome sequences: past, present, future. *Curr. Opin. Plant Biol.* **48**, 1–8 (2019).
8. Hu, L. et al. The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. *Nat. Commun.* **10**, 4702 (2019).
9. Rendón-Anaya, M. et al. The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen

Downloaded from https://academic.oup.com/hr/article/doi/10.1038/s41438-021-00471-9/6446648 by guest on 05 August 2026

influenced gene space adaptation. *Proc. Natl Acad. Sci. USA* **116**, 17081–17089 (2019).
10. Strijk, J. S. et al. The soursop genome and comparative genomics of basal angiosperms provide new insights on evolutionary incongruence. Preprint at *bioRxiv* https://doi.org/10.1101/639153 (2019).
11. Chaw, S. M. et al. Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nat. Plants* **5**, 63–73 (2019).
12. Chen, J. et al. Liriodendron genome sheds light on angiosperm phylogeny and species–pair differentiation. *Nat. Plants* **5**, 18–25 (2018).
13. Wickett, N. J. et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl Acad. Sci. USA* **111**, 4859–4868 (2014).
14. Zeng, L. et al. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nat. Commun.* **5**, 4956 (2014).
15. Gitzendanner, M. A., Soltis, P. S., Wong, G. K.-S., Ruhfel, B. R. & Soltis, D. E. Plastid phylogenomic analysis of green plants: a billion years of evolutionary history. *Am. J. Bot.* **105**, 291–301 (2018).
16. Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., Soltis, D. E. & Burleigh, J. G. From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evol. Biol.* **14**, 23 (2014).
17. Li, H. T. et al. Origin of angiosperms and the puzzle of the Jurassic gap. *Nature Plants* **5**, 461–470 (2019).
18. Wang, Y. L. & Zhang, S. Z. Studies on the microsporogenesis and development of the male gametophyte of Magnolia championii Benth. *J. Wuhan. Bot. Res* **26**, 547–553 (2008).
19. Hirayama, K., Ishida, K. & Tomaru, N. Effects of pollen shortage and self-pollination on seed production of an endangered tree, Magnolia stellata. *Ann. Bot.* **95**, 1009–1015 (2005).
20. Yang, X., Yang, Z. L., Wang, J., Tan, G. Y. & He, Z. S. Floral syndrome and breeding system of endangered species Magnolia officinalis subsp. biloba. *Chinese J. Ecol.* **3**, 551–556 (2012).
21. Wang, X. et al. Development of EST-SSR markers and their application in an analysis of the genetic diversity of the endangered species Magnolia sinostellata. *Mol. Genet. Genomics* **294**, 135–147 (2019).
22. Jiang, W., Cao, J., Li, G. & Weng, M. Development of new ornamental tree species of Magnolia family in China and its application in landscaping. *Acta Agriculturae Shanghai* **21**, 68–73 (2005).
23. Zhao, L. The terpenoid biosynthesis pathway in Magnolia and their significance for taxonomy in the genus. *Guihaia* **4**, 7 (2005).
24. Ho, K. Y., Tsai, C. C., Chen, C. P., Huang, J. S. & Lin, C. C. Antimicrobial activity of honokiol and magnolol isolated from Magnolia officinalis. *Phytother. Res.* **15**, 139–141 (2001).
25. China Pharmacopoeia Committee. *Pharmacopoeia of the People's Republic of China* The first Division of 2000 English edn (Chemical Industry Press, 2000).
26. Qu, L., Qi, Y., Fan, G. & Wu, Y. Determination of the volatile oil of Magnolia biondii pamp by GC–MS combined with chemometric techniques. *Chromatographia* **70**, 905–914 (2009).
27. Zhao, W., Zhou, T., Fan, G., Chai, Y. & Wu, Y. Isolation and purification of lignans from Magnolia biondii pamp by isocratic reversed-phase two-dimensional liquid chromatography following microwave-assisted extraction. *J. Sep. Sci.* **30**, 2370–2381 (2015).
28. Chen, Y., Gao, B. C., Qiao, L. & Han, G. Q. Study on the hydrophilic components of Magnolia biondii pamp. *Acta Pharm. Sin.* **29**, 506–510 (1994).
29. Porebski, S., Bailey, L. G. & Bernard, R. B. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol. Biol. Report* **15**, 8–15 (1997).
30. Chen, Y. et al. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* **7**, gix120 (2017).
31. Belaghzal, H., Dekker, J. & Gibcus, J. H. Hi-C 2.0: an optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods* **123**, 56–65 (2017).
32. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
33. Moscone, E. A. et al. Analysis of nuclear DNA content in Capsicum (Solanaceae) by flow cytometry and Feulgen densitometry. *Ann. Bot.* **92**, 21–29 (2003).
34. Chang, Y. et al. The draft genomes of five agriculturally important African orphan crops. *Gigascience* **8**, 1–16 (2019).
35. Liu, B. et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv Prepr.* **1308**, 2012 (2013).
36. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive\\r, k\\r, -mer weighting and repeat separation. *Genome Res.* **27**, 722 (2017).
37. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
38. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540 (2019).
39. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
40. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2016).
41. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
42. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
43. Dudchenko, O. et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
44. Dudchenko, O. et al. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. Preprint at *bioRxiv* https://doi.org/10.1101/254797 (2018).
45. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **1303**, 3997v2 (2013).
46. Kim, D. et al. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol. Evol.* **14**, R36 (2013).
47. Chang, Z. et al. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biol.* **16**, 30 (2015).
48. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
49. Jerzy, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
50. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* **25**, 4–10 (2009).
51. Hubley, R. & Smit, A. RepeatModeler. http://www.repeatmasker.org/RepeatModeler/ (2019).
52. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
53. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **1999**, 2 (1999).
54. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome annotation and curation using MAKER and MAKER-P. *Curr. Protoc. Bioinformatics* **48**, 4–11 (2014).
55. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506 (2005).
56. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinforma.* **7**, 62 (2006).
57. Johnson, A. D. et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938–2939 (2008).
58. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494 (2013).
59. Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **44**, W54–W57 (2016).
60. Kalvari, I. et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* **46**, D335–D342 (2017).
61. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
62. Vitales, D., D'Ambrosio, U., Gálvez, F., Kovařík, A. & Garcia, S. Third release of the plant rDNA database with updated content and information on telomere composition and sequenced plant genomes. *Plant Syst. Evol.* **303**, 1115–1121 (2017).
63. Aoki, K. F. & Kanehisa, M. Using the KEGG database resource. *Curr. Protoc. Bioinforma.* **11**, 1–12 (2005).
64. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
65. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
66. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

