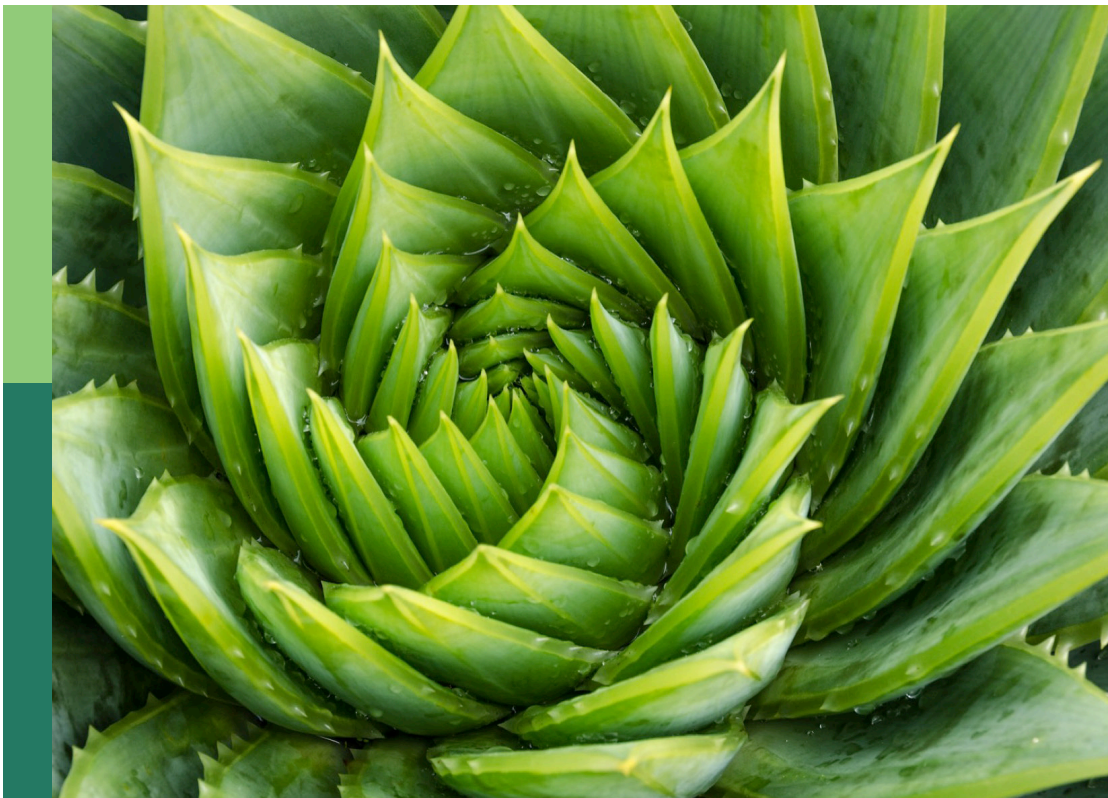
frontiers | Research Topics

Highly contiguous plant genome assembly and transcriptional regulation

Edited by
Kai-Hua Jia, Yongpeng Ma, Wei Zhao
and Guanjing Hu

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-8325-4207-1
DOI 10.3389/978-2-8325-4207-1

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Highly contiguous plant genome assembly and transcriptional regulation

Topic editors

Kai-Hua Jia — Shandong Academy of Agricultural Sciences, China

Yongpeng Ma — Kunming Institute of Botany, Chinese Academy of Sciences (CAS), China

Wei Zhao — Beijing Forestry University, China

Guanjing Hu — Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, China

Citation

Jia, K. -H., Ma, Y., Zhao, W., Hu, G., eds. (2024). *Highly contiguous plant genome assembly and transcriptional regulation*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-4207-1

Table of contents

frontiers | Frontiers in Plant Science

TYPE Editorial
PUBLISHED 13 December 2023
DOI 10.3389/fpls.2023.1331498

OPEN ACCESS

EDITED AND REVIEWED BY
Jihong Hu,
Northwest A&F University, China

*CORRESPONDENCE
Kai-Hua Jia
kaihuajia_saas@163.com

RECEIVED 01 November 2023
ACCEPTED 04 December 2023
PUBLISHED 13 December 2023

CITATION
Tian X-C and Jia K-H (2023) Editorial: Highly contiguous plant genome assembly and transcriptional regulation.
Front. Plant Sci. 14:1331498.
doi: 10.3389/fpls.2023.1331498

Editorial: Highly contiguous plant genome assembly and transcriptional regulation

Xue-Chan Tian[1,2] and Kai-Hua Jia[1]*

[1]Institute of Crop Germplasm Resources, Shandong Academy of Agricultural Sciences, Jinan, China, [2]National Engineering Research Center of Tree Breeding and Ecological Restoration, Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, National Engineering Laboratory for Tree Breeding, Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants, Ministry of Education, College of Biological Sciences and Technology, Beijing Forestry University, Beijing, China

KEYWORDS

genome assembly, Transcriptional regulation, secondary metabolites, TO-GCN, miRNA

Editorial on the Research Topic
Highly contiguous plant genome assembly and transcriptional regulation

Introduction

The past decade has seen a paradigm shift in plant genomics, with the advent of novel sequencing technologies and bioinformatic tools transforming our understanding of genetic complexity and diversity. The plant genome highly influences its functional and regulatory mechanisms, which in turn determine the plant's growth, development, and capacity to withstand environmental changes. Advances in long-read and high-fidelity sequencing technologies have revolutionized how plant genomes can be studied, enabling the generation of novel biologically accurate genomes. Understanding the complexity of the plant genome is vital, as it underpins the effective application of genetic and genomic resources to address challenges such as food security, climate change, and preserving biodiversity. In this context "*Highly Contiguous Plant Genome Assembly and Transcriptional Regulation*", the Research Topic of this Research Topic, underscores the significance of advancing plant genomics research to navigate the challenges of biological and environmental change. The Research Topic presents an assembly of remarkable studies and insights about cutting-edge genome sequencing methods, high-resolution plant genome assembly and the complexity of plant transcriptional regulation.

Advancements in highly contiguous plant genome assembly

This Research Topic exemplify the trend of integrating multiple sequencing technologies to achieve high-resolution plant genome assemblies. For instance, the *Quercus acutissima* (Liu et al.), *Alpinia oxyphylla* (Pan et al.) and *Melastoma candidum* (Zhong et al.) genome were effectively assembled employing a combination of PacBio

long read, Hi-C, and Illumina short read technologies. Likewise, the *Lindera megaphylla* genome (Tian et al.) was assembled through the utilization of ONT (Oxford Nanopore Technologies) long-read sequencing supplemented with Hi-C scaffolding technologies. The *Thalia dealbata* genome (Tang et al.) genome, a pioneering sequenced genome in the Marantaceae family, was assembled using PacBio HiFi reads and Hi-C technology. Additionally, a high-quality chromosome-scale reference genome of *Castanopsis hystrix* (Huang et al.) was obtained using a similar combination of Illumina and PacBio HiFi reads with Hi-C technology. Moreover, the innovative trio-binning method employed in the chili pepper genome assembly adroitly addressed the commonplace challenge of haplotype-switching associated with traditional plant genome assembly methods (Delorean et al.). The successful application of the trio-binning method underscores our evolving understanding and technological advancements in unraveling the complexities inherent in plant genomes.

Functional genomic insights

Following genome assembly, the research featured in this Research Topic offers a comprehensive perspective on functional genomics across a range of plant species. In Lauraceae family, the species revealed unique gene duplications and microsynteny related to isoquinoline alkaloids, elucidating the genetic mechanisms of wood decay resistance. The genome of *Melastoma candidum* provided critical insights into trichome evolution, with whole genome duplications playing a significant role in the expansion of trichome-related genes, thereby highlighting the impact of genomic alterations in morphological diversity. The exploration of *Alpinia oxyphylla* delved into genomic, transcriptomic, and metabolic profiles, correlating genomic variations with the synthesis of pharmacodynamic compounds like nootkatone. *Castanopsis hystrix* exhibited gene family expansions and contractions, pivotal for adapting to tropical and subtropical climates, enriching our understanding of the genomic foundations of environmental adaptation in forest trees. The genome of *Thalia dealbata*, a notable wetland plant, contributed substantially to the understanding of evolutionary adaptations to wetland environments and phylogenomic research in Marantaceae and Zingiberales.

Unraveling the complexity of plant transcriptional regulation

Another prominent theme across these studies is the intricate nature of plant transcriptional regulation. In *Epimedium pubescens*, Xu et al. employed high-temporal-resolution transcriptome analysis coupled with metabolite profiling, uncovering the biosynthesis pathways of bioactive compounds. Chen et al. employed a miRNA-TF-mRNA network to investigate wood development in rubber trees, shedding light on phenylpropanoid and lignin biosynthesis, thus enriching our knowledge of molecular regulatory mechanisms. Likewise, Cao et al.'s investigation of secondary trunk development in *Ginkgo biloba* identified key regulatory pathways and contributed further to our understanding of evolutionary and developmental biology in the context of plant genomics. By presenting these distinct yet connected studies, this Research Topic contributes to expanding our understanding of the vast and intricate world of plant transcriptional regulation.

Conclusion

The articles presented in this Research Topic contribute significantly to our understanding of highly contiguous plant genome assembly and transcriptional regulation. The genome assembly studies provide valuable resources for further ecological and evolutionary studies in *Alpinia oxyphylla, Capsicum annuum, Castanopsis hystrix, Ginkgo biloba*, *Lindera megaphylla, Melastoma candidum*, *Quercus acutissima* and *Thalia dealbata*. The gene regulation studies shed light on the molecular mechanisms underlying the biosynthesis of phenolic and flavonoid glycosides, trichome color variation, and secondary trunk development. These findings have implications for plant adaptation, diversification, and ecological interactions. Overall, this Research Topic highlights the importance of comprehensive genomic studies in unraveling the complexities of plant biology and provides a foundation for future research in this field.

Author contributions

X-CT: Writing – original draft, Writing – review & editing. K-HJ: Writing – original draft, Writing – review & editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 24 November 2022
DOI 10.3389/fpls.2022.1068802

OPEN ACCESS

EDITED BY
Kai-Hua Jia,
Shandong Academy of Agricultural Sciences, China

REVIEWED BY
Ke Qiang Yang,
Shandong Agricultural University, China
Yongqi Zheng,
Chinese Academy of Forestry, China

*CORRESPONDENCE
Wei Li
bjfuliwei@bjfu.edu.cn
Wenqing Li
190199191@qq.com

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 13 October 2022
ACCEPTED 02 November 2022
PUBLISHED 24 November 2022

CITATION
Liu D, Xie X, Tong B, Zhou C, Qu K, Guo H, Zhao Z, El-Kassaby YA, Li W and Li W (2022) A high-quality genome assembly and annotation of *Quercus acutissima* Carruth. *Front. Plant Sci.* 13:1068802. doi: 10.3389/fpls.2022.1068802

A high-quality genome assembly and annotation of *Quercus acutissima* Carruth

Dan Liu[1,2], Xiaoman Xie[2], Boqiang Tong[2], Chengcheng Zhou[1], Kai Qu[1], Haili Guo[2], Zhiheng Zhao[1], Yousry A. El-Kassaby[3], Wei Li[1]* and Wenqing Li[2]*

[1]National Engineering Research Center of Tree Breeding and Ecological Restoration, State Key Laboratory of Tree Genetics and Breeding, College of Biological Sciences and Technology, Beijing Forestry University, Beijing, China, [2]Shandong Provincial Center of Forest and Grass Germplasm Resources, Jinan, China, [3]Department of Forest and Conservation Sciences, The University of British Columbia, Vancouver, BC, Canada

Introduction: *Quercus acutissima* is an economic and ecological tree species often used for afforestation of arid and semi-arid lands and is considered as an excellent tree for soil and water conservation.

Methods: Here, we combined PacBio long reads, Hi-C, and Illumina short reads to assemble *Q. acutissima* genome.

Results: We generated a 957.1 Mb genome with a contig N50 of 1.2 Mb and scaffold N50 of 77.0 Mb. The repetitive sequences constituted 55.63% of the genome, among which long terminal repeats were the majority and accounted for 23.07% of the genome. *Ab initio*, homology-based and RNA sequence-based gene prediction identified 29,889 protein-coding genes, of which 82.6% could be functionally annotated. Phylogenetic analysis showed that *Q. acutissima* and *Q. variabilis* were differentiated around 3.6 million years ago, and showed no evidence of species-specific whole genome duplication.

Conclusion: The assembled and annotated high-quality *Q. acutissima* genome not only promises to accelerate the species molecular biology studies and breeding, but also promotes genome level evolutionary studies.

KEYWORDS

***Quercus acutissima*, genome assembly, gene annotation, phylogenetic analysis, gene families**

Introduction

As one of the largest genera in Fagaceae, *Quercus* (oak) contains more than 400 widely distributed species in Asia, Europe, Africa, and North America (Simeone et al., 2016). Oaks have various utilities, including timber, bioenergy, and dyes production (Sasaki et al., 2014; Wu et al., 2014; Li et al., 2018). According to molecular classification, the genera *Quercus* has been divided into two subgenera, *Quercus* and *Cerris* (Denk and Grimm, 2010; Denk et al., 2017; Deng et al., 2018; Hipp et al., 2018). The subgenera *Quercus* includes five groups (sections): *Ponticae*, *Virentes*, *Protobalanus* (intermediate Oak), *Quercus* (white oak), and *Lobatae* (red oak), while *Cerris* includes three groups (sections): *Ilex*, *Cerris* and *Cyclobalanopsis* (Denk et al., 2017). Within the *Quercus* genera, the evolutionary profiles of plastid genomes have been elucidated in *Q. acutissima*, *Q. aliena*, *Q. aquifolioides*, *Q. baronii*, *Q. dolicholepis*, *Q. edithiae*, *Q. fabri*, *Q. glauca*, and 10 other *Quercus* plastomes (Li et al., 2021). However, only four species with whole genome sequences have been published, including *Q. lobata* (Sork et al., 2016), *Q. suber* (Ramos et al., 2018), *Q. robur* (Plomion et al., 2016), and *Q. acutissima* (Fu et al., 2022). Although the genome data of *Q. acutissima* have been published, the continuity of the assembly still needs improvement (Fu et al., 2022).

As an important ecological and economic tree species, *Q. acutissima* Carruth is widely distributed in East Asia, especially in southeast China (18° - 41° N, 91° - 123°E) (Li et al., 2018; Yang et al., 2019). The silvics of *Q. acutissima* is usually mixed or secondary monocultures, which are also distributed in a scattered manner in harsh environments (Aldrich et al., 2003; Zhang et al., 2013). *Q. acutissima* timber provides excellent building material and charcoal production in many Asian countries, including China, Japan, and Korea (Zhang et al., 2013). At present, research on *Q. acutissima* is mainly focused on propagation, eco-physiology, selection, and genetic diversity (Dong, 2008; Wang et al., 2009; Liao, 2012; Zhang et al., 2013). In northern China, *Q. acutissima* forest ecosystems have been degraded due to human disturbance, threatening the species genetic resources (Aldrich et al., 2003; Zhang et al., 2013). Thus, planning breeding and conservation programs for *Q. acutissima* native populations is crucial, and the understanding of the species genome-wide evolution, gene function, and molecular breeding are important elements to supporting these goal (Greene and Morris, 2001).

Here, the *Q. acutissima* genome was sequenced and *de novo* assembled using PacBio long reads, Hi-C reads, and Illumina short reads. We performed structural gene annotation, repetitive sequences identification, and executed comparative genomics with other plant genomes. Our results are expected to improve our understanding of the evolution and diversification of genes in *Q. acutissima*, laying the foundation for novel genes discovery and ultimately contributing to the development of novel properties for the species breeding programs.

Materials and methods

Plant materials, DNA extraction and genome sequencing

Fresh *Q. acutissima* leaves were collected from a tree growing in the Shandong Provincial Center of Forest and Grass Germplasm Resources (36.62°N, 117.16°E), immediately frozen in liquid nitrogen, and stored at -80°C until further use. Plant specimens (barcode number SDF1001228) and total genomic DNA (code ld001qa001) were stored in Shandong Provincial Center of Forest and Grass Germplasm Resources. Total genomic DNA was extracted from leaf tissue using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. After obtaining high-quality purified genomic DNA samples, PCR free SMRT bell library was constructed and sequenced by PacBio sequencing platform, and we obtained 154.41 Gb of subreads with 160× coverage. We also constructed a Hi-C library and a paired-end library with an insert size of 350 bp and sequenced using the Illumina HiSeq X Ten platform.

Genome assembly, quality evaluation, and construction of pseudomolecule chromosomes

Before *Q. acutissima* genome *de novo* assembly, we used high-quality Illumina paired-end reads to estimate the genome size and heterozygosity with genomescope software (Vurture et al., 2017). Four software, including Canu (v2.1.1, default parameters) (Koren et al., 2017), FALCON (Chin et al., 2016), SmartDenovo (Istace et al., 2017), and WTDBG (Ruan and Li, 2019) were used to perform preliminary assembly of the genome. After the assembly of the third generation subreads, due to the presence of sequencing errors, a certain amount of error information existed such as short insertion-deletion mutations (Indel) and single-nucleotide polymorphism (SNP). Thus, we used the Illumina sort reads to polish this genome with BWA (v0.7.9a, parameter, -k 30) (Li and Durbin, 2009), and Pilon software (v1.22, default parameters) (Walker et al., 2014). Additionally, based on the OrthoDB (Kriventseva et al., 2019) database, we performed a BUSCO (version 3.0.1, default parameters) (Simão et al., 2015) assessment using single-copy orthologous genes to confirm the genome assembly quality. Quality control of the alignment reads was performed using the Phase Genomics Hi-C alignment quality control tool and scaffolding was carried out with Phase Genomics Proximo Hi-C

genome scaffolding platform to obtain chromosome-level assembly.

Genome annotation

We used a combination of *de novo* prediction and homology-based searches to annotate the genome tandem and interspersed repeats. First, RepeatModeler software (Flynn et al., 2020) was used to build the *de novo* repeat sequence library, and then we used RepeatMasker (Tarailo-Graovac and Chen, 2009), and Tandem Repeat Finder (Gary, 1999) software for repeat sequences prediction. Second, based on Repbase (Jurka et al., 2005), we used RepeatMasker to search homologous repeat sequences.

After repetitive sequence masking, we used three methods to predict gene structure. First, homology prediction was conducted by comparing homologous proteins from plant genomes, including *Q. lobata* (Sork et al., 2016), *Q. suber* (Ramos et al., 2018), *Q. robur* (Plomion et al., 2016), *Fagus sylvatica* (Mishra et al., 2018), and *Casuarina equisetifolia* (Ye et al., 2018) using Blast v2.2.28 and the GeneWise web resource v2.2.0 (Birney et al., 2004). Second, we used Augustus (Stanke et al., 2004), SNAP (https://github.com/KorfLab/SNAP), and GeneMark (Ter-Hovhannisyan et al., 2008) to *ab initio* gene prediction. Third, the PASA software (Roberts et al., 2011) was used to predict gene structure by aligning EST/cDNA sequences with the genome. Combining the above results, using the evincemodeler (EVM) (Haas et al., 2008) to integrate the gene set predicted by the three strategies into a nonredundant and more complete gene set.

We used the NCBI protein database, GO (Mi et al., 2019), KEGG (release 84.0) (Kanehisa et al., 2016), NR (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz), PFAM (Finn et al., 2014), and eggNOG-mapper (Cantalapiedra et al., 2021) to annotate gene function. The *E*-value cutoff was set to 1e-5 for BLAST searches.

Gene families and phylogenetic analysis

We downloaded (https://www.ncbi.nlm.nih.gov/) and performed a comparative genomic investigation of *Q. acutissima* with *Q. robur*, *Q. mongolica*, *Q. lobata*, *Q. variabilis*, *Q. suber*, *Castanea mollissima*, *Castanea crenata*, *Castanopsis tibetana*, *Fagus sylvatica*, *Juglans regia*, *Cyclocarya paliurus*, *Carya illinoinensis*, *Morella rubra*, *Corylus mandshurica*, *Carpinus viminea*, *Betula pendula*, and *Vitis vinifera*. The software OrthoFinder2 v2.3.1 (Emms and Kelly, 2019) was used to identify homoeologous gene clusters. IQ-TREE v1.6.7 (Nguyen et al., 2015) was used to construct a phylogenetic tree based on single copy homoeologous genes. The MAFFT v7.4.07 (Katoh and Standley, 2013) was used to align homoeologs before transforming aligned protein sequences into codon alignment. The concatenated amino acid sequences were trimmed using trimAL v1.4 (Capella-Gutiérrez et al., 2009) with -gt 0.8 -st 0.001 -cons 60. Divergence times were estimated using the MCMCTree software (Yang, 2007) in the PAML v4.9h (Guindon et al., 2010) package with the BRMC method (Sanderson, 2003; Blanc and Wolfe, 2004), and the correction times were taken from the TimeTree (Kumar et al., 2017): 109.0-123.5 MYA split time between *V. vinifera* and *B. pendula*, 56.8-95.0 MYA split time between *Q. suber* and *B. pendula*, and 35.7-83.5 MYA split time between *J. regia* and *B. pendula*. Based on the clustering analysis of gene families and dating, gene family expansion and contraction analyses were performed using CAFÉ (De Bie et al., 2006).

Synteny and WGD analysis

Syntenic blocks containing at least five genes were identified using the python version of MCScan (Huang et al., 2009; Schmutz et al., 2010) between *Q. mongolica*, *Q. variabilis*, *Q. acutissima*, *C. mollissima*, and *C. tibetana*. Genome circular plot was produced using Circos (Krzywinski et al., 2009). KaKs_Calculator 2.0 (Wang et al., 2010) was used to calculate *Ka, Ks*, and the *Ka/Ks* ratio by implementing the YN model.

GO enrichment analysis

GO enrichment analysis was performed using the R package clusterProfiler (Yu et al., 2012). The p values were adjusted for multiple comparisons using the method of Benjamini and Hochberg ($p < 0.05$ was considered significant).

Results

Genome sequencing and assembly

We sequenced Q. acutissima genome and generated a total of 154.41 Gb PacBio long reads with N50 of 24,256 bp (Table S1). The genome size and heterozygosity were estimated to be 750 Mb and 2.77% using K-mer analysis, respectively (Figure S1). To accurately assemble the Q. acutissima genome, we compared multiple assembly strategies in the primary step, and based on contiguity metrics including the total number of assembled contigs, N50, contigs' maximum length, and the best assembly from Canu was selected for further polishing and scaffolding with

Hi-C data. The assembled genome size was 957.09 Mb, including 1,507 contigs with an N50 length of 1.20 Mb and 15 scaffolds with N50 length 77.04 Mb (Table 1). The longest 12 scaffolds correspond to 12 pseudo-chromosomes (Figure 1).

Assessment of genomic integrity

The completeness and accuracy of the genome assembly were evaluated using BUSCO. The high BUSCO complete ratio (98.00%) corroborated the genome assembly excellent quality (Table S2). The guanine-cytosine (GC) depth analysis showed that there was no obvious left-right chunking in the GC-depth plot (Figure S2) and the average GC content was 35.18% (Table S3). Approximately 99.84% of the Illumina short reads could be successfully mapped to the genome assembly (Figure S3, Table S4). These results suggest that the assembly of the *Q. acutissima* genome is highly accurate and continuous.

TABLE 1 *Quercus acutissima* genome assembly statistics.

Assembly features	
Number of contigs	1,507
Contig N50 (Mb)	1.20
Number of scaffolds	15
Scaffold N50 (Mb)	77.04
Number of genes	29,889
Average gene length (bp)	4,476.10
Average exons per gene	4.92
Average exon length (bp)	253.60
Average intron length (bp)	824.50
Average Coding sequences length(bp)	1,247.79
Total size of repeat sequences (Mb)	532.33

Genome annotation

Through an integrative approach, we identified 546.67 Mb repetitive sequences, accounting for 57.13% of genome (Table 1, Table S5). The Long terminal repeat retrotransposons (LTR-RTs) from the largest proportion (23.07%) of the repeat (Table S6).

A total of 29,889 protein-coding genes were identified, their average lengths and coding sequences were 4,476.10 and 1,247.79 bp, respectively (Table 1). Based on the comparison between predicted gene sets with the annotation databases, a total of 24,689 (82.6%) genes were functionally annotated (Table S7).

Gene family and phylogenetic relationships

To assess the palaeohistory of *Q. acutissima*, we performed comparative genomic analyses incorporating *Q. acutissima* along with 16 other genomes and one outgroup (*V. vinifera*) (Figure 2). Out of the 28,312 gene families, only 10 were found to be unique to the *Q. acutissima* genome, and fewer than 60 gene families were unique to other *Quercus* (Table S8). Construction of the phylogenetic tree confirmed the evolutionary relationship within *Quercus*, and the divergence between *Q. variabilis* and *Q. acutissima* was estimated at 3.6 MYA (Figure 2). Expanded gene families provide the raw material for adaptation and trait evolution. We then examined the rates and direction of change in gene family size among taxa using CAFE (Han et al., 2013). The results showed

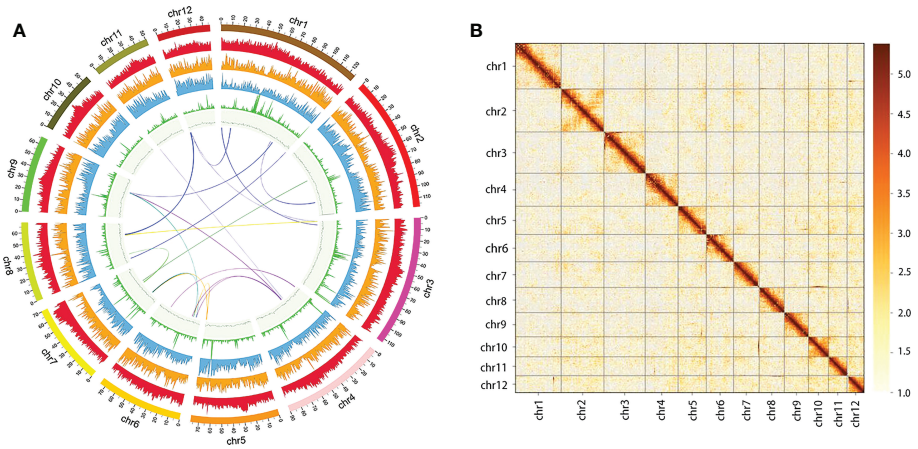


FIGURE 1
Q. acutissima genome features. **(A)** The genome circle plot (from the outer circle to the inner one, Class I transposable element (TE) density, Class II TE density, coding gene density, tandem repeat percentage, guanine-cytosine (GC) content, and co-linear block, respectively). **(B)** Twelve pseudo-molecules scaffolding with Hi-C data.

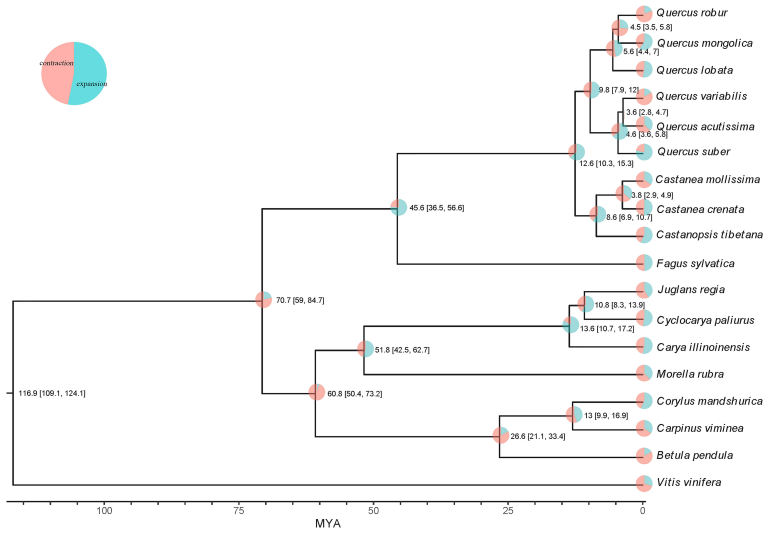


FIGURE 2
Maximum likelihood phylogenetic tree and expanded and contracted gene families in *Q. acutissima*. The numbers at the branch node in the tree indicate the divergence time and 95% confidence interval.

that *Q. acutissima* exhibited larger numbers of contracted gene families (2,390) than expanded (3,897) (Table S9, Figure 2). These expanded families are mainly related to ion transport, such as ion transport, ion transmembrane transport, inorganic ion transmembrane transport (Table S10), while the contracted gene families were mainly enriched to glycosinolate biosynthetic process, sesquiterpene metabolic and biosynthetic process, monoterpenoid metabolic and biosynthetic process (Table S11).

Whole-genome duplication and synteny analysis

Whole genome duplication (WGD) events are widespread and play a vital role in plant genome adaptation and evolution (Xue et al., 2020), and are an important source of gene family expansion. After multiple sequence alignment of sequences in synteny blocks within *Q. acutissima* and other species, the synteny analysis showed that *Q. acutissima* had a 1:1 syntenic relationship with other Fagaceae, and there was little rearrangement of chromosomes, which indicated that the evolution of Fagaceae was very conserved and no independent WGD events occurred in *Q. acutissima* (Figure 3, Figures S4-S11).

Discussion

Q. acutissima, Fagaceae, is an economically and ecologically important tree species with wide distribution in China (Li et al., 2018; Zhang et al., 2020). Here, we generated a *Q. acutissima* genome at the chromosome-level. The assembled genome size is approximately 956.9 Mb, which is larger than the genome we assessed using the *K*-mer method, this may be due to the presence of chimerism in our assembly. The development of PacBio sequencing has resulted in a considerable increase in contig N50 sizes compared to previous sequencing technologies (Wei et al., 2020). The assemble length of contig N50 sizes can represent the genome assembling quality (Yang et al., 2021), consequently, our genome has high assembly contiguity. High heterozygosity and repetition rates are responsible for the inability to assemble high-quality genomes (Gao et al., 2020; Wang et al., 2021). *Q. acutissima* heterozygous rate was 2.77%, which is higher than that of *Q. lobata* (1.25%) (Sork et al., 2016) and *Q. suber* (1.62%) (Ramos et al., 2018). It is worth noting that 98% of complete BUSCO core genes were detected in the assembled genome, which is higher than that of *Q. lobata* (94%) (Sork et al., 2016) and comparable to *Q. suber* genome (97%) (Ramos et al., 2018). In summary, *Q. acutissima* assembly is relatively accurate and complete, which will provide a valuable genome resource for understanding the species evolution and enhance its genetic improvement.

The genus *Quercus* (Fagaceae), which includes 400-500 species, is distributed in Asia, Africa, Europe, and North America (Simeone et al., 2016; Bent, 2020). As a member in this genus, *Q. acutissima* genome information can fill genome research gap and promote the species evolutionary biology research. Following the statistical analysis of repeat in the genome, we found that the repeat regions accounted for 57.13%, the numbers of repetitive and ncRNA sequences were relatively high in *Q. acutissima* compared with other *Quercus*

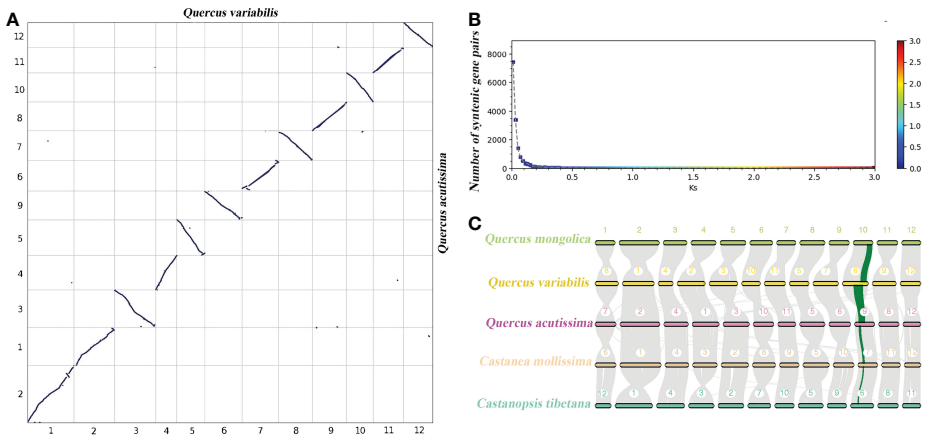


FIGURE 3
Syntenic dot plot and synteny analysis between *Q. acutissima* and other evaluated species. **(A)** Syntenic dot plot between the *Q. acutissima* and *Q. variabilis* genome. **(B)** Ks distribution between the *Q. acutissima* and *Q. variabilis* genome. **(C)** Synteny analyses among the genomes of *Q. acutissima, Q. mongolica, Q. variabilis, C. mollissima* and *C. tibetana*. Synteny blocks between paired chromosomes are connected by gray lines; one representative orthologous block (green lines) is noted.

species. To understand the evolutionary development of *Q. acutissima*, we analyzed its evolution and divergence times. The syntenic analysis indicated that *Q. acutissima* did not experienced a recent WGD event. In plants, WGD events can lead to genome size variation, gene family expansion, chromosomal rearrangement, and species evolution (El Baidouri and Panaud, 2013; Wang et al., 2021). We found high collinearity relationship between *Q. acutissima* and *Q. variabilis* chromosomes, suggesting the conservative nature of their karyotypes.

In summary, we obtained high-quality *Q. acutissima* genome sequences using Pacbio, Hi-C and Illumina reads. The development of sequencing technologies, analytical methods, and statistical algorithms continue to promote the efficiency and accuracy of genome sequencing and assembly (Xue et al., 2020; Wei et al., 2020; Wu et al., 2020). *Q. acutissima* genome includes high quality chromosomal-level assembly and many important genes, offering novel insights into genome evolution, functional innovation, and key regulatory pathways in wood formation and production of high-value metabolites, and providing excellent genetic resources for comparative genome studies among *Quercus* species.

Data availability statement

The data presented in the study are deposited in the CNGB Sequence Archive (CNSA, https://db.cngb.org/cnsa/) of China National GeneBank DataBase (CNGBdb) repository, accession number CNP0003530, CNP0002992.

Author contributions

WeiL and WenL designed and supervised the study. DL, XX, BT, CZ, and KQ collected the samples and extracted the genomic DNA and RNA. DL, CZ, KQ, HG and ZZ performed genome assembly and bioinformatics analysis. YE did English editing and retouching. DL wrote the original manuscript. WeiL and WenL reviewed and edited this manuscript. All authors read and approved the final manuscript.

Funding

This research was funded by the Collection and arrangement of genetic resources and genetic diversity evaluation of *Quercus acutissima* of Biosafety and Genetic Resources Management Project of State Forestry and Grassland Administration, grant number KJZXSA202111; 'Collection, Conservation, and Accurate Identification of Forest Tree Germplasm Resources' of Shandong Provincial Agricultural Elite Varieties Project, grant number 2019LZGC018; Project of National Forest Germplasm Resources Sharing Service Platform Construction and Operation, grant number 2005-DKA21003.

Acknowledgments

We are grateful for the generous grant from the National Engineering Research Center of Tree Breeding and Ecological Restoration and Shandong Provincial Center of Forest and Grass Germplasm Resources that made this work possible.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1068802/full#supplementary-material

References

Aldrich, P. R., Parker, G. R., Ward, J. S., and Michler, C. H. (2003). Spatial dispersion of trees in an old-growth temperate hardwood forest over 60 years of succession. *For. Ecol. Manage.* 180, 475–491. doi: 10.1016/s0378-1127(02)00612-6

Bent, J. S. (2020). *Quercus: classification ecology and uses* (America: Nova Science Publishers Inc).

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and genomewise. *Genome Res.* 14, 988–995. doi: 10.1101/gr.1865504

Blanc, G., and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16, 1667–1678. doi: 10.1105/tpc.021345

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation orthology assignments and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* 38, msab293. doi: 10.1093/molbev/msab293

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13, 1050. doi: 10.1038/nmeth.4035

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: A computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271. doi: 10.1093/bioinformatics/btl097

Deng, M., Jiang, X. L., Hipp, A. L., Manos, P. S., and Hahn, M. (2018). Phylogeny and biogeography of East Asian evergreen oaks (Quercus section cyclobalanopsis fagaceae): Insights into the Cenozoic history of evergreen broad-leaved forests in subtropical Asia. *Mol. Phylogenet Evol.* 119, 170–181. doi: 10.1016/j.ympev.2017.11.003

Denk, T., and Grimm, G. W. (2010). The oaks of western Eurasia: traditional classifications and evidence from two nuclear markers. *Taxon* 59, 351–366. doi: 10.1002/tax.592002

Denk, T., Grimm, G. W., Manos, P. S., Min, D., and Hipp, A. L. (2017). *An updated infrageneric classification of the oaks: review of previous taxonomic schemes and synthesis of evolutionary patterns* (Germany: Springer International Publishing), 13–38. doi: 10.1007/978-3-319-69099-5_2

Dong, Y. (2008). *Study on variation among quercus acutissima population and selection of its families and clones* (China: Shandong Agricultural University).

El Baidouri, M., and Panaud, O. (2013). Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome Biol. Evol.* 5, 954–965. doi: 10.1093/gbe/evt025

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 1–14. doi: 10.1186/s13059-019-1832-y

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkh121

Flynn, J. M., Hubley, R., Goubert, C., Goubert, C., Rosen, J., Clark, A. G., et al. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* 117, 9451–9457. doi: 10.1073/pnas.1921046117

Fu, R., Zhu, Y., Liu, Y., Feng, Y., Lu, R. S., Li, Y., et al. (2022). Genome-wide analyses of introgression between two sympatric Asian oak species. *Nat. Ecol. Evol.* 6, 924–935. doi: 10.1038/s41559-022-01754-7

Gao, S., Wang, B., Xie, S., Xu, X., Zhang, J., Pei, L., et al. (2020). A high-quality reference genome of wild cannabis sativa. *Hortic. Res.* 7, 73. doi: 10.1038/s41438-020-0295-3

Gary, B. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Greene, S. L., and Morris, J. B. (2001). The case for multiple-use plant germplasm collections and a strategy for implementation. *Crop Sci.* 41, 886–892. doi: 10.2135/cropsci2001.413886x

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 3, 307–321. doi: 10.1093/sysbio/syq010

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments *Genome Biol* Vol. 9, R7. doi: 10.1186/gb-2008-9-1-r7

Han, M. V., Thomas, G. W. C., Lugomartinez, J., and Hahn, M. W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol. Biol. Evol.* 30, 1987–1997. doi: 10.1093/molbev/mst100

Hipp, A. L., Manos, P. S., González, R. A., Hahn, M., Kaproth, M., McVay, J. D., et al. (2018). Sympatric parallel diversification of major oak clades in the americas and the origins of Mexican species diversity. *New Phytol.* 217, 439–452. doi: 10.1111/nph.14773

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., et al. (2009). The genome of the cucumber cucumis sativus l. *Nat. Gentic.* 41, 1275. doi: 10.1038/ng.475

Istace, B., Friedrich, A., Agata, L., Faye, S., Payen, E., Beluche, O., et al. (2017). *De novo* assembly and population genomic survey of natural yeast isolates with the Oxford nanopore MinION sequencer. *Gigascience* 6, 1–13. doi: 10.1093/gigascience/giw018

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase update a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110, 462–467. doi: 10.1159/000084979

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462. doi: 10.1093/nar/gkv1070

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly *via* adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., et al. (2019). OrthoDB v10: sampling the diversity of animal plant fungal protist bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 47, D807–D811. doi: 10.1093/nar/gky1053

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi: 10.1101/gr.092759.109

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: a resource for timelines timetrees and divergence times. *Mol. Biol. Evol.* 34, 1812–1819. doi: 10.1093/molbev/msx116

Liao, J. (2012). *Somatic embryogenesis and rapid propagation technology of quercus acutissima Carr* (China: Nanjing Forestry University).

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, X., Li, Y., Sylvester, S. P., Zang, M., El-Kassaby, Y. A., and Fang, Y. (2021). Evolutionary patterns of nucleotide substitution rates in plastid genomes of quercus. *Ecol. Evol.* 11, 13401–13414. doi: 10.1002/ece3.8063

Li, X., Li, Y., Zang, M., Li, M., and Fang, Y. (2018). Complete chloroplast genome sequence and phylogenetic analysis of quercus acutissima. *Int. J. Mol. Sci.* 19, 2443. doi: 10.3390/ijms19082443

Mi, H., Huang, X., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P. D. (2019). PANTHER version 14: More genomes a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* 47, D419–D426. doi: 10.1093/nar/gky1038

Mishra, B., Gupta, D. K., Pfenninger, M., Hickler, T., Langer, E., Nam, B., et al. (2018). A reference genome of the European beech (Fagus sylvatica l.). *GigaScience* 7, giy063. doi: 10.1093/gigascience/giy063

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Plomion, C., Aury, J., Amselem, J., Alaeitabar, T., Barbe, V., Belser, C., et al. (2016). Decoding the oak genome: public release of sequence data assembly annotation and publication strategies. *Mol. Ecol. Resour.* 16, 254–265. doi: 10.1111/1755-0998.12425

Ramos, A. M., Usié, A., Barbosa, P., Barros, P. M., Capote, T., Chaves, I., et al. (2018). The draft genome sequence of cork oak. *Sci. Data.* 5, 180069. doi: 10.1038/sdata.2018.69

Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-seq. *Bioinformatics* 27, 2325–2329. doi: 10.1093/bioinformatics/btr355

Ruan, J., and Li, H. (2019). Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* 17, 155–158. doi: 10.1038/s41592-019-0669-3

Sanderson, M. J. (2003). R8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19, 301–302. doi: 10.1093/bioinformatics/19.2.301

Sasaki, C., Kushiki, Y., Asada, C., and Nakamura, Y. (2014). Acetone-butanol-ethanol production by separate hydrolysis and fermentation (SHF) and simultaneous saccharification and fermentation (SSF) methods using acorns and wood chips of quercus acutissima as a carbon source. *Ind. Crop Prod.* 62, 286–292. doi: 10.1016/j.indcrop.2014.08.049

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178. doi: 10.1038/nature08670

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Simeone, M. C., Grimm, G. W., Papini, A., Vessella, F., Cardoni, S., Tordoni, E., et al. (2016). Plastome data reveal multiple geographic origins of quercus group ilex. *PeerJ* 4, e1897. doi: 10.7717/peerj.1897

Sork, V. L., Fitz-Gibbon, S. T., Puiu, D., Crepeau, M., Gugger, P. F., Sherman, R., et al. (2016). First draft assembly and annotation of the genome of a california endemic oak quercus lobata née (Fagaceae). *G3 (Bethesda Md.)* 6, 3485–3495. doi: 10.1534/g3.116.030411

Stanke, M, Steinkamp, R, Waack, S, and Morgenstern, B. (2004). AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Res* 32. doi: 10.1093/nar/gkh379

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinf.* 4, 4–10. doi: 10.1002/0471250953.bi0410s25

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O., and Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* 18, 1979–1990. doi: 10.1101/gr.081612.108

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. doi: 10.1093/bioinformatics/btx153

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* 9, e112963. doi: 10.1371/journal.pone.0112963

Wang, J., Xu, S., Mei, Y., Cai, S., Gu, Y., Sun, M., et al. (2021). A high-quality genome assembly of morinda officinalis a famous native southern herb in the lingnan region of southern China. *Hortic. Res.* 8, 135. doi: 10.1038/s41438-021-00551-w

Wang, B., Yu, M., Sun, H., Cheng, X., Dan, Q., and Fang, Y. (2009). Photosynthetic characters of quercus acutissima from different provenances under effects of salt tress. *Chin. J. Appl. Ecol.* 20, 1817–1824.