Downloaded from https://academic.oup.com/hr/article/doi/10.1038/s41438-021-00471-9/6446648 by guest on 05 August 2026

67. Finn, R. D. et al. The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–D288 (2007).
68. Letunic, I., Doerks, T. & Bork, P. SMART 6: recent updates and new developments. *Nucleic Acids Res.* **37**, D229–D232 (2009).
69. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.* **8**, 1551 (2013).
70. Attwood, T. K. et al. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.* **31**, 400–402 (2003).
71. Corpet, F., Servant, F., Gouzy, J. & Kahn, D. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.* **28**, 267–269 (2000).
72. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
73. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).
74. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577 (2007).
75. Lanfear, R., Calcott, B., Ho, S. Y. & Guindon, S. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* **29**, 1695–1701 (2012).
76. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
77. Yang, Z. PAML4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
78. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. Cafe: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).
79. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e49 (2012).
80. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
81. Tang, H. et al. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18**, 1944–1954 (2008).
82. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, 29–37 (2011).
83. Tamura, K. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2013).
84. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2014).
85. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–1359 (2012).
86. Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* **10**, 71–73 (2013).
87. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
88. Vanburen, R. et al. Single-molecule sequencing of the desiccation-tolerant grass oropetium thomaeum. *Nature* **527**, 508 (2015).
89. Chen, Y. et al. The Litsea genome and the evolution of the laurel family. *Nat. Commun.* **11**, 1675 (2020).
90. Lu, J. et al. Analysis of the chemical constituents of essential oil from Magnolia biondii by GC-MS. *J. Chin. Med. Mater.* **31**, 1649–1651 (2008).

Downloaded from https://academic.oup.com/hr/article/doi/10.1038/s41438-021-00471-9/6446648 by guest on 05 August 2026