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. genomics. *Proteom. Bioinforma.* 8, 77–80. doi: 10.1016/s1672-0229(10)60008-3

Wei, S., Yang, Y., and Yin, T. (2020). The chromosome-scale assembly of the willow genome provides insight into salicaceae genome evolution. *Hortic. Res.* 7, 45. doi: 10.1038/s41438-020-0268-6

Wu, S., Sun, W., Xu, Z., Zhai, J., Li, X., Li, C., et al. (2020). The genome sequence of star fruit (Averrhoa carambola). *Hortic. Res.* 7, 95. doi: 10.1038/s41438-020-0307-3

Wu, T., Wang, G. G., Wu, Q., Cheng, X., Yu, M., Wang, W., et al. (2014). Patterns of leaf nitrogen and phosphorus stoichiometry among quercus acutissima provenances across China. *Ecol. Complex.* 17, 32–39. doi: 10.1016/j.ecocom.2013.07.003

Xue, T., Zheng, X., Chen, D., Liang, L., Chen, N., Huang, Z., et al. (2020). A high-quality genome provides insights into the new taxonomic status and genomic characteristics of cladopus chinensis (Podostemaceae). *Hortic. Res.* 7, 46. doi: 10.1038/s41438-020-0269-5

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yang, B., He, F., Zhao, X., Wang, H., Xu, X., He, X., et al. (2019). Composition and function of soil fungal community during the establishment of quercus acutissima (Carruth.) seedlings in a cd-contaminated soil. *Environ. Manage.* 246, 150–156. doi: 10.1016/j.jenvman.2019.05.153

Yang, C., Ma, L., Xiao, D., Liu, X., Jiang, X., Ying, Z., et al. (2021). Chromosome-scale assembly of the sparassis latifolia genome obtained using long-read and Hi-c sequencing. *G3 (Bethesda Md.).* 11, jkab173. doi: 10.1093/g3journal/jkab173

Ye, G., Zhang, H., Chen, B., Nie, S., Liu, H., Gao, W., et al. (2018). *De novo* genome assembly of the stress tolerant forest species casuarina equisetifolia provides insight into secondary growth. *Plant J* 97, 779–794. doi: 10.1111/tpj.14159

Yu, G., Wang, L. G., Han, Y., and He, Q. (2012). clusterProfiler: an r package for comparing biological themes among gene clusters. *Omics* 16, 284–287. doi: 10.1089/omi.2011.0118

Zhang, Y. Y., Fang, Y. M., Yu, M. K., Li, X. X., and Xia, T. (2013). Molecular characterization and genetic structure of quercus acutissima germplasm in China using microsatellites. *Mol. Biol. Rep.* 40, 4083–4090. doi: 10.1007/s11033-013-2486-6

Zhang, R. S., Yang, J., Hu, H. L., Xia, R. X., Li, Y. P., Su, J. F., et al. (2020). A high level of chloroplast genome sequence variability in the sawtooth oak quercus acutissima. *Int. J. Biol. Macromol.* 152, 340–348. doi: 10.1016/j.ijbiomac.2020.02.201

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 27 January 2023
DOI 10.3389/fpls.2023.1126319

OPEN ACCESS

EDITED BY
Kai-Hua Jia,
Shandong Academy of Agricultural Sciences, China

REVIEWED BY
Jian-Feng Mao,
Beijing Forestry University, China
Yongpeng Ma,
Kunming Institute of Botany (CAS), China

*CORRESPONDENCE
Seping Dai
daiseping@126.com
Renchao Zhou
zhrench@mail.sysu.edu.cn

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 17 December 2022
ACCEPTED 09 January 2023
PUBLISHED 27 January 2023

CITATION
Zhong Y, Wu W, Sun C, Zou P, Liu Y, Dai S and Zhou R (2023) Chromosomal-level genome assembly of *Melastoma candidum* provides insights into trichome evolution. *Front. Plant Sci.* 14:1126319. doi: 10.3389/fpls.2023.1126319

Chromosomal-level genome assembly of *Melastoma candidum* provides insights into trichome evolution

Yan Zhong[1], Wei Wu[1], Chenyu Sun[1], Peishan Zou[2], Ying Liu[1], Seping Dai[2]* and Renchao Zhou[1]*

[1]State Key Laboratory of Biocontrol and Guangdong Provincial Key Laboratory of Plant Resources, School of Life Sciences, Sun Yat-sen University, Guangzhou, China, [2]Guangzhou Institute of Forestry and Landscape Architecture, Guangzhou, China

Melastoma, consisting of ~100 species diversified in tropical Asia and Oceania in the past 1-2 million years, represents an excellent example of rapid speciation in flowering plants. Trichomes on hypanthia, twigs and leaves vary markedly among species of this genus and are the most important diagnostic traits for species identification. These traits also play critical roles in contributing to differential adaptation of these species to their own habitats. Here we sequenced the genome of *M. candidum*, a common, erect-growing species from southern China, with the aim to provide genomic insights into trichome evolution in this genus. We generated a high-quality, chromosome-level genome assembly of *M. candidum*, with the genome size of 256.2 Mb and protein-coding gene number of 40,938. The gene families specific to, and significantly expanded in *Melastoma* are enriched for GO terms related to trichome initiation and differentiation. We provide evidence that *Melastoma* and its sister genus *Osbeckia* have undergone two whole genome duplications (WGDs) after the triplication event (γ) shared by all core eudicots. Preferential retention of trichome development-related transcription factor genes such as C2H2, bHLH, HD-ZIP, WRKY, and MYB after both WGDs might provide raw materials for trichome evolution and thus contribute to rapid species diversification in *Melastoma*. Our study provides candidate transcription factor genes related to trichome evolution in *Melastoma*, which can be used to evolutionary and functional studies of trichome diversification among species of this genus.

KEYWORDS

***Melastoma candidum*, genome assembly, trichome evolution, whole genome duplication, transcription factor**

1 Introduction

Melastoma is a shrub genus distributed in tropical Asia and Oceania, with Southeast Asia as its species diversification center. This genus comprises about 100 species (Chen, 1984; Wong, 2016), which were estimated to be formed in the past 1-2 million years (Renner and Meyer, 2001), thus represents an exceptional example of

rapid species diversification in plants. All species of *Melastoma* have an erect-growing habit except *M. dodecandrum*, which is the only creeping species and also the first diverging species in this genus (Dai et al., 2019). Species of *Melastoma* are mainly recognized by trichomes in the hypanthia, young stems and leaves, which show a very rich diversity in shape, size, density and color among species (Wong, 2016). For example, the trichomes on the hypanthia include stellate hairs, scales, bristles, soft hairs and so on (Figure 1).

Trichomes possess protective functions and defense mechanisms against biotic and abiotic stresses such as herbivores, pathogens, and ultraviolet (UV) irradiation (Kang et al., 2010; Riddick and Simmons, 2014; Bickford, 2016; Rakha et al., 2017). It also plays an important role in biological functions such as development, seed dispersal, adaptation to extreme temperatures, and signal transmission (Hegebarth et al., 2016; Zhao and Chen, 2016; Zhou et al., 2017). Previous studies in *Melastoma* suggested that trichomes of different species or populations might contribute to their differential adaptation to heterogenous habitats (Ng et al., 2019). For example, *M. candidum*, always found in open habitats, has densely covered scales in the hypanthia and densely covered hairs in the leaves, which can resist the (UV) irradiation. In contrast, *M. sanguineum* usually occurs in shady understory, has sparse bristles in its hypanthia and glabrous leaves (Liu et al., 2014; Ng et al., 2019). In *M. normale*, populations with red and white trichomes in the young stems (twigs) exhibit higher fitness in their own habitats with high and low sunlight intensities, respectively, indicating differential adaptation[1]. Therefore, trichomes appear to be a key trait in *Melastoma*, providing various ecological opportunities in facilitating rapid species diversification in this genus. Under the ecology opportunity hypothesis (Schluter, 2000; Stankowski and Streisfeld, 2015), the ancestral species may have evolved some key ecologically related traits to take advantage of available resources.

A variety of factors, such as regulatory genes, non-coding RNAs, hormones and environment, are involved in regulating plant trichome initiation, growth and differentiation (Wang et al., 2019; Wang et al., 2021). Previous studies found that many transcription factors including R2R3-MYB, bHLH, WD40, HD-ZIP, WRKY and C2H2, play a critical role in trichome development of *Arabidopsis* and cotton (Yang and Ye, 2013; Wang et al., 2021). However, trichome development is regulated by different mechanisms in different plants, especially in the multicellular trichomes produced by most plants. For example, although the bHLH transcription factors are essential for the initiation of trichomes differentiation in *Arabidopsis*, it has no effect in tobacco (*Nicotiana tabacum*) and tomato (*Solanum lycopersicum*) (Lloyd et al., 1992).

Genes functioning in the initiation and differentiation of trichomes have been characterized in model plants like *Arabidopsis* (Szymanski et al., 2000), cotton (Zhao and Chen, 2016) and tomato (Rakha et al., 2017), but similar studies have been rarely conducted in non-model plants, including *Melastoma* in which trichomes play an important role in species diversification and ecological adaptation. Many genomic processes including whole genome duplication and gene family expansion can provide raw materials for the evolution of new traits and adaption to novel environments in plants (Li et al., 2016; Feng et al., 2020; Wu et al., 2020). Based on the remarkable trichome diversity in *Melastoma*, we predict that trichome-related genes might have been expanded in the genome of this genus. To date, only the genome of *M. dodecandrum* has been reported (Hao et al., 2022), however, the annotation of this genome is not complete (see Results) and no analyses on trichome evolution have been performed in that study. Here we report the sequencing, assembly, annotation and characterization of the genome of *M. candidum*, an erect-growing species widely distributed in southern China, northern Vietnam and Okinawa of Japan (Chen, 1984). We aimed to connect the genomic features in *Melastoma* and thus to understand trichome evolution in this genus.

2 Results

2.1 Genome assembly and annotation

The genome size of *M. candidum* was estimated to be about 257.1 Mb based on a K-mer (k=21) analysis of Illumina sequencing data (Figure S1). Using PacBio long reads, we generated a genome assembly of 256.2 Mb, which represents 99.7% of the estimated genome size and consists of 266 scaffolds. 98.0% (251.2 Mb) of the scaffold sequences were anchored to the 12 pseudochromosomes based on the Hi-C data (Figure 2). The N50 and N90 of the scaffolds were 20.5 Mb and 13.7 Mb, respectively (Table 1). The PacBio long reads and Illumina short reads have mapping rates of 96.0% and 97.9%, and cover 99.8% and 99.6% of the genome, respectively (Table S1). The total mapping rate of Illumina RNA-seq reads to the genome was 95.6% (Table S1). 96.9% of 1614 Benchmarking Universal Single-Copy Orthologs (BUSCOs) genes in the embryophyta_odb10 and 92.9% of 2326 BUSCOs genes in the eudicots_odb10 datasets were recovered in our genome assembly (Table S2). The LTR Assembly Index (LAI) across the genome is 25.5. All the genome continuity, completeness and accuracy assessment results above suggest that the genome assembly of *M. candidum* is of high quality.

Repetitive sequences account for 31.5% of the genome (Table 1). Most of them are long terminal repeat retrotransposons (LTR), covering 23.1% of the genome (Table S3). The two major superfamilies, Ty3/*Gypsy* and Ty1/*Copia*, account for 11.7% and 7.8% of the genome, respectively. The DNA transposons take up 6.3% of the genome. We predicted 40,938 protein-coding genes in the *M. candidum* genome (Table 1), by combining *de novo* prediction, transcriptome evidence and homology-based approaches. 91.3% genes could be annotated in at least one of the functional annotation databases (Table 1; Table S4). The average exon and intron sizes were 279 bp and 228 bp, respectively (Table 1). In addition, 1,818 non-coding RNAs including 188 miRNAs, 233 rRNAs, 699 tRNAs, and 698 snRNAs were identified. 96.0% and 93.5% of the BUSCOs genes in the two datasets mentioned above were recovered based on our genome annotation (Table S5).

1 Huang G., Wu W., Chen Y., Zhi X., Zou P., Ning Z., Fan Q., Liu Y., Deng D., Zeng K., Zhou R. Balancing selection on an MYB transcription factor maintains the twig trichome color variation in Melastoma normale. unpublished.

Although the BUSCO assessment revealed comparable gene recovery rates between the genomes of *M. candidum* and *M. dodecandrum*, we found that the *M. candidum* genome has higher proportion of single-copy genes (72.4%) and lower proportion of duplicated genes (20.5%) of complete BUSCOs than the *M. dodecandrum* genome (69.4% and 23.9%, respectively) in the eudicots_odb10 dataset (Figure S2). Based on the genome annotation, *M. candidum* has 5,257 (12.8% of the predicted 40,938 genes) more genes than *M. dodecandrum*, in which 35,681 genes were predicted. Meanwhile, the BUSCO assessment with protein mode showed that, of the 2,326 genes in the eudicots_odb10 dataset, *M. candidum* recovered 231 more genes than *M. dodecandrum* (Figure S2). The 231 genes are either fragmented (60) or missing (171) in *M. dodecandrum*. The similar situation was observed in the embryophyta_odb10 dataset (Figure S3). Taken together, this suggests incomplete gene annotation for the *M. dodecandrum* genome, given very low divergence between the two species (see below).

2.2 Phylogenetic position and short species diversification history of *Melastoma*

The topology of the constructed maximum likelihood tree of 13 species including *M. candidum* based on sequences of 346 single copy genes is consistent with previous studies (Myburg et al., 2014; Hao et al., 2022) and confirms that Myrtales is sister to the ancestor of Fabids and non-Myrtales Malvids (Figure 3), but is not consistent with its position shown in APGIV (2016). *Melastoma* is sister to *Osbeckia*, which is in agreement with previous studies (Veranso-Libalah et al., 2017). In the tree, the *Melastoma* and *Osbeckia* clade is then sister to *Eucalyptus*, another species with available genome from Myrtales. The ancestral branch leading to *Melastoma* and *Osbeckia* (0.306) is roughly twice as long as the *Eucalyptus* branch (0.154) (Figure S4), suggesting accelerating evolution for the branch leading to *Melastoma* and *Osbeckia* after diverging from *Eucalyptus*. This is likely the consequence of much shorter generation time of *Melastoma* and *Osbeckia* compared with the tree genus *Eucalyptus*. Within

FIGURE 1
Trichomes on the hypanthia of *Melastoma*. From left to right, the first row: *Melastoma saigonense* (Vietnam), *M. beccarianum* (Malaysia), *M. dendrisetosum* (China), *M. ultramaficum* (Malaysia); the second row: *M. sabahense* (Malaysia), *M. normale* (China), *M. candidum* (China), *M. affine* (China),; the third row: *M. setigerum* (Indonesia), *M. sanguineum* (Cambodia), *M. sanguineum* (China), *M. penicillatum* (China); the fourth row: *M.* sp. (Vietnam), *M. laevifolium* (Malaysia), *M. kudoi* (China), *M. dodecandrum* (China).

Melastoma, *M. candidum* and *M. dodecandrum* have extremely short branch length (0.005 and 0.014), indicating very recent divergence between them. Considering that *M. dodecandrum* is the first-diverging species of the genus *Melastoma*, the whole genus should have a short evolutionary history of species diversification. The divergence time between *M. candidum* and *M. dodecandrum* was dated back to 4.6 Ma (Figure 3), larger than the previous estimation of 1-2 Ma (Renner and Meyer, 2001). However, whether 1-2 Ma or 4.6 Ma is a fairly short evolutionary time for the formation of about 100 species in this genus, both supporting rapid speciation.

2.3 Genomic synteny between *M. candidum* and *M. dodecandrum*

Genomic synteny analysis between the two species shows that the 12 chromosomes are in a relatively good one-to-one correspondence between them despite the existence of some structural variations (Figure S5). There are 960 syntenic blocks between *M. candidum* and *M. dodecandrum*, with the number of gene pairs in these blocks ranging from 5 to 2252. A total of 57,880 gene pairs were identified in these blocks, involving 30,890 genes of *M. candidum* and 29,263 genes of *M. dodecandrum*. The inter-species syntenic blocks in the *M. candidum* genome are totally 249.3 Mb in length and contain 40,448 genes (including genes not in the gene pairs between the two species), while the counterparts in the *M. dodecandrum* genome are totally 272.6 Mb and contain 33,504 genes. The two genomes have 7.8 Mb and 13.9 Mb of non-syntenic regions, respectively. We also found that most of the syntenic blocks show a 2:2 correspondence between *M. candidum* and *M. dodecandrum* (Figure S6), indicating the existence of whole genome duplication in the two species.

2.4 Two whole genome duplications were shared by *Melastoma* and *Osbeckia*

1621 syntenic blocks were identified within the genome of *M. candidum*. The number of gene pairs in these blocks range from 5 to 492, with a mean of 27. The number of genes in these blocks is 27,956, covering 68.3% of the annotated genes in the genome. The synonymous substitution rate (Ks) distribution for all paralogous gene pairs in the syntenic blocks of the *M. candidum* genome have two peaks, very close to the two peaks identified in the genome of *M. dodecandrum* and the transcriptome of *Osbeckia opipara* (Figure 4). This implies that the two WGDs, the recent σ event at Ks = 0.256-0.280 and the more ancient ρ event at Ks = 0.927-1.022, were shared by the three species. Both of the two WGDs occurred after diverging from *Eucalyptus* (Hao et al., 2022). Because the γ triplication event is shared by all the core eudicots (Jiao et al., 2012), including *Eucalyptus* (Myburg et al., 2014), the two WGDs, both with smaller Ks peak values, must have happened after the γ event. The peaks of the Ks distribution of orthologous gene pairs between *Osbeckia opipara* and either species of *Melastoma* were much less than that for the recent WGD (Figure 4), further supporting the inference that the two WGDs occurred prior to the divergence of *Melastoma* and *Osbeckia*. The distribution of Ks between

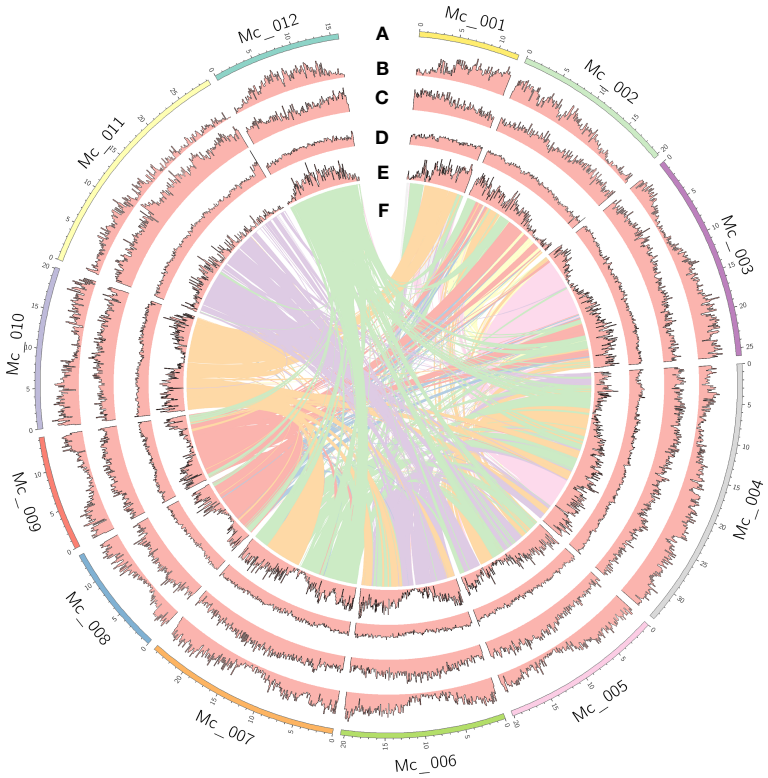


FIGURE 2
Genome features of *Melastoma candidum*. Tracks displayed are: **(A)** 12 pseudochromosomes; **(B)** gene density; **(C)** density of repeats; **(D)** GC content; **(E)** density of genes in the syntenic blocks; **(F)** inter-chromosome synteny.

10.3389/fpls.2023.1126319

TABLE 1 Summary of genome assembly and annotation for *Melastoma candidum*.

Assembly features	
Genome-sequencing depth (×)	212
Assembly genome size (Mb)	256.2
Estimated genome size (Mb)	257.1
GC content	42.9%
Scaffolds number	266
Scaffold N50 (bp)	20,460,156
Scaffold L50	5
Scaffold N90 (bp)	13,725,599
Scaffold L90	11
Contig N50 (bp)	2,018,067
Contig N90 (bp)	447,033
Annotation features	
Number of predicted genes	40,938
Mean gene length (bp)	2387.2
Mean exon length (bp)	278.9
Mean intron length (bp)	227.7
Mean of exon number per gene	5.2
Repeat content (% of the genome assembly)	80.6 Mb (31.5%)
Number of functionally annotated genes	37,393

N50: sequence length of the shortest contig/scaffold at 50% of the total genome length
L50: the smallest number of contigs/scaffolds whose length sum makes up half of genome size
N90: sequence length of the shortest contig/scaffold at 90% of the total genome length
L90: the smallest number of contigs/scaffolds whose length sum makes up 90% of genome size

orthologous gene pairs in the syntenic blocks between the two species of *Melastoma* has a peak at Ks = 0.023 (Figure 4), again suggesting very recent divergence between them.

2.5 *Melastoma* specific gene families contain genes related to trichome development

Homology clustering of protein sequences of the 12 species including *M. candidum* and *M. dodecandrum* implemented in OrthoFinder2 produced 28,371 orthologous groups. Of the 40,938 predicted genes in *M. candidum*, 36,924 (90.2%) were assigned to 17,108 gene families (the percentage of unassigned genes is 9.8%), in which 503 gene families comprising 1,358 predicted genes were specific to *M. candidum* (Table S6). There are 2,744 gene families specific to the two species of *Melastoma*, including 4,130 *M. candidum* genes and 3,549 *M. dodecandrum* genes.

The unique gene families of *M. candidum* among the 11 species (excluding *M. dodecandrum* in this analysis) were significantly enriched for 189 GO terms (Supplementary Excel file 1). Many of these GO terms (Category: "Biological Process") were associated with cell differentiation, including seed trichome differentiation (GO:0090376) and seed trichome elongation (GO:0090378), and

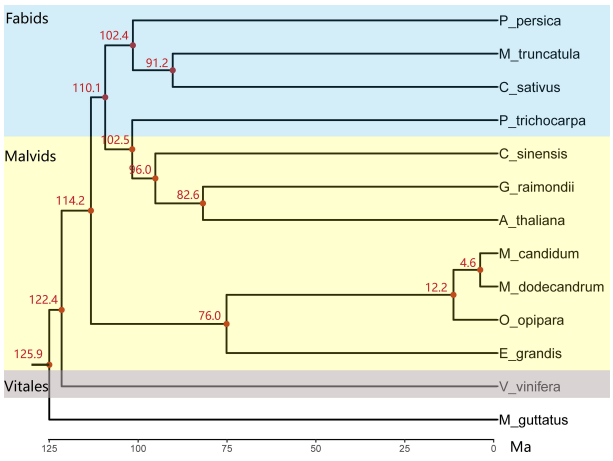


FIGURE 3
The chronogram tree of 12 rosids species and one outgroup based on the concatenated sequences of 346 single-copy genes. All nodes have a 100% bootstrap support value. The three different background colors (top to bottom) represent Fabids, Malvids and Vitales based on APG IV (2016), respectively. Species abbreviations: A_thaliana, *Arabidopsis thaliana*; C_sinensis, *Cirtus sinensis*; C_sativus, *Cucumis sativus*; E_grandis, *Eucalyptus grandis*; G_raimondii, *Gossypium raimondii*; M_truncatula, *Medicago truncatula*; M_dodecandrum, *Melastoma dodecandrum*; M_candidum, *Melastoma candidum*; O_opipara, *Osbeckia opipara*; P_trichocarpa, *Populus trichocarpa*; P_persica, *Prunus persica*; V_vinifera, *Vitis vinifera*; M_guttatus, *Mimulus guttatus*.

environmental resistance, including response to red or far red light (GO:0009639) and shade avoidance (GO:0009641) (Table 2). These enriched gene families include some transcription factors (Table S7), such as bHLH, HD-ZIP, and WRKY, which were known to be implicated in trichome development in *Arabidopsis* (Chalvin et al., 2020). These genes specific to *Melastoma* may contribute to trichome evolution in *Melastoma*.

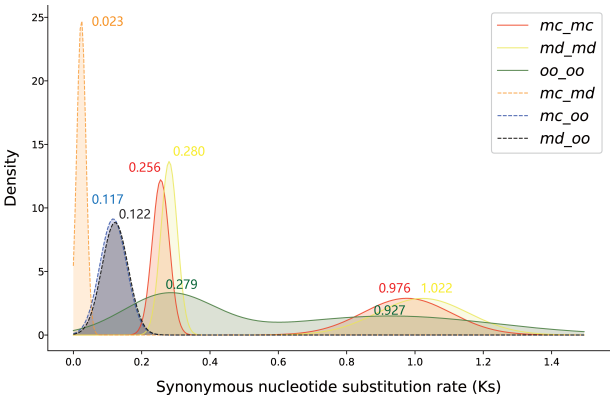


FIGURE 4
The frequency density distribution of synonymous substitution rate (Ks) of paralogous gene pairs in the syntenic blocks within, and orthologous gene pairs between, the three species, *Melastoma candidum*, *M. dodecandrum* and *Osbeckia opipara*. The values of Ks peaks are labeled. mc, *Melastoma candidum*; md, *M. dodecandrum*; oo, *Osbeckia opipara*.

TABLE 2 GO Enrichment analysis result of the gene families unique to *Melastoma*.

Category	GO term	Description	P value
Cell differentiation	GO:0090376	seed trichome differentiation	4.75e-4
	GO:0090378	seed trichome elongation	4.75e-4
	GO:0048863	stem cell differentiation	4.28e-6
	GO:0035987	endodermal cell differentiation	3.21e-4
Environment resistance	GO:0009639	response to red or far red light	2.62e-6
	GO:0009641	shade avoidance	2.37e-5
	GO:0009611	response to wounding	1.66e-4
	GO:0071236	cellular response to antibiotic	5.96e-6

2.6 Expanded gene families in *Melastoma* contain trichome-related transcription factors

We identified 3,027 expanded and 467 contracted gene families in *M. candidum* (Figure S7), among which 1,176 were significantly expanded ($P < 0.05$) and 287 were significantly contracted ($P < 0.05$). At the node of the last common ancestor leading to *M. candidum* and *M. dodecandrum*, 726 (4,120 genes) and 441 (864 genes) gene families were significantly expanded and contracted ($P < 0.05$), respectively. Enrichment analysis for the significantly expanded gene families in the common ancestor of *M. candidum* and *M. dodecandrum* identified some GO terms related to cell differentiation, including trichome differentiation (GO:0010026) and epithelial cell differentiation (GO:0030855), and response to environment, including defense response to fungus (GO:0050832) and response to antibiotic (GO:0046677) (Table 3; Supplementary Excel file 2). Similar to the results above, a high proportion (up to 68.6%) of genes belonging to these enriched GO terms are transcription factors, including bHLH, C2H2, HD-ZIP, WRKY, MYB, and MYB_related (Tables 3, S8).

Among the 18 transcription factor families with more than 10 members and fitting the normal distribution in gene number among the 12 species, 11 are largest, and 7 are the second or third largest in *M. candidum* (Table S9), including all six trichome-related transcription factor families shown in Table 3. Ten transcription factor families in *M. candidum* have a Z-score > 1.5, including HD-ZIP, MYB, WRKY, and bHLH (Table S10), which are also trichome-related transcription factors with enriched GO terms listed in Table 3. The Z-score results are consistent with the gene family evolution analysis, suggesting that *M. candidum* has more trichome-related transcription factor families than 10 other species.

2.7 Preferential retention of trichome-related genes following the WGDs

According to the criteria in Methods, 8,608 genes (21.0% of the total genes) have been retained from the σ event, much larger than those (3,858 genes; 9.4%) retained from the ρ event. Among the transcription factor families in *M. candidum*, including bHLH, C2H2, WRKY, HD-ZIP, MYB and MYB-related, roughly half (40.6%-50.3%) of members in these transcription factor families are retained after WGDs in *M. candidum* (Table 4). We identified the retention of transcription factors in the two WGDs, including 98 bHLH, 59 C2H2, 65 WRKY, 34 HD-ZIP, 93 MYB, and 30 MYB-related genes following the σ event (Table 4). For the ρ event, a smaller number of genes were identified, including 43 bHLH, 27 C2H2, 33 WRKY, 7 HD-ZIP, 35 MYB, and 22 MYB-related genes (Table 4). Whether for all the transcription factor families as a whole or for individual families, the number of genes retained after the σ event far exceeds that retained after the ρ event, which may be explained by more recent occurrence of the σ event.

However, GO terms for trichome morphogenesis (GO:0010090) and trichome differentiation (GO:0010026) were enriched in genes retained after the ρ event, but not in genes retained after the σ event (Table 5; Supplementary Excel file 3, 4). In addition, genes retained after the ρ event was enriched for morphogenesis of a polarized epithelium and regulation of morphogenesis of an epithelium, whereas genes retained after the σ event was enriched for epithelium development and regulation of cell differentiation (Table 5; Supplementary Excel file 3, 4). This suggests that although more genes were retained after the σ event, trichome formation and development-related genes were mainly expanded and retained after the ρ event, and downstream genes for epithelial cell development and regulation were also retained after the σ event. The two WGD events together contributed to the retention of trichome-related genes and provided the genetic materials for trichome evolution in *Melastoma* as well as *Osbeckia*, thus facilitating their adaptation to different habitats.

3 Discussion

Trichomes are the product of epidermal cell differentiation (Hülskamp, 2004; Yang and Ye, 2013). Plant trichome development is coordinated and regulated by a complex network of regulatory genes, non-coding RNA, hormones, and environmental factors (Wang et al., 2019; Wang et al., 2021). In *Arabidopsis*, two acting together models, activator-depletion and activator-inhibitor models, have been well-established and proposed for explaining the molecular regulatory mechanisms of trichome development (Wang et al., 2021). Many transcription factors, including MYB-bHLH-WD40 complex, are involved and form the hub of regulating plant trichome initiation, growth and differentiation (Yang and Ye, 2013; Wang et al., 2019). The positive regulators are represented by the R2R3-MYB transcription factor GLABRA1 (GL1) and its counterparts MYB23, the bHLH transcription factor GL3 and its close homolog ENHANCER OF GLABRA3 (EGL3), and the WD40-repeat transcription factor TRANSPARENT TESTA GLABRA1 (TTG1) (Li et al., 2009b; Zhao et al., 2012). GL1/MYB23, GL3/EGL3 and TTG1 are combined to form a MYB-bHLH-TTG1 complex (Serna and Martin, 2006). This regulatory complex simulates trichome initiation by promoting the expression of GL2 and TTG2, which encode a homeodomain-leucine zipper (HD-Zip) and a WRKY transcription factor, respectively (Johnson et al., 2002). GL3-

TABLE 3 GO enrichment analysis result of significantly expanded gene families in the common ancestor of *Melastoma candidum* and *M. dodecandrum*.

Category	GO term	Description	P value	Gene number	TF %[a]	Trichome-related transcription factors (number)
Cell differentiation	GO:0010026	trichome differentiation	5.34e-3	31	48.4%	bHLH (2); C2H2 (2); HD-ZIP (3); MYB (5)
	GO:0060429	epithelium development	5.63e-6	29	20.7%	bHLH (2); C2H2 (3); HD-ZIP (1)
	GO:0030855	epithelial cell differentiation	6.05e-6	19	31.6%	bHLH (2); C2H2 (2); HD-ZIP (1)
	GO:0045595	regulation of cell differentiation	1.08e-3	44	15.9%	HD-ZIP (3); C2H2 (2); MYB (1)
	GO:0000904	cell morphogenesis involved in differentiation	8.36e-3	60	25.0%	bHLH (2); C2H2 (2); HD-ZIP (3); MYB (5)
Response to environment	GO:0050832	defense response to fungus	1.41e-8	83	42.2%	WRKY (13); MYB (12)
	GO:0002833	positive regulation of response to biotic stimulus	1.66e-8	44	34.1%	WRKY (13); MYB (2)
	GO:0032101	regulation of response to external stimulus	1.77e-3	53	28.3%	WRKY (13); MYB (2)
	GO:0031347	regulation of defense response	4.31e-3	61	24.6%	WRKY (13); MYB (2)
	GO:0046677	response to antibiotic	1.26e-7	86	68.6%	WRKY (13); MYB (35); MYB_related (4)

[a]The percentage of transcription factors (TF) in enriched GO terms.

dependent depletion of TTG1 in trichome neighboring cells is the core foundation of activator-depletion model. The negative regulatory factors mainly consist of genes encoding single-repeat R3 MYB proteins, including CAPRICE (CPC), TRIPTYCHON (TRY) (Szymanski et al., 2000). They combine with GL3/EGL3 and TTG1 by competing with GL1/MYB23 to form an inactivating complex, thereby inhibiting trichome formation (Esch et al., 2003; Yang et al., 2020). ENHANCER OF TRY AND CPC 1 (ETC1) and ETC2 act as enhancers of TRY and CPC. Activator-inhibitor model is explained by TRY/CPC to form an inactive TRY/CPC-GL3/EGL3-TTG1 complex, which negatively regulates trichome formation by replacing the transcription factor GL1/MYB23. These transcription factors form the hub of regulating trichome initiation and differentiation. In addition, cytokinin (CK) increases trichome formation through C2H2 transcription factors (Wang et al., 2021).

The trichome of *Arabidopsis* has been intensively studied as a model for cell differentiation, but trichome development in different plants and organs can be regulated by different mechanisms. For example, homologs of GL1, GL2, TTG1, and HD-ZIP in cotton were reported to have similar function to those in *Arabidopsis*, but the negative regulators like single-repeat R3 MYB transcription factors have not been identified in cotton (Zhang et al., 2010; Yang and Ye, 2013). It is unknown whether the activator-inhibitor model is effective in cotton fiber (trichome on its seed) development. Also, the two types of multicellular trichomes, short- and long-stalked, produced in tobacco are explained by different developmental mechanisms (Payne et al., 1999). The multicellular trichomes including those of *Melastoma* may be controlled by mechanisms that are more complex than those of unicellular trichomes such as *Arabidopsis* and cotton. *Melastoma* has the advantage to dissect the mechanisms because a high level of trichome diversity is displayed among species of this genus.

TABLE 4 The number of genes in the genome and the number of genes retained after two WGDs (σ and ρ events) for six transcription factor families in *Melastoma candidum*.

	bHLH	C2H2	WRKY	HD-ZIP	MYB	MYB-related
# of genes in the genome	229	158	147	81	260	101
# of genes retained after σ+ρ events	114	71	74	35	111	41
# of genes retained after σ event	98	59	65	34	93	30
# of genes retained after ρ event	43	27	33	7	35	22

The high-quality genome assembly of *M. candidum* makes it feasible to provide genomic insights into the evolution of trichomes, a key trait in contributing to species diversification and ecological adaptation in *Melastoma* (Wong, 2016; Ng et al., 2019; Huang et al. unpublished data). Based on the remarkable trichome diversity in *Melastoma*, we predict the expansion of trichome-related genes in the *M. candidum* genome, which can provide the raw materials for trichome evolution and thus effective response to diverse biotic and abiotic stresses. Our genomic analysis results are consistent with this prediction.

First, GO enrichment analysis of gene families specific to, and significantly expanded in *Melastoma* found that both were enriched for GO terms related to cellular differentiation (including trichome differentiation) and environmental response. This involves six transcription factor families, including C2H2, bHLH, HD-ZIP, WRKY, MYB, and MYB-related, which are key regulators of trichome formation and differentiation in plants such as *Arabidopsis* (Wang et al., 2019; Wang et al., 2021), as detailed below. Large and frequent changes in gene family size among species might be associated with some important morphological, physiological, and behavioral differences among them (Demuth and Hahn, 2009).

10.3389/fpls.2023.1126319

TABLE 5 GO enrichment analysis result of the genes retained after two WGDs in *Melastoma candidum*.

WGD	GO term	Description	P value
σ event	GO:0045595	regulation of cell differentiation	1.77e-3
	GO:0060429	epithelium development	1.05e-3
	GO:0090558	plant epidermis development	1.27e-3
ρ event	GO:0010090	trichome morphogenesis	4.67e-3
	GO:0010026	trichome differentiation	1.60e-3
	GO:1905330	regulation of morphogenesis of an epithelium	1.22e-3
	GO:0001738	morphogenesis of a polarized epithelium	2.75e-3

Meanwhile, we provided genomic evidence that two WGDs (the σ and ρ events) happened in, and shared by *Melastoma* and *Osbeckia*. After a WGD event, one of the duplicated genes may be lost or become pseudogenes or both duplicates may be retained *via* sub-functionalization or neo-functionalization (Edger and Pires, 2009; Li et al., 2016; Li et al., 2021). Given that both *Melastoma* and *Osbeckia* have a great variety of trichomes, we propose the hypothesis that the two WGDs allowed the expansion and retention of trichome-related genes. Our enrichment analysis of genes duplicated and retained by the two WGDs found that the two WGD events contributed to preferential retention of many trichome-related transcription factors, and GO terms including trichome morphogenesis and trichome differentiation were enriched after the ρ event. This supports our hypothesis that the ρ event allowed the expansion and retention of genes promoting trichome development, and the event σ further duplicated and retained trichome-associated genes. Therefore, both WGD events have contributed to retention of trichome-related genes.

In summary, trichome-associated transcription factors were identified in *Melastoma*-specific, significantly expanded, and preferentially retained genes after two WGDs. These transcription factors have been shown to be central components of the regulatory network for trichome formation and differentiation in model plants such as *Arabidopsis* and cotton. Therefore, we suggest that these expanded genes, especially duplicated and retained transcription factors in the ρ event, provide the raw genetic materials for trichome evolution and further contribute to ecological adaptation of *Melastoma*.

4 Conclusion

We assembled and annotated a high-quality, chromosome-level genome of *M. candidum*. Genomic data support very recent divergence between *M. candidum* and *M. dodecandrum* (Ks peak of the orthologous gene pairs at 0.02) and good synteny of 12 chromosomes between them. Two WGD events were identified in, and shared by *Melastoma* and *Osbeckia*, two sister genera both with a high level of trichome diversity. We found that the gene families involved in trichome initiation and differentiation were significantly expanded, and meanwhile, trichome-related genes, especially related transcription factor genes, were preferentially retained following the two WGDs, which together may greatly contribute to trichome evolution in *Melastoma*. Since trichomes in species of *Melastoma* contribute to their adaptation to diverse environments, the expansion and retention of trichome-related genes may promote rapid species diversification in this genus. The *Melastoma* genome also provides an ideal genomic resource for ecological and evolutionary studies in this genus, particularly transcription factor genes in association with trichome evolution.

5 Materials and methods

5.1 Plant materials and sequencing

One individual of *Melastoma candidum* from Wenchang, Hainan, China was collected and transplanted in the campus of Sun Yat-sen University (SYSU) and used for *de novo* genome sequencing. Total DNA was extracted from fresh leaves of this individual using the modified cetyltrimethylammonium bromide (CTAB) protocol (Doyle, 1991). RNA was isolated from leaves, flowers, young branches, fruits and two whorls of stamens of the individual using the method described in (Fu et al., 2004).

Four DNA libraries with the insert sizes of 180 bp, 300 bp, 500 bp and 800 bp were constructed and then sequenced on an Illumina Hiseq2000 platform. A PacBio library with an insertion size 20 Kb was also constructed and sequenced on the PacBio RSII sequencer with DNA Sequencing Kit 2.0 (Pacific Biosciences, CA, USA) (Table S11). Transcriptome libraries using RNA isolated from the six tissues mentioned above were constructed and then sequenced separately on an Illumina Hiseq2000 platform. Moreover, a Hi-C library following a standard procedure (Lieberman-Aiden et al., 2009) was constructed, and then sequenced on an Illumina HiSeq X Ten sequencer. All the details of these sequencing data were shown in Table S11. In addition, the fresh leaves of one individual of *Osbeckia opipara* sampled from Chishui, Guizhou, China was used for transcriptome sequencing using the same method for *Melastoma*.

5.2 Genome size estimation

Illumina reads were filtered by fastp 0.20.1 (Chen et al., 2018) and FastUniq (Xu et al., 2012) with default parameters. All clean reads were supplied to Jellyfish v.2.3.0 (Marçais and Kingsford, 2011) to calculate Kmer (k = 21) frequency. The genome size, as well as the heterozygosity and repeat content were then estimated in GenomeScope 1.0 (Vurture et al., 2017).

5.3 Genome assembly, annotation and quality assessment

A two-step procedure was implemented to assemble the draft genome of *M. candidum*. First, *de novo* assembly was implemented using ALLPATH-LG v52488 (Gnerre et al., 2011) with default settings except for the two parameters: ploidy set to 2, and estimated genome size set to 257 Mb. At this stage, only Illumina reads were supplied

into the assembler, and corrected with the embedded modules PreCorrect and FindErrors with 24-kmer read stacks. Next, the pre-assembled contigs of *M. candidium* were scaffolded by SSPACE v3.0 (Boetzer et al., 2011), and the gaps were closed with Gapclose v1.12 (Luo et al., 2012). The PacBio subreads were corrected with LoRDEC v0.9 (Salmela and Rivals, 2014) and were then used to fill gaps and scaffolding all the available scaffolds with PBJelly v14.1.15 (Patel and Jain, 2012).

After mapping the clean Hi-C reads against the scaffolds using BWA v0.7.12-r1039 (Li and Durbin, 2009) with default parameters, we corrected, clustered, sorted, and anchored the scaffolds >1 kb into 12 pseudomolecules using Juicer v1.6 (Durand et al., 2016) and 3D-DNA (Dudchenko et al., 2017). Then, Juicebox Assembly Tools (https://github.com/aidenlab/Juicebox) was used to manually review the scaffolds and plot the contact maps. The final genome of 12 pseudochromosomes was obtained with the run-asm-pipeline-post-review.sh script in 3D-DNA. Finally, we clipped scaffolds < 1 kb in length.

EDTA v1.9.6 (Ou et al., 2019) was used to identify repetitive sequences with default parameters. Noncoding RNA including rRNA, tRNA, miRNA, snoRNA were predicted using INFERNAL v 1.1.4 (Nawrocki and Eddy, 2013) by searching the *M. candidum* genome against the RNA family database release 14.7 (RFAM v 14.7) with the parameters "-Z 512 –cut_ga –rfam –nohmmonly –fmt 2" (Kalvari et al., 2021).

Protein-coding genes were predicted using a combination of homologous-sequence search, *ab initio* gene prediction, and transcriptome-based prediction implemented in a genome annotation tool GETA v 2.4.5 (https://github.com/chenlianfu/geta). Illumina RNA-seq reads from different tissues were mapped to the genome assembly using HISAT2 (Kim et al., 2019) and were used for transcriptome-based prediction. Protein sequences from six eudicots (*Arabidopsis thaliana*, *Cirtus sinensis*, *Gossypium raimondii*, *Medicago truncatula*, *Populus trichocarpa*, and *Vitis vinifera*) and plant protein sequences from UniProtKB/Swiss-Prot (https://www.uniprot.org/) were used for homology-based prediction with GeneWise (https://www.ebi.ac.uk/~birney/wise2/). *ab initio* prediction was performed in Augustus v3.3.3 (Stanke and Morgenstern, 2005), trained with intron and exon information generated above. These prediction results were integrated and then were searched against the Pfam database for screening to get the final gene prediction result. Functional annotation of genes was performed with InterproScan (Jones et al., 2014), eggnog-mapper (http://eggnog-mapper.embl.de/), PANNZER2 (Törönen et al., 2018), and Mercator4 v3.0 (Schwacke et al., 2019).

We used four approaches to assess the quality of genome assembly and annotation of *M. candidum*. First, genome continuity and completeness were assessed using QUAST v5.1.0 (Gurevich et al., 2013) to count the scaffold N50, L50, N90 and L90 of the genome. Second, Illumina DNA reads, PacBio reads and RNA-seq reads were mapped to the genome using BWA-MEM (Li, 2013), Minimap2 (Li, 2018) and HISAT2 v2.1.0 (Kim et al., 2019), respectively. The accuracy of the genome was assessed by the analysis of sequencing depth, percentage of mapped reads and genome coverage using SAMtools (Li et al., 2009a), bamdst (https://github.com/shiquan/bamdst) and Qualimap 2 (Okonechnikov et al., 2015). Third, the completeness of genome assembly and annotation was assessed using BUSCO v5.1.3 (Simão et al., 2015) with both eudicots_odb10 and embryophyta_odb10 databases. The *M. dodecandrum* genome published before (Hao et al., 2022) was also assessed with the same method. Finally, the continuity of the genome was also assessed by LTR Assembly Index (LAI) using LTR_retriever v2.9.0 (Ou et al., 2018; Ou and Jiang, 2018).

5.4 Transcriptome assembly and assessment of *Osbeckia opipara*

Illumina reads of *Osbeckia opipara* were first trimmed for quality using fastp 0.20.1 (Chen et al., 2018). The clean reads were *de novo* assembled to 159,724 transcripts using Trinity v2.11.0 (Haas et al., 2013) with default parameters. The transcripts were clustered using cd-hit-est v4.8.1 (Li and Godzik, 2006)with the identity parameter set to 0.95% and the longest transcript for each cluster was selected using the script get_longest_isoform_seq_per_trinity_gene.pl in Trinity, which led to the output of 64,927 unigenes. The BUSCO assessment of these unigenes revealed 84.9% and 89.7% of the complete BUSCOs in the eudicots_odb10 and embryophyta_odb10 dataset, respectively. TransDecoder v5.5.0 was employed to predict coding regions of these unigenes and then to translate them into amino acid sequences. Only sequences > 150 aa in length were kept for subsequent analyses.

5.5 Phylogeny construction and divergence time estimation

Using OrthoFinder2 (Emms and Kelly, 2019), gene families of *M. candidum* and other 12 related species, namely, *M. dodecandrum*, *Arabidopsis thaliana*, *Cirtus sinensis*, *Cucumis sativus*, *Euclyptus grandis*, *Gossypium raimondii*, *Medicago truncatula*, *Mimulus guttatus*, *Osbeckia opipara Populus trichocarpa*, *Prunus persica*, and *Vitis vinifera* (Table S12), were clustered with default parameters. One-to-one orthogroups among *M. candidum* and other 12 species were identified as single copy genes. For each single copy gene, protein sequences of these species were aligned using MAFFT v7.0 (Katoh and Standley, 2013), and then their corresponding nucleotide sequences were aligned using the pal2nal.pl script (Suyama et al., 2006). All the nucleotide sequence alignments were concatenated into a supermatrix, and then subject to substitution model test using ModelFinder (Kalyaanamoorthy et al., 2017) with the Bayesian information criterion (BIC). The maximum-likelihood tree between *M. candidum* and other 12 species was constructed under GTR+F+R4 model using IQ-TREE v2.0.3 (Nguyen et al., 2014) with 1000 ultrafast bootstrap replicates (Hoang et al., 2017) and *Mimulus guttatus* as an outgroup.

The divergence time in the ML tree was estimated by mcmctree program in the PAML package (Yang, 2007) under a relaxed clock model with independent rates constraints and two calibration points, one between Asterids and Rosids (111-131 million years ago, Ma), and the other between *Arabidopsis* and *Populus* (98-117 Ma) from TimeTree (http://www.timetree.org).

5.6 Enrichment analysis of *Melastoma* specific gene families

For this analysis, we redid the gene family clustering for the 12 species in OrthoFinder2 by excluding *Osbeckia opipara* because it has only transcriptome data and genes are not adequate to represent those in its genome. We extracted the gene families specific to *Melastoma* by combining gene families both specific to the two species of *Melastoma* and specific to *M. candidum*. Enrichment levels of Gene Ontology (GO) terms were evaluated by comparing genes in the *Melastoma* specific gene families with the genomic background (all annotated genes of *M. candidum*) in the clusterProfiler v4.2.2 package (Wu et al., 2021) of R. Statistical significance was tested by Fisher's exact test (Fisher, 1922) and adjusted *P* values were calculated according to the Benjamini and Hochberg (false discovery rate) method (Benjamini and Hochberg, 1995). We used default parameters except for pAdjustMethod = "BH", pvalueCutoff = 0.05, and qvalueCutoff = 0.2.

5.7 Gene family expansion and contraction analysis

Gene family clustering results for the 12 species in the preceding section were used for this analysis. We analyzed changes in gene family size across a specified chronogram tree of the 12 species using 601 single copy genes of these species in CAFÉ 5 (Mendes et al., 2020). Gene gain and loss rates were modeled using a birth and death process. Poisson distribution was specified as root frequency distribution model. Gene families >100 members were filtered with the script clade_and_size_filter.py. Evolutionary rates were estimated using different K values (evolutionary rate categories) ranging from 2 to 8. Birth and death rate with the maximum likelihood value (K = 4) were used to infer ancestral states of gene family sizes for each node and changes along each branch in the phylogenetic tree. Gene families with significant expansion and contraction were determined with a threshold conditional P-value ($P < 0.05$). Changes of gene family size along each branch were labeled in the phylogenetic tree. GO enrichment analysis of the significantly expanded gene families in the common ancestor of the two species of *Melastoma* was performed using the methods and parameters mentioned above.

5.8 Genomic synteny analysis and whole genome duplication identification

All-versus-all alignment of the protein sequences of *M. candidum* was constructed using the blastp algorithm (Altschul et al., 1997). To detect the signature of whole genome duplication (WGD), the icl module in WGDI v0.5.1 (Sun et al., 2022) was employed to define syntenic blocks with minimum gene number of five and evalue threshold of 1e-5 in the blast search. For each gene pair in the syntenic blocks, synonymous nucleotide substitution rate (Ks) was calculated by the ks module in WGDI with the YN00 model. Tandems and blocks with significance more than 0.1 were filtered and by the kp module with parameters "tandem = true; pvalue = 0.1". To avoid random errors and the effect of synonymous substitution saturation, we retained gene pairs with the Ks values > 0.05 and Ks values < 1.50, which is the upper limit of the divergence between *Melastoma* and *Eucalyptus* (Hao et al., 2022). According to color of dotplots and troughs value (0.6) of Ks frequency distribution, syntenic blocks from the older and younger WGDs were separated by the kp module with parameters "homo = 0,0.5; ks_area = 0.6,1.5" and "homo = 0.5,1; ks_area = 0.05,0.6", respectively. The frequency distribution of Ks for each of the WGDs was fitted individually a normal distribution with Gaussian model using the pf module with "mode=median", and then the kf module was used to make a plot based on the fitted parameters. Intra-genomic syntenic blocks of *M. dodecandrum* were analyzed with the same method.

For the *O. opipara* transcriptome, the gene families were constructed using the mclblastline pipeline (Enright et al., 2002), and each gene family was compared using MUSCLE (Edgar, 2004), and finally the codeml module in the PAML package (Yang, 2007) was used to calculate the Ks values with the YN00 model. The script KSPloter.py (https://github.com/EndymionCooper/KSPlotting) was employed to execute the above process with mode 1 (-R M1).

Meanwhile, we extracted single copy genes of *M. candidum*, *M. dodecandrum* and *O. opipara* as a representative of orthologs to calculate pairwise Ks of gene pairs among the three species. For each gene pair, the protein sequences were aligned and then converted into coding sequence alignments by ParaAT (Zhang et al., 2012). The Ks value was calculated using KaKs_Calculator 2.0 (Wang et al., 2010) with the YN00 model. Gene pairs with the Ks values > 0.05 and < 1.50 were retained. The frequency distribution of Ks for each peak is constructed with 200 bins and is fitted a normal distribution with a Gaussian model.

The density of genes, repeats, genes within the syntenic blocks, and GC content in the 12 pseudochromosomes of *M. candidum* were calculated in a 100-kb sliding window with BEDTools v2.30.0 (Quinlan and Hall, 2010) and were plotted with syntenic curves between chromosomes using Circos v 0.69-8 (Krzywinski, 2009). Syntenic regions between the 12 pseudochromosomes of *M. candidum* and *M. dodecandrum* were identified and plotted using the MCScan pipeline (Tang et al., 2008).

5.9 Gene retention analysis after the WGDs

Based on the phylogenetic tree constructed above, gene duplication events were identified with parameters "-M msa -T raxml" in OrthoFinder2. We firstly conducted gene trees for each orthogroup using maximum likelihood method. Then, reconciliation of all the nodes of gene trees with corresponding nodes in the species tree was executed to obtain resolved gene trees and to further infer gene duplication events.

Firstly, orthogroups containing four and more genes and at least one gene from non-*Melastoma* species were kept. Then, we extracted gene duplication events specific to *Melastoma* and screened the gene family trees to accurately identify gene duplication events with the two criteria: 1) Both of the two child branches of each gene duplication event have genes from *M.*

candidum, and 2) The bootstrap support values are not less than 0.5. We further eliminated tandem duplications when two duplicated genes located within the range of five genes. Finally, we got 1,026 gene duplication events.

Pairwise protein sequences for all duplicated genes were aligned and then converted into nucleotide sequence alignments using ParaAT. Ks value for pairwise comparisons at the duplication node (one gene in one child branch and the other gene in another child branch) were calculated using the YN00 model implemented in KaKa_Calculator2.0. The mean of Ks for each gene duplication event were then calculated. We filtered out gene duplication events with Ks mean value < 0.05 and > 1.50. To further validate if the duplicate genes are still located on syntenic blocks, we extracted duplicate genes in the syntenic blocks based on the WGD analysis results. To distinguish genes duplicated by the two WGDs, we defined that duplicated gene pairs in the syntenic blocks produced by the σ event belong to the retained genes after the σ event and so do gene pairs that were produced by the ρ event. GO enrichment analysis of the retained genes after the two WGDs was performed using the same methods mentioned above.

5.10 Transcription factors retention analysis

The transcription factor families for the two species of *Melastoma* were annotated PlantTFDB v5.0 (Jin et al., 2016; Tian et al., 2019), and transcription factors of 10 other species were download from PlantTFDB v5.0. We examined whether the gene number of each family from the 12 species fits a normal distribution using the Shapiro-Wilk Test. Only transcription factor families with a normally distributed gene number across the 12 species were retained, and those with very few members (< 10) in any species were removed. For each screened family, we calculated the z-score according to the formula $z = (x-\mu)/\sigma$, in which x, μ and σ represent the gene number of the family in *M. candidum*, the mean and the standard deviation of gene numbers of this family in the 12 species, respectively. We then analyzed the retention of transcription factor genes after the two WGDs. Using the same method as the gene duplication retention analysis described above, we identified the retained transcription factor genes of *M. candidum* for each of the two WGDs.

Data availability statement

All raw reads of *Melastoma candidum*, including the Pacbio, Hi-C, Illumina DNA-seq, and RNA-sequencing, were deposited in the National Center for Biotechnology and Information (NCBI) short read archive repository under the accession numbers SRR22574044-SRR22574048. The genome assembly and annotation of *Melastoma candidum* was deposited in NCBI GenBank under the accession number: JAKZET000000000. (BioProject accession: PRJNA811312). The RNA-sequencing raw reads of *Osbeckia opipara* was deposited in the NCBI short read archive repository under the accession number SRR22557471. The *de novo* assembly was deposited in NCBI Transcriptome Shotgun Assembly Sequence Database under the accession number: GKED00000000. (BioProject accession: PRJNA909408).

Author contributions

RZ and SD planned the projects. YZ analyzed data and wrote the manuscript. CS and PZ performed the experiments. WW and YL participated in plant sampling and sequencing. WW participated in data analysis, and YZ and RZ revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was financially supported by the National Natural Science Foundation of China (32170217, 31670210 and 31811530297).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1126319/full#supplementary-material

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

APGIV (2016). An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. *Botan. J. Linn. Soc.* 181, 1–20. doi: 10.1111/boj.12385

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B (Methodological)* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bickford, C. P. (2016). Ecophysiology of leaf trichomes. *Funct. Plant Biol.* 43, 807–814. doi: 10.1071/FP16095

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27, 578–579. doi: 10.1093/bioinformatics/btq683

Chalvin, C., Drevensek, S., Dron, M., Bendahmane, A., and Boualem, A. (2020). Genetic control of glandular trichome development. *Trends Plant Sci.* 25, 477–487. doi: 10.1016/j.tplants.2019.12.025

Chen, J. (1984). *Melastomataceae* (Beijing: Science Press).

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Dai, J.-H., Lin, C.-W., Zhou, Q.-J., Li, C.-M., Zhou, R.-C., Liu, Y., et al. (2019). The specific status of *Melastoma kudoi* (Melastomataceae, melastomeae). *Botan. Stud.* 60, 1–11. doi: 10.1186/s40529-019-0253-2

Demuth, J. P., and Hahn, M. W. (2009). The life and death of gene families. *Bioessays* 31, 29–39. doi: 10.1002/bies.080085

Doyle, J. (1991). DNA Protocols for plants. *In. Mol. techniques taxonomy: Springer.* p, 283–293. doi: 10.1007/978-3-642-83962-7_18

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). *De novo* assembly of the *Aedes aegypti* genome using Hi-c yields chromosome-length scaffolds. *Science* 356, 92–95. doi: 10.1126/science.aal3327

Durand, N. C., Shamim, M. S., Machol, I, Rao, S. S. P., Huntley, M. H., Lander, E. S., et al. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-c experiments. *Cell Syst.* 3, 95–98. doi: 10.1016/j.cels.2016.07.002

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Edger, P. P., and Pires, J. C. (2009). Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.* 17, 699–717. doi: 10.1007/s10577-009-9055-9

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238. doi: 10.1186/s13059-019-1832-y

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584. doi: 10.1093/nar/30.7.1575

Esch, J. J., Chen, M., Sanders, M., Hillestad, M., Ndkium, S., Idelkope, B., et al. (2003). A contradictory GLABRA3 allele helps define gene interactions controlling trichome development in arabidopsis. *Development* 130, 5885–5894. doi: 10.1242/dev.00812

Feng, C., Wang, J., Wu, L., Kong, H., Yang, L., Feng, C., et al. (2020). The genome of a cave plant, *Primulina huaijiensis*, provides insights into adaptation to limestone karst habitats. *New Phytol.* 227, 1249–1263. doi: 10.1111/nph.16588

Fisher, R. A. (1922). On the interpretation of $\chi 2$ from contingency tables, and the Calculation of P. *J. Royal Statist. Soc.* 85, 87–94. doi: 10.2307/2340521

Fu, X., Deng, S., Su, G., Zeng, Q., and Shi, S. (2004). Isolating high-quality RNA from mangroves without liquid nitrogen. *Plant Mol. Biol. Rep.* 22, 197–197. doi: 10.1007/BF02772728

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, FJ., Burton, JN., Walker, BJ., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci.* 108, 1513–1518. doi: 10.1073/pnas.1017351108

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi: 10.1093/bioinformatics/btt086

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Hao, Y., Zhou, Y.-Z., Chen, B., Chen, G.-Z., Wen, Z.-Y., Zhang, D., et al. (2022). The *Melastoma dodecandrum* genome and the evolution of myrtales. *J. Genet. Genomics* 49, 120–131. doi: 10.1016/j.jgg.2021.10.004

Hegebarth, D., Buschhaus, C., Wu, M., Bird, D., and Jetter, R. (2016). The composition of surface wax on trichomes of *Arabidopsis thaliana* differs from wax on other epidermal cells. *Plant J.* 88, 762–774. doi: 10.1111/tpj.13294

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2017). UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522. doi: 10.1093/molbev/msx281

Hülskamp, M. (2004). Plant trichomes: a model for cell differentiation. *Nat. Rev. Mol. Cell Biol.* 5, 471–480. doi: 10.1038/nrm1404

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J. E., Mckain, M. R., Mcneal, J., et al. (2012). A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* 13, R3. doi: 10.1186/gb-2012-13-1-r3

Jin, J., Tian, F., Yang, D.-C., Meng, Y.-Q., Kong, L., Luo, J., et al. (2016). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* 45, gkw982. doi: 10.1093/nar/gkw982

Johnson, C. S., Kolevski, B., and Smyth, D. R. (2002). TRANSPARENT TESTA GLABRA2, a trichome and seed coat development gene of *Arabidopsis*, encodes a WRKY transcription factor. *Plant Cell* 14, 1359–1375. doi: 10.1105/tpc.001404

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Kalvari, I., Nawrocki, E. P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., et al. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* 49, D192–D200. doi: 10.1093/nar/gkaa1047

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Kang, J.-H., Liu, G., Shi, F., Jones, A. D., Beaudry, R. M., Howe, G. A., et al. (2010). The tomato odorless-2 mutant is defective in trichome-based production of diverse specialized metabolites and broad-spectrum resistance to insect herbivores. *Plant Physiol.* 154, 262–272. doi: 10.1104/pp.110.160192

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915. doi: 10.1038/s41587-019-0201-4

Krzywinski, M. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi: 10.1101/gr.092759.109

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 1303, 3997. doi: 10.48550/arXiv.1303.3997

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293. doi: 10.1126/science.1181369

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009a). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, S. F., Milliken, O. N., Pham, H., Seyit, R., Napoli, R., Preston, J., et al. (2009b). The *Arabidopsis* MYB5 transcription factor regulates mucilage synthesis, seed coat development, and trichome morphogenesis. *Plant Cell* 21, 72–89. doi: 10.1105/tpc.108.063503

Li, Z., Defoort, J., Tasdighian, S., Maere, S., Van de Peer, Y., De Smet, R., et al. (2016). Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* 28, 326–344. doi: 10.1105/tpc.15.00877

Li, Z., McKibben, M. T.W., Finch, G. S., Blischak, P. D., Sutherland, B. L., Barker 2021, M. S., et al. (2021). Patterns and processes of diploidization in land plants. *Annu. Rev. Plant Biol.* 72, 387–410. doi: 10.1146/annurev-arplant-050718-100344

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Liu, T., Chen, Y., Chao, L., Wang, S., Wu, W., Dai, S., et al. (2014). Extensive hybridization and introgression between *Melastoma candidum* and m. *sanguineum*. *PloS One* 9, e96680. doi: 10.1371/journal.pone.0096680

Lloyd, A. M., Walbot, V., and Davis, R. W. (1992). *Arabidopsis* and *Nicotiana* anthocyanin production activated by maize regulators *R* and *C1*. *Science* 258, 1773–1775. doi: 10.1126/science.1465611

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* 1, 2047–2217X-2041-2018. doi: 10.1186/2047-217X-1-18

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Mendes, F. K., Vanderpool, D., Fulton, B., and Hahn, M. W. (2020). CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* 36, 5516–5518. doi: 10.1093/bioinformatics/btaa1022

Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., Grimwood, J., et al. (2014). The genome of *Eucalyptus grandis*. *Nature* 510, 356–362. doi: 10.1038/nature13308

Nawrocki, E. P., and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29, 2933–2935. doi: 10.1093/bioinformatics/btt509

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2014). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Ng, W. L., Wu, W., Zou, P., and Zhou, R. (2019). Comparative transcriptomics sheds light on differential adaptation and species diversification between two melastoma species and their f-1 hybrid. *AoB Plants* 11, 1–14. doi: 10.1093/aobpla/plz019

Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2015). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 32, 292–294. doi: 10.1093/bioinformatics/btv566

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Res.* 46, e126. doi: 10.1093/nar/gky730

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R.A., Hellinga, A. J., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 20, 275. doi: 10.1186/s13059-019-1905-y

Ou, S., and Jiang, N. (2018). LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 176, 1410–1422. doi: 10.1104/pp.17.01310

Patel, R. K., and Jain, M. (2012). NGS QC toolkit: a toolkit for quality control of next generation sequencing data. *PloS One* 7, e30619. doi: 10.1371/journal.pone.0030619

Payne, T., Clement, J., Arnold, D., and Lloyd, A. (1999). Heterologous myb genes distinct from GL1 enhance trichome production when overexpressed in *Nicotiana tabacum*. *Development* 126, 671–682. doi: 10.1242/dev.126.4.671

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rakha, M., Bouba, N., Ramasamy, S., Regnard, J.-L., and Hanson, P. (2017). Evaluation of wild tomato accessions (*Solanum* spp.) for resistance to two-spotted spider mite (*Tetranychus urticae* Koch) based on trichome type and acylsugar content. *Genet. Resour. Crop Evol.* 64, 1011–1022. doi: 10.1007/s10722-016-0421-0

Renner, S. S., and Meyer, K. (2001). Melastomeae come full circle: biogeographic reconstruction and molecular clock dating. *Evolution* 55, 1315–1324. doi: 10.1111/j.0014-3820.2001.tb00654.x

Riddick, E. W., and Simmons, A. M. (2014). Do plant trichomes cause more harm than good to predatory insects? *Pest Manage. Sci.* 70, 1655–1665. doi: 10.1002/ps.3772

Salmela, L., and Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics* 30, 3506–3514. doi: 10.1093/bioinformatics/btu538

Schluter, D. (2000). *The ecology of adaptive radiation* (Oxford, United Kingdom: Oxford University Press).

Schwacke, R., Ponce-Soto, G. Y., Krause, K., Bolger, A. M., Arsova, B., Hallab, A., et al. (2019). MapMan4: a refined protein classification and annotation framework applicable to multi-omics data analysis. *Mol. Plant* 12, 879–892. doi: 10.1016/j.molp.2019.01.003

Serna, L., and Martin, C. (2006). Trichomes: different regulatory networks lead to convergent structures. *Trends Plant Sci.* 11, 274–280. doi: 10.1016/j.tplants.2006.04.008

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33, W465–W467. doi: 10.1093/nar/gki458

Stankowski, S., and Streisfeld, M. A. (2015). Introgressive hybridization facilitates adaptive divergence in a recent radiation of monkeyflowers. *Proc. R. Soc. B: Biol. Sci.* 282, 20151666. doi: 10.1098/rspb.2015.1666

Sun, P., Jiao, B., Yang, Y., Shan, L., Li, T., Li, X., et al. (2022). WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant* 15, 1–11. doi: 10.1016/j.molp.2022.10.018

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, W609–W612. doi: 10.1093/nar/gkl315

Szymanski, D. B., Lloyd, A. M., and Marks, M. D. (2000). Progress in the molecular genetic analysis of trichome initiation and morphogenesis in *Arabidopsis*. *Trends Plant Sci.* 5, 214–219. doi: 10.1016/S1360-1385(00)01597-1

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., Paterson, A. H., et al. (2008). Synteny and collinearity in plant genomes. *Science* 320, 486–488. doi: 10.1126/science.1153917

Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J., and Gao, G. (2019). PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res.* 48, D1104–D1113. doi: 10.1093/nar/gkz1020

Törönen, P., Medlar, A., and Holm, L. (2018). PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res.* 46, W84–w88. doi: 10.1093/nar/gky350

Veranso-Libalah, M. C., Stone, R. D., Fongod, A. G. N., Couvreur, T. L. P., and Kadereit, G. (2017). Phylogeny and systematics of African melastomateae (Melastomataceae). *Taxon* 66, 584–614. doi: 10.12705/663.5

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. doi: 10.1093/bioinformatics/btx153

Wang, X., Shen, C., Meng, P., Tan, G., and Lv, L. (2021). Analysis and review of trichomes in plants. *BMC Plant Biol.* 21, 70. doi: 10.1186/s12870-021-02840-x

Wang, Z., Yang, Z., and Li, F. (2019). Updates on molecular mechanisms in the development of branched trichome in *Arabidopsis* and nonbranched in cotton. *Plant Biotechnol. J.* 17, 1706–1722. doi: 10.1111/pbi.13167

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinf.* 8, 77–80. doi: 10.1016/S1672-0229(10)60008-3

Wong, K. M. (2016). *The genus melastoma in Borneo: including 31 new species* (Kota Kinabalu, Sabah, Malaysia: Natural History Publications).

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* 2, 100141. doi: 10.1016/j.xinn.2021.100141

Wu, S., Han, B., and Jiao, Y. (2020). Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Mol. Plant* 13, 59–71. doi: 10.1016/j.molp.2019.10.012

Xu, H., Luo, X., Qian, J., Pang, X., Song, J., Qian, G., et al. (2012). FastUniq: a fast *de novo* duplicates removal tool for paired short reads. *PloS One* 7, e52249. doi: 10.1371/journal.pone.0052249

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yang, F. S., Nie, S., Liu, H., Shi, T. L., Tian, X. C., Zhou, S. S., et al. (2020). Chromosome-level genome assembly of a parent species of widely cultivated azaleas. *Nat. Commun.* 11, 5269. doi: 10.1038/s41467-020-18771-4

Yang, C., and Ye, Z. (2013). Trichomes as models for studying plant cell differentiation. *Cell. Mol. Life Sci.* 70, 1937–1948. doi: 10.1007/s00018-012-1147-6

Zhang, F., Zuo, K., Zhang, J., Liu, X., Zhang, L., Sun, X., et al. (2010). An L1 box binding protein, GbML1, interacts with GbMYB25 to control cotton fibre development. *J. Exp. Bot.* 61, 3599–3613. doi: 10.1093/jxb/erq173

Zhang, Z., Xiao, J., Wu, J., Zhang, H., Liu, G., Wang, X., et al. (2012). ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* 419, 779–781. doi: 10.1016/j.bbrc.2012.02.101

Zhao, Q., and Chen, X.-Y. (2016). Development: A new function of plant trichomes. *Nat. Plants* 2, 16096. doi: 10.1038/nplants.2016.96

Zhao, H., Wang, X., Zhu, D., Cui, S., Li, X., Cao, Y., et al. (2012). A single amino acid substitution in IIIf subfamily of basic helix-loop-helix transcription factor AtMYC1 leads to trichome and root hair patterning defects by abolishing its interaction with partner proteins in *Arabidopsis*. *J. Biol. Chem.* 287, 14109–14121. doi: 10.1074/jbc.M111.280735

Zhou, L. H., Liu, S. B., Wang, P. F., Lu, T. J., Xu, F., Genin, G. M., et al. (2017). The *Arabidopsis* trichome is an active mechanosensory switch. *Plant Cell Environ.* 40, 611–621. doi: 10.1111/pce.12728

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 08 March 2023
DOI 10.3389/fpls.2023.1122549

OPEN ACCESS

EDITED BY
Guanjing Hu,
Agricultural Genomics Institute at Shenzhen (CAAS), China

REVIEWED BY
Qi Wu,
Chengdu University, China
Xiwen Li,
Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, China

*CORRESPONDENCE
Jinxing Lin
linjx@bjfu.edu.cn
Jian-Feng Mao
jianfeng.mao@bjfu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 13 December 2022
ACCEPTED 10 February 2023
PUBLISHED 08 March 2023

CITATION
Tian X-C, Guo J-F, Yan X-M, Shi T-L, Nie S, Zhao S-W, Bao Y-T, Li Z-C, Kong L, Su G-J, Mao J-F and Lin JX (2023) Unique gene duplications and conserved microsynteny potentially associated with resistance to wood decay in the Lauraceae. *Front. Plant Sci.* 14:1122549. doi: 10.3389/fpls.2023.1122549

Unique gene duplications and conserved microsynteny potentially associated with resistance to wood decay in the Lauraceae

Xue-Chan Tian[1†], Jing-Fang Guo[1†], Xue-Mei Yan[1], Tian-Le Shi[1], Shuai Nie[1], Shi-Wei Zhao[1], Yu-Tao Bao[1], Zhi-Chao Li[1], Lei Kong[1], Guang-Ju Su[2], Jian-Feng Mao[1,3*] and Jinxing Lin[1*]

[1]National Engineering Research Center of Tree Breeding and Ecological Restoration, State Key Laboratory of Tree Genetics and Breeding, Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants, Ministry of Education, College of Biological Sciences and Technology, Beijing Forestry University, Beijing, China, [2]National Tree Breeding Station for Nanmu in Zhuxi, Forest Farm of Zhuxi County, Hubei, China, [3]Department of Plant Physiology, Umeå Plant Science Centre, Umeå University, Umeå, Sweden

Wood decay resistance (WDR) is marking the value of wood utilization. Many trees of the Lauraceae have exceptional WDR, as evidenced by their use in ancient royal palace buildings in China. However, the genetics of WDR remain elusive. Here, through comparative genomics, we revealed the unique characteristics related to the high WDR in Lauraceae trees. We present a 1.27-Gb chromosome-level assembly for *Lindera megaphylla* (Lauraceae). Comparative genomics integrating major groups of angiosperm revealed Lauraceae species have extensively shared gene microsynteny associated with the biosynthesis of specialized metabolites such as isoquinoline alkaloids, flavonoid, lignins and terpenoid, which play significant roles in WDR. In Lauraceae genomes, tandem and proximal duplications (TD/PD) significantly expanded the coding space of key enzymes of biosynthesis pathways related to WDR, which may enhance the decay resistance of wood by increasing the accumulation of these compounds. Among Lauraceae species, genes of WDR-related biosynthesis pathways showed remarkable expansion by TD/PD and conveyed unique and conserved motifs in their promoter and protein sequences, suggesting conserved gene collinearity, gene expansion and gene regulation supporting the high WDR. Our study thus reveals genomic profiles related to biochemical transitions among major plant groups and the genomic basis of WDR in the Lauraceae.

KEYWORDS

Lauraceae, *Lindera megaphylla*, wood decay resistance (WDR), tandem and proximal duplications (TD/PD), gene microsynteny

Introduction

Wood is an exceptionally useful biomaterial, with myriad uses in construction, pulp and paper, and as a biofuel. Moreover, wood is a renewable material. One problem with using wood as a renewable biomaterial is that many microbes and insects have evolved to use wood as an energy source, producing enzymes that break down the components of the wood. Some species have evolved mechanisms to resist microbial damage and oxidation; many species with high wood decay resistance (WDR), such as teak (*Tectona grandis*), redwood (*Sequoia sempervirens*), and mahogany (*Swietenia mahagoni*) are rare and extremely valuable. Therefore, understanding the genetic basis and molecular mechanisms of WDR has the potential to provide effective information for improving WDR in commercially grown tree species. Wood is mainly composed of cellulose, hemicellulose, and lignin, which provide structural support for trees and resistance to microbial attack (Nascimento et al., 2013). Generally, lignin, a phenolic compound that is extremely resistant to degradation by certain fungi and plant diseases, acts as the basal component of wood durability by covering and protecting cellulose (Vance et al., 1980; Mounguengui et al., 2016). Further, trees resistant to decay exhibit significant production or accumulation of some bioactive compounds that function as antifungal compounds, antioxidants, or insect antifeedants, and are the main factors contributing to WDR (Nascimento et al., 2013). WDR is influenced by alkaloids such as indols and beta-carboline alkaloids, which have strong antifungal activity, as well as berberine and palmatine, which have shown good antifeedant and antioxidant activities (Kawaguchi et al., 1989; Anouhe et al., 2018; Ekeuku et al., 2020; Imenshahidi and Hosseinzadeh, 2020). Moreover, flavonoids are phenolic compounds with strong fungicidal activity, natural antioxidants and are excellent free radical scavengers, which have a significant effect on improving WDR (Schultz and Nicholas, 2000). In addition, terpenoids, including triterpenoids, diterpenoids, sesquiterpenoids, and monoterpenoids, have important antifungal, antifeeding, and antioxidant abilities, and contribute greatly to WDR (Park et al., 2000; Isman, 2002).

Lauraceae, a family of the order Laurales in the Magnoliids, includes about 67 genera and over 2,500 species (Anouhe et al., 2018). Lauraceae species are economically important, playing important roles in timber production, medicine, spice production, and ecological afforestation (Anouhe et al., 2018). A distinguishing feature of most Lauraceae species is the extremely high decay resistance of wood, including resistance to fungi, insect erosion, and oxidation (Jagels et al., 2005). Nanmu species, a group of tree species belonging to the Lauraceae family, are characterized by their straight trunks, fragrant and dense wood, and most notably by their super WDR (Jiao et al., 2022). Given these valuable traits, Nanmu wood is a precious natural resource that has historically been exploited, for example, for the construction of royal palaces (Xie et al., 2015). Generally, most species of the *Phoebe* and *Machilu* genera are recognized as Nanmu (e.g., *Phoebe zhennan* and *Machilu nanmu*) (Jiao et al., 2022). Another tree, *Lindera megaphylla*, has all superior qualities of the generally accepted Nanmu species, and was extensively used for the construction of royal buildings in Beijing in the Qing dynasty (Figure S1). *L. megaphylla* accumulates a variety of alkaloids (Chou et al., 1994) that promote resistance to microbial infection and herbivore attack, increasing the antifeeding and antioxidant activities of its wood (Kawaguchi et al., 1989; Ekeuku et al., 2020). *L. megaphylla* also has a wide range of medicinal properties due to alkaloid accumulation (Cao et al., 2016). In addition, the wood of some other Lauraceae species, e.g., *Cinnamomum* (Zhou et al., 2019) and *Litsea* species, have good natural durability and are highly valuable in construction, furniture, sculpture, and other building applications. With the development of society, there is increasing demand for naturally durable wood. However, genetic studies on the natural durability of wood, especially of Lauraceae species, are limited. Therefore, it is of great significance to identify the genes of biosynthetic pathways related to WDR, to investigate whether the WDR-related gene families have expanded significantly, and to reveal whether there are unique and conserved characteristics of WDR-related genes in Lauraleae species.

The phylogenetic location of Magnoliids remains to be further clarified. *Lindera megaphylla* belongs to Lauraceae, which together with Canellales, Piperales, and Magnoliales, constitutes the Magnoliids, including 9,000 species (The Angiosperm Phylogeny Group et al., 2016). Although multiple genomes of Magnoliids have been published, the relationship between magnoliids, eudicots, and monocots remains discordant. For example, the gene sequence-based phylogenomic analyses of *Liriodendron chinense* (Chen et al., 2019), *Piper nigrum* (Hu et al., 2019), *Persea americana* (Rendon-Anaya et al., 2019) and *Phoebe bournei* (Chen et al., 2020a) supported the Magnoliids as sister to the monocots-eudicots clade, while analyses of *Cinnamomum kanehirae* (Chaw et al., 2019a), *Chimonanthus salicifolius* (Lv et al., 2020) and *Chimonanthus praecox* (Shang et al., 2020a) supported Magnoliids as sister clade of eudicots. In addition, the phylogenomic analyses of *Litsea cubeba* suggested that the definite evolutionary relationships between Magnoliids, monocots, and eudicots remains to be resolved due to the possibility of incomplete lineage sorting (ILS) (Chen et al., 2020b). Microsynteny, gene colocality or collinearity, is the local conservation of gene order or gene neighborhood. Microsynteny provides valuable information to infer gene and genome evolution (Bowers et al., 2003; Van de Peer, 2004; Dewey, 2011), and is significant in phylogenetic inferences (Zhao and Schranz, 2019; Zhao et al., 2021c).

Here, we generated a chromosome-level genome assembly of *L. megaphylla* with long-read sequencing and Hi-C scaffolding technologies. The wood of *L. megaphylla* is dense and durable, making it an ideal material for construction, furniture, and shipbuilding. We conducted phylogenomic reconstruction of main angiosperm groups based on multiple strategies of concatenation, coalescent-based, and network-based microsynteny. Further, through the comparative genomics, especially shared gene microsynteny among major angiosperm lineages, we identified unique gene duplications and conserved microsynteny associated with isoquinoline alkaloids (IA),

flavonoids, lignin, and terpenoids biosynthesis in Lauraceae species, which may be associated with outstanding wood durability in Lauraceae trees. The genome resources and findings presented here provide a basis for further evolutionary or functional studies in Lauraceae species, and for additional exploration of Lauraceae wood decay resistance.

Results

L. megaphylla genome sequencing, assembly, and gene annotation

As a first step to understand genomics of WDR in Lauraceae species with significant WDR, we sequenced the genome *L. megaphylla*. According to *k*-mer analysis, the genome size of *L. megaphylla* (Figure S1) was estimated to be ~1.3 Gb, with a 0.5% heterozygosity rate (Figure S2 and Note S1 for details). We generated 178.78 Gb (10.3 million reads, roughly 130× coverage) of Oxford Nanopore Technologies (ONT) long reads (Table S1) for primary assembly, 160.28 Gb (1068 million reads, 120× coverage, PCR-free library) of Illumina paired-end reads for correction and polishing, and 223.23 Gb (1488.194 million reads, 170× coverage) of Hi-C paired-end reads for scaffolding (Figure S3 and Table S1). A final genome assembly of 1.27 Gb was obtained, which consisted of 486 scaffolds, including 12 chromosome-level scaffolds, with a scaffold N50 of 104 Mb (Figure 1; Table 1; Tables S2, 3). The high confidence of the genome assembly was supported by high ten-fold minimum genome coverages of 95.1% (Illumina) and 99.6% (ONT), as well as the high mapping rates of 99.2% (Illumina) and 81.3% (ONT) reads. A 90.7% (1,306 complete genes) Benchmarking Universal Single Copy Orthologs (BUSCO) recovery score (Simão

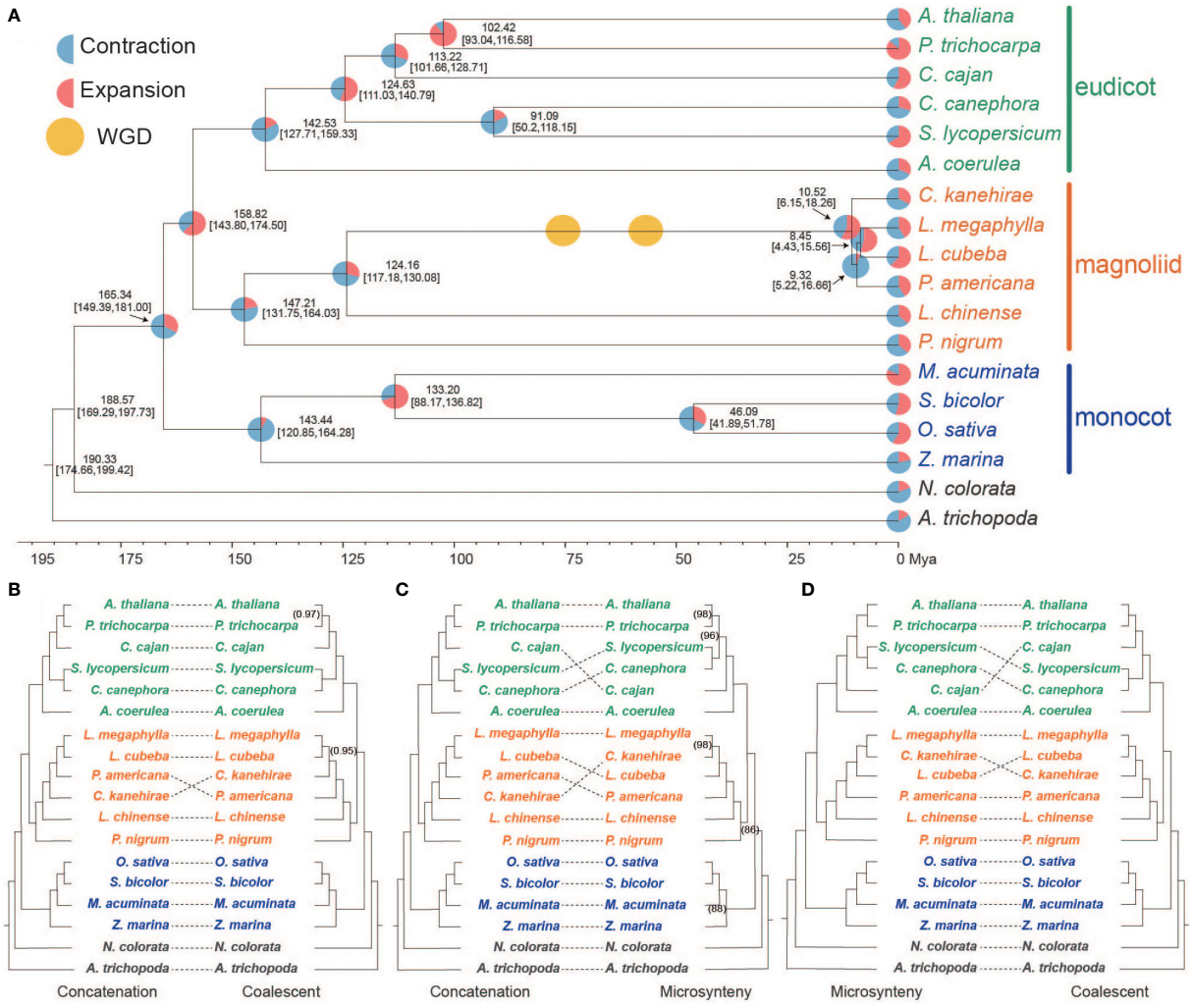


FIGURE 1
Phylogenomic analysis of three major angiosperm groups. **(A)** Phylogenetic tree of 18 plant species generated by the concatenation-based method. Pie charts indicate the predicted expansion (red) and contraction (blue) of the gene family. The numbers represent divergence time of each node (Mya, million years ago), and values in brackets are 95% confidence intervals for the time of divergence. The yellow circle shows the WGD events identified in Lauraceae species. **(B)** Comparison of phylogenetic trees produced by the concatenation- and multi-species coalescent (MSC)-based methods. **(C)** Comparison of phylogenetic trees generated using concatenation- and microsynteny-based methods. **(D)** Comparison of phylogenetic trees produced using the microsynteny- and MSC-based methods.

TABLE 1 Statistics of the *Lindera megaphylla* genome assembly and annotation.

Sequencing	
Raw bases of WGS-ONT Sequel (Gb)	178.78
Raw bases of WGS-Illumina (Gb)	160.28
Raw bases of Hi-C (Gb)	223.23
Raw bases of mRNAseq (Gb)	144.90
Assembly	
Genome size (Mb)	1,268.60
Number of scaffolds	486
N50 of scaffolds (bp)	104,721,408
L50 of scaffolds	5
Chromosome-scale scaffolds (bp)	1,206,404,078 (95.10%)
Number of contigs	1,407
N50 of contigs (bp)	2,612,587
L50 of contigs	125
Number of Gap	921
BUSCO (genome)	90.70%
GC content of the genome (%)	39.44%
Annotation	
Number of predicted genes	34,216
Number of predicted protein-coding genes	32,586
Average gene length (bp)	7,693.79
Average CDS length (bp)	1,250.89
Average exon per transcript	5.22
Number of tRNAs	579
Number of rRNAs	248
Repeat sequences (bp)	849,656,470 (66.98%)
BUSCO (gene set)	91.70%

et al., 2015) and a high LTR Assembly Index (LAI) (Ou et al., 2018) score of 12.40 revealed a high completeness in the final assembly (Table S4).

A total of 32,586 protein-coding genes were predicted from the final assembly (Table S5). The average lengths of total gene regions, transcripts, coding sequences, exons, and introns were 7,693.8, 1,410.1, 1,250.9, 270, and 1,094.7 bp, respectively (Table S5). In addition, we annotated 579 tRNAs, 248 rRNAs (including five 28S, six 18S, and 237 5S rRNAs), and 803 other non-coding RNAs (Table S6). The strongly supported gene annotation was evidenced by a 91.7% complete BUSCO score, as well as by 85.9% of the predicted genes (29,400 genes) with an annotation edit distance (AED) lower than 0.5 (Table S4 and Figure S4). More results of genome annotation are available in Note S4 and Table S7.

We identified 34,888 gene families, of which 6,340 are shared among all 18 species (Table S8) (see "Methods" section for details). And 885 expansion gene families in Lauraceae were enriched in isoquinoline alkaloid biosynthesis, flavonol biosynthesis, phenylpropanoid catabolism, lignin catabolic processes, and sulfur compound transport (Figure S5). All of these processes are tightly associated with resistance to bacteria and fungi, insect attacks, and high wood durability. The expanded gene families in *L. megaphylla* were also enriched in isoquinoline alkaloid biosynthesis, positive regulation of flavonoid biosynthesis, and isoflavone 7-O-glucosyltransferase activity (Figure S6). Similarly, these processes are all tightly associated with wood decay resistance.

Results of transposable element and other repeat annotation are available in Note S5, Figure S7 and Tables S9, S10.

Phylogenetic placement of Magnoliids

To determine the phylogenetic position of the Magnoliids relative to monocots and eudicots, phylogenetic trees were constructed using three distinct methods (concatenation-, coalescent-, and microsynteny-based approaches). For the concatenation-based approach, we constructed a phylogenetic tree using 885 low-copy orthologs from 18 species, with *Amborella trichopoda* and *Nymphaea colorata* as the outgroup (Figure 1A) (see "Methods" section). Results showed that the Maximum likelihood (ML) trees placed the Magnoliids as sister to the eudicots (Figure 1A). Phylogenetic analysis indicated that divergence time between Magnoliids and eudicots was 158.8 million years ago (Mya), with 95% confidence intervals of 143.8-174.5 Mya (Figure 1A), which overlaps with the *C. kanehirae* genome (136-209 Mya) (Chaw et al., 2019a). Lauraceae divergence was 124.16 Mya (Figure 1A), which was approximately equal to *Phoebe bournei* (Chen et al., 2020a). In addition, *L. megaphylla* diverged from *C. kanehirae* and *L. cubeba* around 10.52 Mya and 8.45 Mya, respectively (Figure 1A).

To reduce the influence of incomplete lineage sorting (ILS) on the determination of phylogenetic position, we also performed coalescent-based analyses of gene trees from the 855 low-copy gene families with ASTRAL-Pro (version 1.1.2) (Zhang et al., 2020a). The result from the coalescent-based analysis with strongly supported topology was highly consistent with the results of the concatenation-based method, placing Magnoliids as a sister group to eudicots after their divergence from monocots (Figure 1B). In addition, to reduce the interference caused by gene duplication and loss, ancestor hybridization, and lateral gene transfer in the homology assessment of plants, a novel method for phylogenetic tree reconstruction based on genome-wide synteny network data has been proposed (Zhao and Schranz, 2019; Zhao et al., 2021c). This method, microsynteny or gene order conservation, has been considered to be a valuable and alternative phylogenetic character in addition to sequence-based characters (Zhao et al., 2021c). The microsynteny-based analysis results confirmed that Magnoliids and eudicots are sister groups, which was topologically identical to the results of the above two methods (Figures 1C, D). These results strongly support that Magnoliids and eudicots are sister branches of monocots.

10.3389/fpls.2023.1122549

Microsynteny sharing and functional implications

To examine the lineage-specific microsynteny profile of major plant groups (Magnoliids, monocots, and eudicots), the genome synteny cluster obtained from microsynteny-based analysis of 16 species excluding *N. colorata* and *A. trichopoda* was analyzed. Interestingly, the Lauraceae species *L. megaphylla*, *L. cubeba*, and *C. kanehirae* had the largest number of microsyntenic clusters, with 15,347, 14,879, and 14,830 from each species, respectively (Figure S8A). The number of microsyntenic clusters shared by Magnoliids-eudicots (3,840) was significantly more than that shared by Magnoliids-monocots (871) and eudicots-monocots (491) (Figure S8B). Based on the heatmap of correlation in shared microsynteny, we observed a strong correlation between Magnoliids and eudicots (see "Methods" section) (Figure 2A). In contrast, the monocots showed a weak correlation with the other two clades, especially *S. bicolor* and *O. sativa*, which belong to the Poaceae (Figure 2A). These data signified a closer relationship between Magnoliids and eudicots.

Next, we examined the functional implications of the shared or group-specific microsyntenic clusters among the three clades by removing the species-specific cluster (see "Methods" section). We discovered 2,839, 1,758, and 1,208 clusters specific to Magnoliids, eudicots, and monocots, respectively (Figure 2B). The number of synteny clusters common to Magnoliids-eudicots was still the largest (358), followed by Magnoliids-monocots (54), and eudicots-monocots (37) (Figures 2B–E, S9). As revealed in the UpSet plot, the Poaceae species *Sorghum bicolor* and *Oryza sativa* shared the largest number of clusters (6,283), followed by *Piper nigrum* and *Musa acuminata* with 3,243 and 1,651 species-specific clusters, respectively (Figure S8C). Four Lauraceae species also shared many clusters (1,460) (Figure S8C). Excluding these species-specific and clade-specific clusters, the six Magnoliids and six eudicot species shared the most clusters (39) (Figure S8C). These results further supported that Magnoliids and eudicots may be most closely related.

Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) enrichment analyses of eudicot-specific microsyntenic clusters showed that they were mainly associated with terms related to a series of signaling pathways (Figure S10). Clusters specific to the Magnoliids were mainly enriched in terms such as isoquinoline alkaloid biosynthesis, ribosome biogenesis, brassinosteroid biosynthetic process, phospholipid biosynthetic process, and secondary metabolite biosynthesis (Figure S10). The microsyntenic clusters in monocots were mainly enriched in terms such as histidine metabolism, chloroalkane limonene and pinene degradation, cell plate assembly, and pyrimidine metabolism

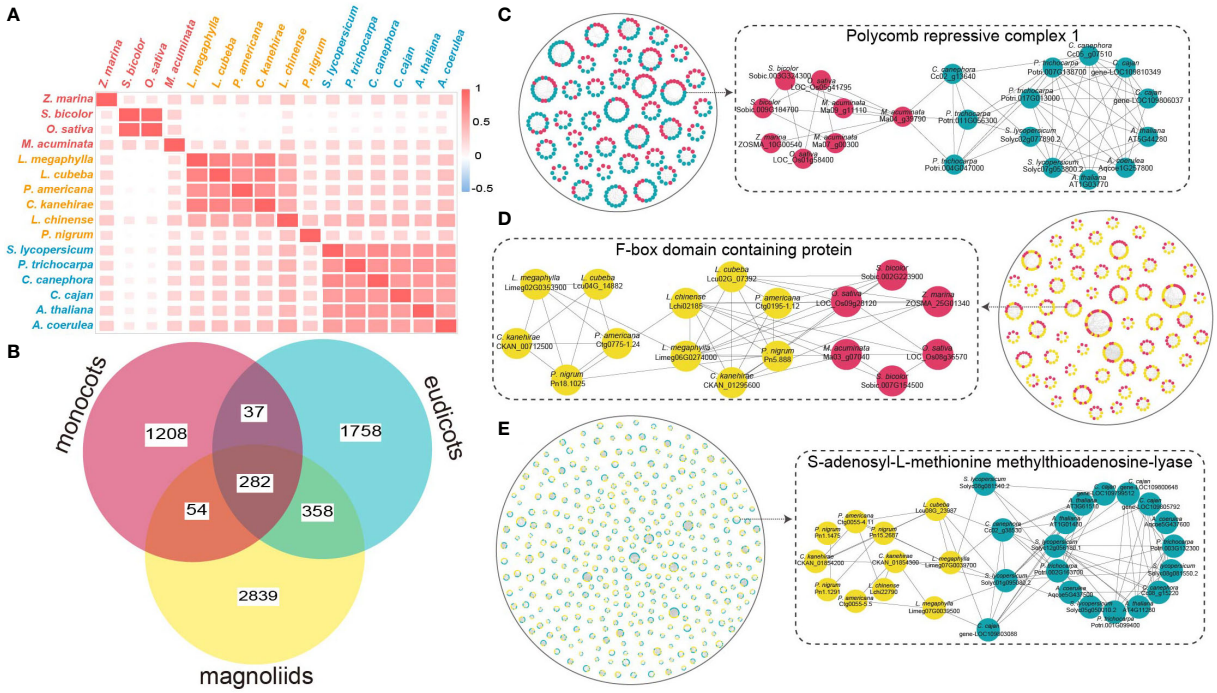


FIGURE 2
Analysis of gene microsyntenic clusters. **(A)** Heatmap of the number of microsyntenic clusters shared among Magnoliid, eudicot, and monocot species. **(B)** Venn diagram showing the microsyntenic clusters shared among Magnoliids, eudicots, and monocots. **(C)** The black solid circle on the left surrounds microsyntenic clusters shared by eudicots and monocots, where blue dots represent eudicots, and red dots represent monocots. The black dotted rectangle on the right highlights an example of a cluster (in a subnetwork) shared between eudicots and monocots. **(D)** The black solid circle on the right surrounds microsyntenic clusters shared by Magnoliids and monocots, where yellow dots represent Magnoliids and red dots represent monocots. The black dotted rectangle on the left highlights an example of a cluster (in a subnetwork) shared between Magnoliids and monocots. **(E)** The black solid circle on the left surrounds microsyntenic clusters shared by Magnoliids and eudicots, where yellow dots represent Magnoliids and blue dots represent eudicots. The black dotted rectangle on the right highlights an example of a cluster (in a subnetwork) shared between Magnoliids and eudicots.

(Figure S10). Remarkably, the synteny clusters shared by Magnoliids, eudicots, and monocots were significantly enriched in sesquiterpenoid, diterpenoid, and triterpenoid biosynthesis (Figure S10). This finding indicates that the genes involved in terpenoid biosynthesis are conserved among plant clades, indicating the importance of terpenoids in various plants. In addition, the unique clusters in Lauraceae were mainly enriched in isoquinoline alkaloid biosynthesis, phenylpropanoid metabolic process, secondary metabolic process and lignin metabolic process, revealing potential links to the super WDR of Lauraceae trees (Figure S11).

Inference of whole-genome duplication in Lauraceae species are available in Note S7 and Figures S12-S14.

Tandem duplicate/proximal duplicate gene duplications in Lauraceae

A total of 28,838, 22,618 and 25,951 duplicated genes originating from whole-genome duplicates (WGD), tandem duplicates (TD), proximal duplicates (PD), dispersed duplicates (DSD), and transposed duplicates (TRD) were annotated in *L. megaphylla*, *C. kanehirae*, and *L. cubeba*, respectively (Figure 3A and Table S11). Aside from 18.39% TD/PD genes in *Aquilegia coerulea*, high TD/PD ratios were found in the Lauraceae species *L. megaphylla* (19.97%), *L. cubeba* (18.36%), *C. kanehirae* (22.45%), and *P. bournei* (18.44%) (Figure 3A and Table S12). Gene families expanded *via* TD/PD duplications in Lauraceae were functionally enriched in GO categories significantly associated with wood decay resistance, such as lignin catabolism, isoquinoline alkaloid biosynthesis, flavonol biosynthesis, and phenylpropanoid catabolism (Figure 3B). KEGG enrichment confirmed this pattern, showing that TD/PD duplications were enriched in isoquinoline alkaloid biosynthesis, flavone and flavonol biosynthesis, phenylpropanoid and flavonoid biosynthesis, monoterpenoid biosynthesis, antibiotic biosynthesis, defense response to bacterium, response to oxidative stress, cyanoamino acid metabolism, tropane, piperidine and pyridine alkaloid biosynthesis, and sulfur metabolism (Figure 3B). In addition to these functions, KEGG and GO analyses also revealed significant enrichment of TD/PD duplications in the biosynthesis of various terpenoids in *L. megaphylla*, including diterpenoid, monoterpenoid, sesquiterpenoid, and triterpenoid biosynthesis (Figure S15) In summary, these results indicate that local gene duplication in Lauraceae contributed to the expansion of secondary metabolite biosynthesis genes related to WDR.

Genes involved in benzylisoquinoline alkaloid biosynthesis

Three different benzylisoquinoline alkaloid (BIA) biosynthesis pathways were annotated in four Lauraceae species (*L. megaphylla*, *L. cubeba*, *C. kanehirae*, and *P. bournei*), including magnoflorine, berberine, and palmatine biosynthesis pathways (Figure 4A and Table S13), all of which were important for improving decay resistance of wood. The termite antifeeding activities of berberine and palmatine have been well demonstrated (Kawaguchi et al., 1989; Park et al., 2000). A total of twelve gene families related to BIA biosynthesis were identified. The enzymes 4OMT, 6OMT, SOMT, and CoOMT belong to the O-methyltransferase (OMT) family, and CYP80G, CYP80B, and CYP719A belong to the cytochrome P450 (CYP) family. These enzymes are mainly found in Magnoliids and *A. coerulea*, but rarely in monocots and other core eudicots

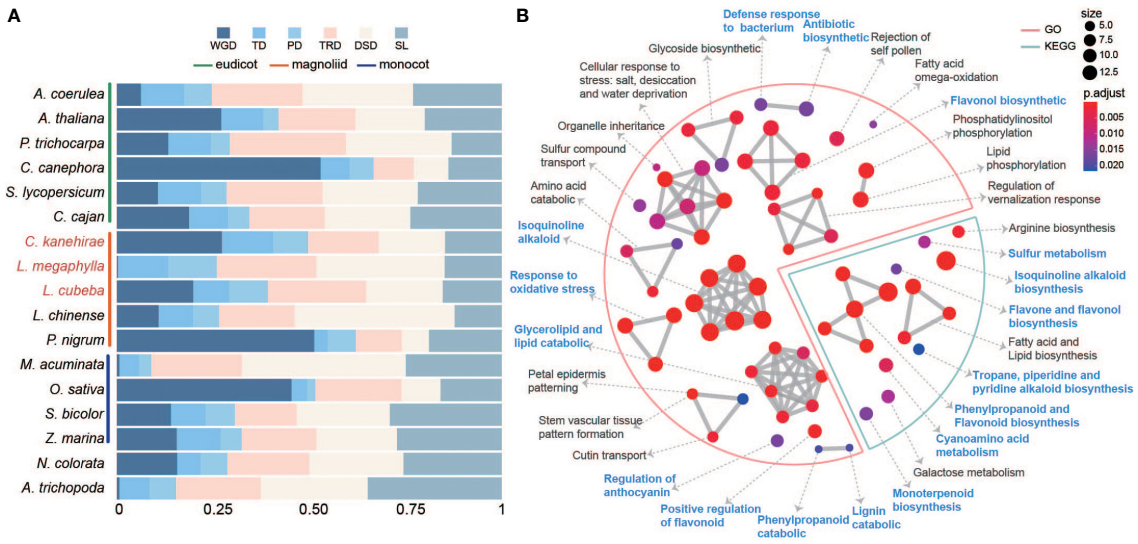


FIGURE 3
The expansion of duplicated genes. **(A)** The stacked bar chart shows the proportion of genes derived from five duplication types (WGD whole-genome duplication, TD tandem duplication, PD proximal duplication, TRD transposed duplication and DSD dispersed duplication). **(B)** GO and KEGG functional enrichment analysis of expanded genes arising from tandem and proximal duplicates (TD/PD) in Lauraceae. The red line represents GO enrichment and the blue line represents KEGG enrichment. Blue letters indicate terms related to wood decay resistance.

10.3389/fpls.2023.1122549

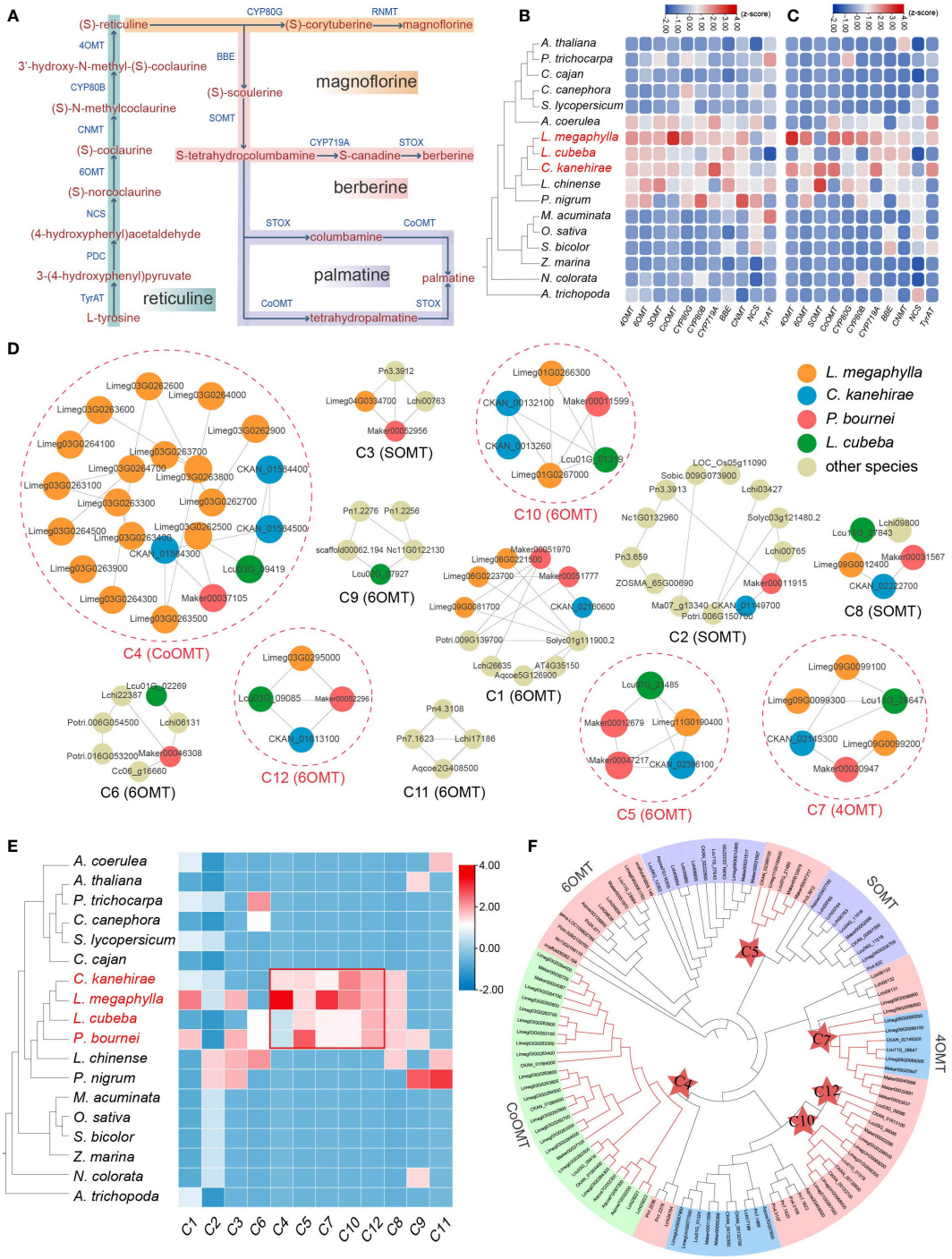


FIGURE 4
Characteristics of benzylisoquinoline alkaloid genes in Lauraceae. **(A)** The biosynthesis pathway of isoquinoline alkaloids. TyrAT, tyrosine aminotransferase; PDC, 4-hydroxyphenylpyruvate decarboxylase; NCS, (*S*)-norcoclaurine synthase; 6OMT, (RS)-norcoclaurine 6-O-methyltransferase; CNMT, (S)-coclaurine-N-methyltransferase; CYP80B, *N*-methylcoclaurine 3′-hydroxylase; 4OMT, 3′-hydroxy-N-methyl-(S)-coclaurine 4′-O-methyltransferase; CYP80G, (S)-corytuberine synthase; RNMT, reticuline N-methyltransferase; BBE, berberine bridge enzyme; SMT, (S)-scoulerine 9-O-methyltransferase; CYP719A, (S)-canadine synthase; THBO, tetrahydroberberine oxidase; CoOMT, columbamine O-methyltransferase. **(B)** Number of annotated genes in each enzyme gene family (*4OMT*, *6OMT*, *SOMT*, *CoOMT*, *CYP80G*, *CYP80B*, *CYP719A*, *BBE*, *NMT*, *NCS*, and *TyrAT*) for each species. **(C)** Proportion of tandem (TD) and proximal (PD) duplication genes in each enzyme gene family (*4OMT*, *6OMT*, *SOMT*, *CoOMT*, *CYP80G*, *CYP80B*, *CYP719A*, *BBE*, *NMT*, *NCS*, and *TyrAT*) for each species. **(D)** Microsyntenic gene clusters associated with subfamilies of the *OMT* gene family (here, *4OMT*, *6OMT*, and *CoOMT*). Circles in dashed red line denote the syntenic clusters (here, C4, C5, C7, C10 and C12) unique to Lauraceae species. **(E)** Heatmap of 12 microsyntenic clusters in **(D)**, five of which are Lauraceae-specific and highlighted by a red square. The color in the heatmap represents the gene number in each cluster for each species. **(F)** Phylogenetic analysis of *OMT* gene families. The red stars represent genes within Lauraceae-specific gene clusters identified in **(D)**.

(Figure 4B). In addition, TD/PD duplication contributed the most BIA biosynthesis genes in Lauraceae, especially the *4OMT*, *6OMT*, *CoOMT*, *CYP80G*, *CYP80B*, *CYP719A*, and *CNMT* ((S)-coclaurine-N-methyltransferase) genes (Figures 4C, S16, and Table S14).

A total of 12 microsyntenic clusters were identified as related to *OMT* gene families (here, *4OMT*, *6OMT*, and *CoOMT*) (Figure 4D). Five of these 12 microsyntenic clusters were specific to Lauraceae, including C5, C10 and C12 associated with *6OMT*, C7 associated with *4OMT*, and C4 with *CoOMT* (Figures 4D–F). These genes on Lauraceae-specific microsyntenic clusters may play an important role in the unique WDR of Lauraceae species.

6OMT is involved in the rate-limiting step of isoquinoline biosynthesis (Robin et al., 2016). Phylogenetic analysis showed that the *6OMT* genes in Lauraceae could be divided into five groups (Figure 5A). Genes of the Lauraceae-specific clusters C5, C10 and C12 were located in groups 2,4 and 5 respectively (Figures 5A–C). Protein sequence analysis found a Lauraceae-specific motif (motif 9) among genes in group 2 (C5). Genes in group 4 (C10) and 5 (C12) shared another Lauraceae-specific motif (motif 12) (Figure 5A). In addition, we identified several conserved motifs unique to Lauraceae through sequence analysis of gene promoters. Motif 1 existed in both C5 and C10 genes and overlapped with the predicted binding sites of bHLH transcription factors (TFs) (Figure 5A). Motif 8 was specific to C5 genes and overlapped with the predicted binding sites of ERF TFs (Figure 5A). Interestingly, six motifs (motif 5, 4, 2, 1, 3, and 6) formed a tandem cluster unique to Lauraceae genes in group 4 (C10). These motifs were the predicted binding sites of GATA, B3, ERF, bHLH, Trihelix, and MYB transcription factors (Figure 5A).

In addition to 6OMT, 4OMT is also an important rate-limiting enzyme in BIA biosynthesis (Inui et al., 2012). The *4OMT* genes in Lauraceae were divided into two groups, with C7 genes located in group 2 (Figures 5D, E). Four Lauraceae-specific and conserved motifs (motifs 10-14) were identified among the protein sequences of these group 2 (C7) genes (Figure 5E). Although no Lauraceae-specific microsyntenic cluster in group 1, phylogenetic analysis results showed that they were located in Lauraceae-specific clades, and two Lauraceae-specific motifs (motif 9 and motif 10) were identified (Figure 5E). Sequence analysis of gene promoters revealed that both groups (group1 and group 2) shared a common

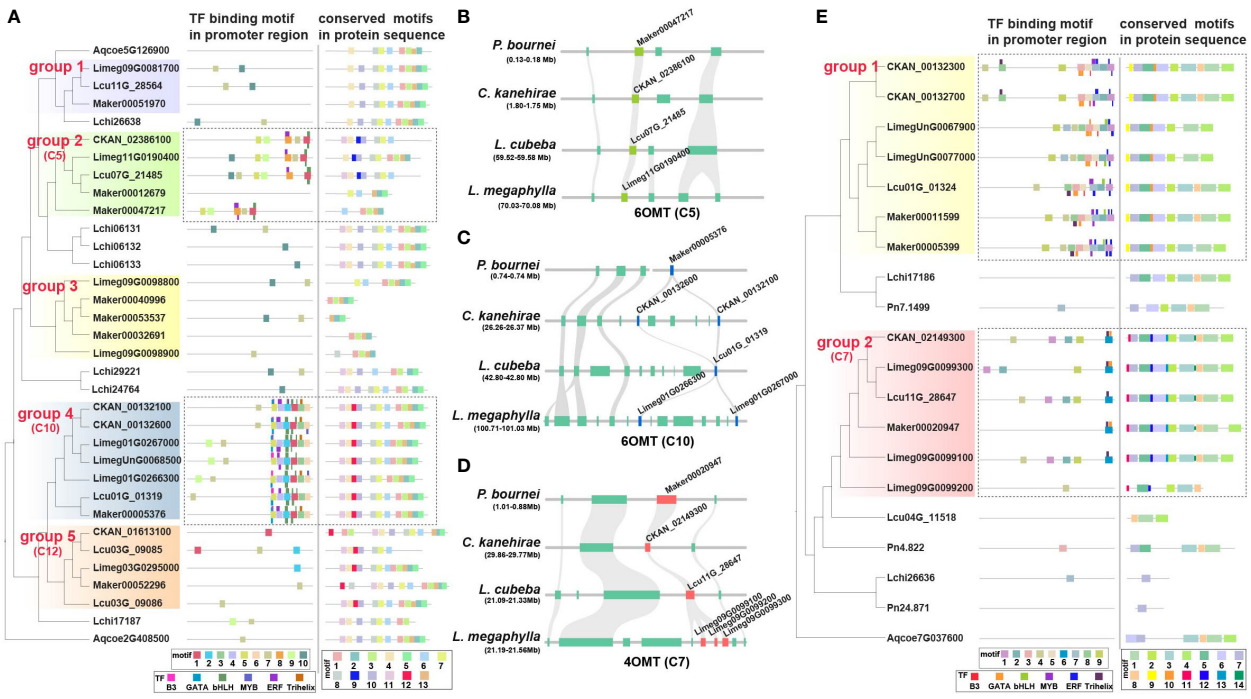


FIGURE 5
Specificity of *6OMT* and *4OMT* genes in Lauraceae. **(A)** The different panels illustrate the phylogenetic tree of the *6OMT* gene family (left), the distribution of motifs in the promoter sequences and the predicted transcription factor binding sites (TFBS) (middle), and the distribution of motifs in protein sequences (right). Thick squares represent motifs and thin ones represent TFBSs. Dashed boxes highlight genes and promoter motifs unique to Lauraceae species. **(B)** The syntenic block containing the *6OMT* gene family within the Lauraceae-specific microsynteny gene cluster (C5), which was identified in Figures 4D, E. This syntenic block was compared among *P. bournei*, *C. kanehirae*, *L. cubeba*, and *L. megaphylla*. Chartreuse squares represent the *6OMT* genes and aquamarine squares represent other genes on the syntenic block. **(C)** The syntenic block containing the *6OMT* gene family within the Lauraceae-specific microsynteny gene cluster (C10), which was identified in Figures 4D, E. This syntenic block was compared among *P. bournei*, *C. kanehirae*, *L. cubeba*, and *L. megaphylla*. Blue squares represent *6OMT* genes and aquamarine squares represent other genes on the syntenic block. **(D)** The syntenic block containing the *4OMT* gene family within the Lauraceae-specific microsynteny gene cluster (C7), which was identified in Figures 4D, E. This syntenic block was compared among *P. bournei*, *C. kanehirae*, *L. cubeba*, and *L. megaphylla*. Red squares represent *4OMT* genes and aquamarine squares represent other genes on the syntenic block. **(E)** The different panels show the phylogenetic tree of the *4OMT* gene family (left),the distribution of motifs in the promoter sequences and the predicted transcription factor binding sites (TFBS) (middle), and the distribution of motifs in protein sequences (right). Thick squares represent motifs and thin ones represent TFBSs. Dashed boxes highlight genes and promoter motifs unique to Lauraceae species.

Lauraceae-specific DNA motif (motif 1), but only group 1 contained potential MYB transcription factor binding sites (TFBSs) (Figure 5D). One Lauraceae-specific promoter motif (motif 6) in group 2 (C7) overlapped with predicted WRKY and bHLH TFBSs (Figure 5E). Interestingly, we identified a unique promoter motif (motif 8) among group 1 genes, and found a conserved cluster formed by six motifs (motif 8, 9, 3, 5, 2, and 1) that overlap with WRKY, bHLH, B3, ERF, MYB, and C2H2 TFBSs. These TFBS clusters may play key roles in coordinating specific gene expression as well as efficient activation and regulation of alkaloid biosynthesis.

Columbamine O-methyltransferase (CoOMT) is a vital enzyme that catalyzes the formation of tetrahydropalmatine, an isoquinoline alkaloid. The C4, a Lauraceae-specific microsyntenic cluster contained all *CoOMT* genes (Figures 4E, S17A). We found that TD/PD duplications occurred before Lauraceae speciation, producing three major *CoOMT* groups (group 1, 2, and 3) (Figures S17A, 17B). In *L. megaphylla*, all members of the *CoOMT* family were found in one TD/PD cluster on chromosome 3 (Figure S17B). Two Lauraceae-specific motifs (motif 9 and motif 10) among *CoOMT* protein sequences were identified (Figure S17A). In addition, we identified two Lauraceae-specific promoter motifs (motif 3 and motif 4), of which motif 3 is the potential TFBS of WRKY, ERF, and MYB TFs, and motif 4 is the potential TFBS of bHLH TFs (Figures S17A, 17C).

Cytochrome P450 monooxygenases (CYPs) play an important role in the structural and functional diversity of alkaloids. The *CYP80B*, *CYP80G*, and *CYP719A* gene families play key oxidative roles in BIA metabolism (Hagel and Facchini, 2013; Nguyen and Dang, 2021). A total of 20 microsyntenic clusters were identified as related to the *CYP* gene family, among which three clusters were unique to Lauraceae (C11, C12, and C5) (Figures S18A, 18B, S19). Specifically, microsyntenic cluster C11 is related to the *CYP719A* family, and C12 and C5 are related to the *CYP80G* family. TD/PD expansion of genes on C5 cluster occurred in all Lauraceae species, especially in *L. megaphylla* (Figure S19A).

CYP719A catalyzes the conversion of (S)-tetrahydro columbamine to (S)-tetrahydroberberine, and is an essential enzyme in berberine biosynthesis (Ikezawa et al., 2003). According to the phylogenetic tree, *CYP719A* genes from Lauraceae can be divided into two groups. All members of C11 were classified into group 2, and these genes are located in a species-specific TD/PD cluster found on *L. megaphylla* chromosome 8 (Figure S20B). A motif unique to Lauraceae (motif 12) was discovered in the protein sequences of these *CYP719A* genes (Figures S20A, 20C). Further, three Lauraceae-specific motifs (motif 1, motif 4, and motif 8) were found in the promoters (Figures S20A, 20B). Among these motifs, motif 1 contains NAC TFBSs, motif 4 contains bHLH and ERF TFBSs, and motif 8 contains MYB TFBSs (Figure S20C). Notably, TFs such as bHLH, NAC, WRKY, and MYB have been implicated in the regulation of BIA biosynthesis in plants (Yamada et al., 2011; Zhou and Memelink, 2016; Deng et al., 2018). Here, we identified Lauraceae-specific and conserved protein sequences, TFBS motifs, and TFBS clusters among BIA biosynthesis genes. It is found that the genes related to BIA biosynthesis in Lauraceae species are significantly different from those in other species. These findings are valuable in the genetic dissection of BIA biosynthesis in Lauraceae species.

Characterization of genes involved in phenolic compound biosynthesis

We next examined the lignin and flavonoid biosynthesis pathways, which are the downstream branches of phenylpropanoid metabolism related to phenol biosynthesis (Figure 6A and Table S15). Phenolic compounds can protect wood from decaying organisms and improve WDR. The key reactions of general phenylpropanoid biosynthesis involve three enzymes: phenylalanine ammonia-lyase (PAL), cinnamate 4-hydroxylase (C4H), and 4-coumarate coenzyme A ligase (4CL). Among these enzymes, we found that the *C4H* and *4CL* genes underwent remarkable TD/PD duplication events in Lauraceae (Figures 6A, B, S21, and Table S16).

Our microsynteny analysis of *C4H* genes revealed a Lauraceae-specific cluster (C4) (Figure 6C). Genes of this C4 cluster were divided into two groups resulted from the Lauraceae-specific TD/PD duplication (Figure 6D). Two Lauraceae-specific protein motifs (motif 12 and motif 13) were identified in these C4 cluster genes (Figures 6E, F, S22). We also found two motifs (motif 2 and motif 4) specific to Lauraceae in the promoter regions of these genes (Figures 6E, S22). Motif 2 overlapped with C2H2 TFBSs and motif 8 with that of MYB and ERF TFs (Figures 6E, S22). Members of all these TF families are involved in the regulation of phenylpropanoid biosynthesis (Ma et al., 2017; Mondal and Roy, 2018; Teng et al., 2018). In addition, among the C4 genes, these two Lauraceae-specific motifs were clustered together with motifs 7, 9, 6, 3, 1, and 8, forming a very distinct cluster of ERF, MYB, bHLH, and ERF TFBSs. This motif cluster was shared among Lauraceae species and *L. chinense* (Figures 6E, S22).

Sequence analysis of promoter regions revealed two motifs (motif 8 and motif 9) unique to Lauraceae of *PAL* genes (Figure 7A). Motif 8 overlapped with TCP TFBSs and motif 9 with that of TCP and GATA TFs (Figure 7A). TCP TFs play an important role in plant defense and have been found to enhance flavonoid biosynthesis of *Arabidopsis thaliana* (Li and Zachgo, 2013; Li, 2014). Moreover, overexpression of a GATA gene can enhance the activity of the phenylpropanoid biosynthesis pathway in *Solanum lycopersicum* (Zhao et al., 2021b). Similar to the *PAL* genes, although there was no Lauraceae-specific collinearity cluster found related to *4CL* genes, a motif in the promoters unique to Lauraceae (motif 7) was identified and overlapped with C2H2 TFBSs (Figure 7B). Moreover, conserved TFBS clusters were also found among the promoters of *PAL* and *4CL* genes. These TFBSs were of TFs belonging to the TCP, BFR-BPC, C2H2, ERF, MYB, GATA, and GRAS families (Figures 7A, B).

The biosynthesis pathways of taxifolin, myricetin, catechin, quercetin, and kaempferol have been annotated in Lauraceae species (Figure 6A). All of these flavonoids have been reported to improve plant WDR (Nascimento et al., 2013). TD/PD duplications accounted for expansions of *F3H* (flavanone 3-hydroxylase) and *F3′5′H* (flavonoid 3′,5′-hydroxylase) genes in Lauraceae (Figures 6B,

10.3389/fpls.2023.1122549

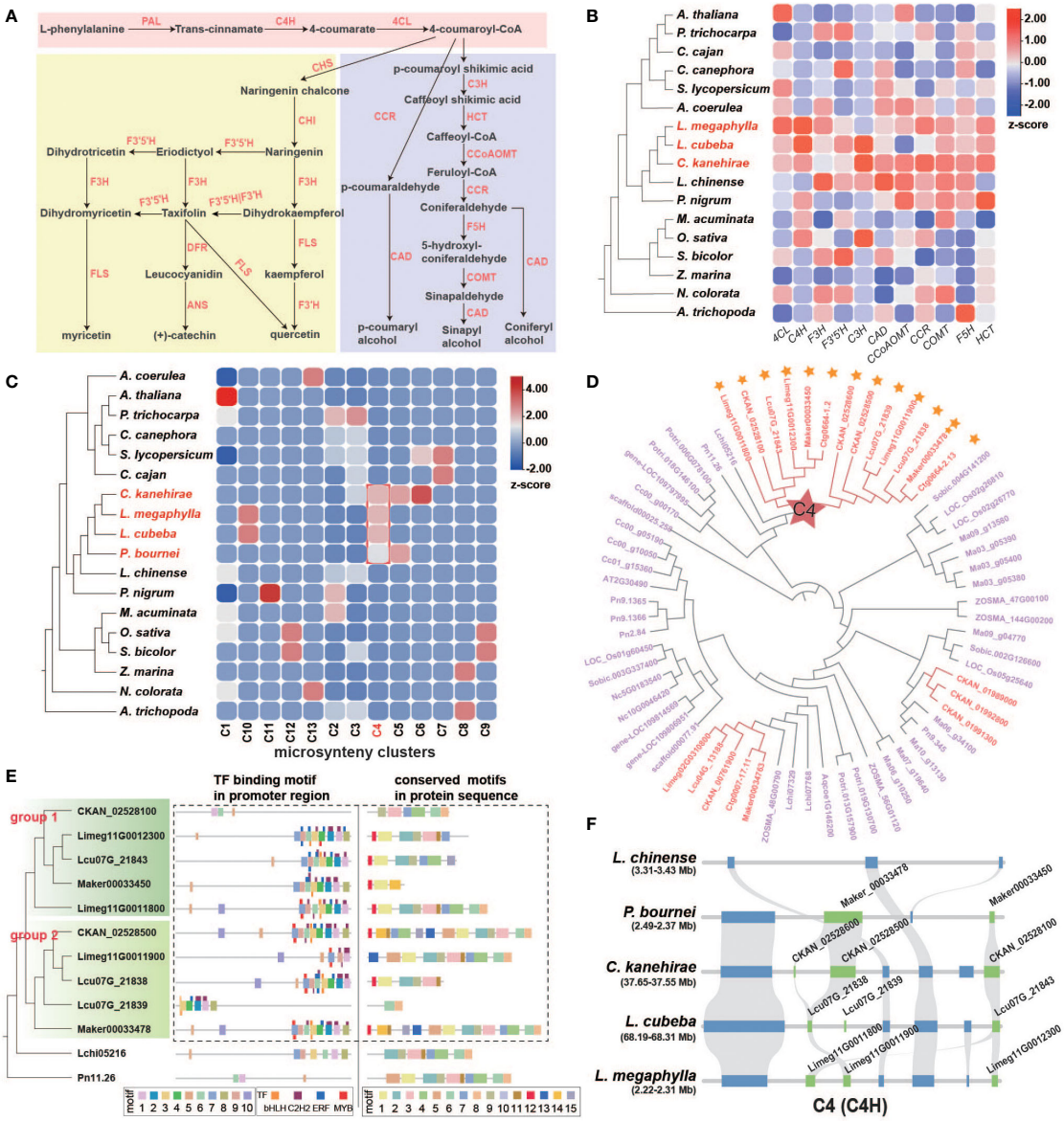


FIGURE 6
Characteristics of flavonoid and lignin genes in Lauraceae. **(A)** Biosynthesis pathways of general phenylpropanoids, flavonoids, and lignin. PAL, phenylalanine ammonia-lyase; C4H, cinnamate-4-hydroxylase; 4CL, 4-coumarate CoA ligase 4; CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone 3-hydroxylase; FLS, flavonol synthase; F3′H, flavonoid 3′-hydroxylase; F3′5′H, flavonoid 3′,5′-hydroxylase; DFR, dihydroflavonol 4-reductase; ANS, anthocyanidin synthase; C3′H, p-coumaroyl shikimate 3′-hydroxylase; CCR, cinnamoyl-CoA reductase; CAD, (hydroxy)cinnamyl alcohol dehydrogenase; HCT, hydroxycinnamoyl-CoA:shikimate/quinate hydroxycinnamoyltransferase; CCoAOMT, caffeoyl-CoAO methyltransferase; F5H, coniferaldehyde/ferulate 5-hydroxylase; COMT, caffeicacid/5-hydroxyferulic acid O-methyltransferase. **(B)** Proportion of tandem and proximal duplication genes in *4CL*, *C4H*, *F3H*, *F3′5′H*, *C3′H*, *CAD*, *CCoAOMT*, *COMT*, *F5H* and *HCT* gene families in each species. **(C)** Heatmap of 13 microsynteny clusters of *C4H* gene families in 18 species. The Lauraceae-specific cluster is highlighted by a red square. Colors in the heatmap indicate gene number in each cluster for each species. **(D)** Phylogenetic analysis of *C4H* gene families. The gene names of Lauraceae species are shown in red, and red stars represent Lauraceae-specific gene clusters identified in **(C)**. Yellow stars show the tandem and proximal duplication (TD/PD) genes. **(E)** The phylogenetic tree of the C4H gene family (left), the distribution of motifs in the promoter sequences and the predicted transcription factor binding sites (TFBS) (middle), and the distribution of motifs in protein sequences (right) are shown. Thick squares represent motifs and thin ones represent TFBSs. Dashed boxes highlight genes and promoter motifs unique to Lauraceae species. **(F)** The syntenic block containing the *C4H* gene family within the Lauraceae-specific microsynteny gene cluster (C4) identified in **(C)**. Here, this syntenic block was compared among *L. chinense*, *P. bournei*, *C. kanehirae*, *L. cubeba*, and *L. megaphylla*. Chartreuse squares represent *C4H* genes and blue squares represent other genes on the syntenic block.

S21, and Table S17). F3H is an important rate-limiting enzyme in flavonoid biosynthesis pathway. Enzymatic gene families of the lignin biosynthesis pathway include *C3′H*, *HCT*, *CCR*, *CAD*, *CCoAOMT*, *F5H*, and *COMT*, all of which were expanded through TD/PD duplications in Lauraceae (Figures 6B, S23, and Table S18). Microsynteny analysis revealed two Lauraceae-specific conserved gene clusters (C9 and C24) associated with lignin pathway genes (*HCT* and *CCR*) (Figures 8A–D). Although no

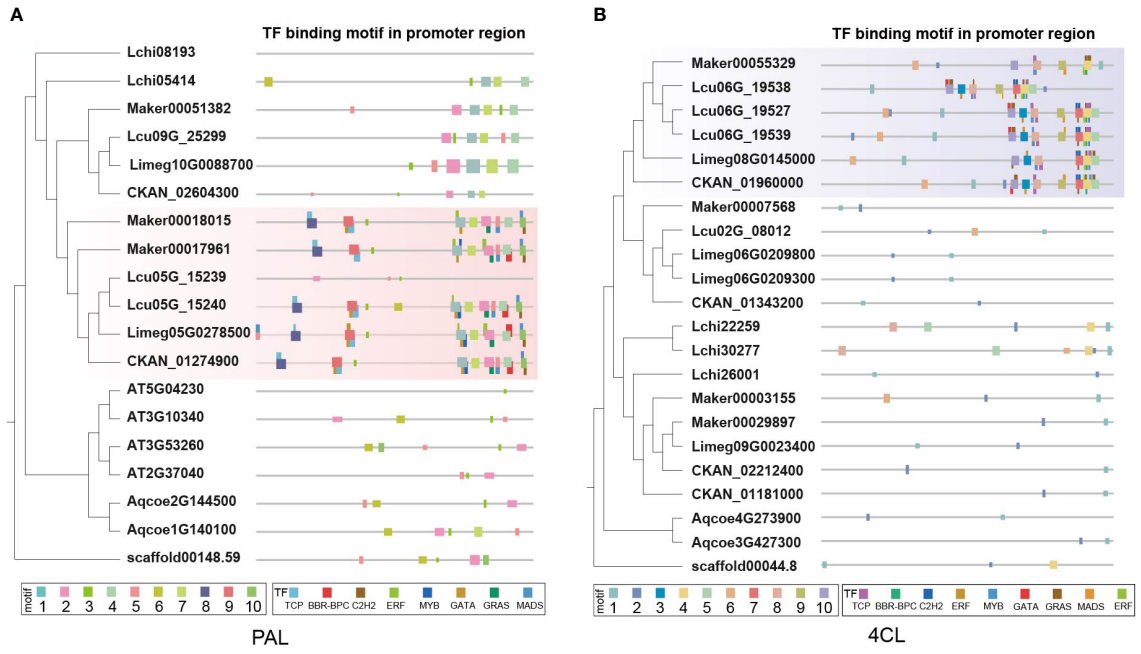


FIGURE 7
Characterization of *PAL* and *4CL* genes in Lauraceae species. **(A)** Different panels represent of the phylogenetic tree of the *PAL* gene family (the left), the distribution of motifs in the promoter sequence and the predicated transcription factor binding sites (TFBS) (the right). Fat squares represent the motifs and thin ones the TFBSs. Red boxes highlight the promoter motifs uniquely found among the Lauraceae species. **(B)** Different panels represent of the phylogenetic tree of the *4CL* gene family (the left), the distribution of motifs in the promoter sequence and the predicated transcription factor binding sites (TFBS) (the right). Fat squares represent the motifs and thin ones the TFBSs. Purple boxes highlight the promoter motifs uniquely found among the Lauraceae species.

Lauraceae-specific motifs and TFBSs were found among genes of the C9 and C24 clusters, all of these genes showed obvious TD/PD expansion (Figures 8B, D).

Remarkable TD/PD duplications were also found for *TPS* gene family of Lauraceae species, which may be associated with the super WDR. Details are available in Note S8, Figures S24-26 and Table S19.

Discussion

Our genomic investigation, especially the gene microsynteny profiling, may contribute to resolving the phylogenetic position of Magnoliids relative to eudicots and monocots, the other two major angiosperm groups. Although multiple assemblies of magnoliid genomes have been published, such as *C. kanehirae* (Chaw et al., 2019a), *L. chinense* (Chen et al., 2019), *P. nigrum* (Hu et al., 2019), *P. americana* (Rendon-Anaya et al., 2019), *P. bournei* (Chen et al., 2020a), *L. cubeba* (Chen et al., 2020b), *C. salicifolius* (Lv et al., 2020), and *C. praecox* (Shang et al., 2020a), the phylogenetic placement of Magnoliids still remains unclear. Our phylogenetic analyses using three different methods (concatenation-, coalescent-, and microsynteny-based approaches) confirmed that Magnoliids are the sister group of eudicots, which is in line with previous genomic analyses (Chaw et al., 2019a; Lv et al., 2020; Shang et al., 2020a) and phylotranscriptomic analyses of 92 streptophytes (Wickett et al., 2014) and 20 representative angiosperms (Zeng et al., 2014). In addition, the microsyntenic clusters of 16 species in Magnoliids, eudicots, and monocots were further analyzed. There were significantly more shared clusters in Magnoliids-eudicots compared with Magnoliids-monocots and eudicots-monocots, which strongly supports the finding that Magnoliids and eudicots are sister groups. The three clades were enriched in different GO and KEGG terms, indicating their functional divergence. The genes of Lauraceae-specific microsyntenic clusters were significantly enriched in terms including isoquinoline alkaloid biosynthesis, phenylpropanoid metabolic and lignin metabolic processes, suggesting that various Lauraceae-specific biochemical processes may influence its wood decay resistance.

A variety of bioactive compounds, including terpenoids, alkaloids, and phenolic compounds such as flavonoids, have been associated with WDR (Nascimento et al., 2013; Anouhe et al., 2018). In addition to the dual fungicidal and antioxidant effects of bioactive compounds, other factors such as lignin content also impact WDR (Vance et al., 1980; Nascimento et al., 2013; Mounguengui et al., 2016). Apart from annotating enzymes involved in the biosynthesis of isoquinoline alkaloids, flavonoids, terpenoids, and lignins, we characterized genes, gene syntenies, gene expansions, and gene promoter motifs specific to Lauraceae, which help to track genomic characters potentially related with the super wood decay resistance.

The biosynthetic pathways of three benzylisoquinoline alkaloids (BIA), namely magnoflorine, berberine, and palmatine, were annotated in Lauraceae species. Both berberine and palmatine exhibit significant antifeedant activity against termites (Kawaguchi et al., 1989). Magnoflorine is an aporphine-type BIA

10.3389/fpls.2023.1122549

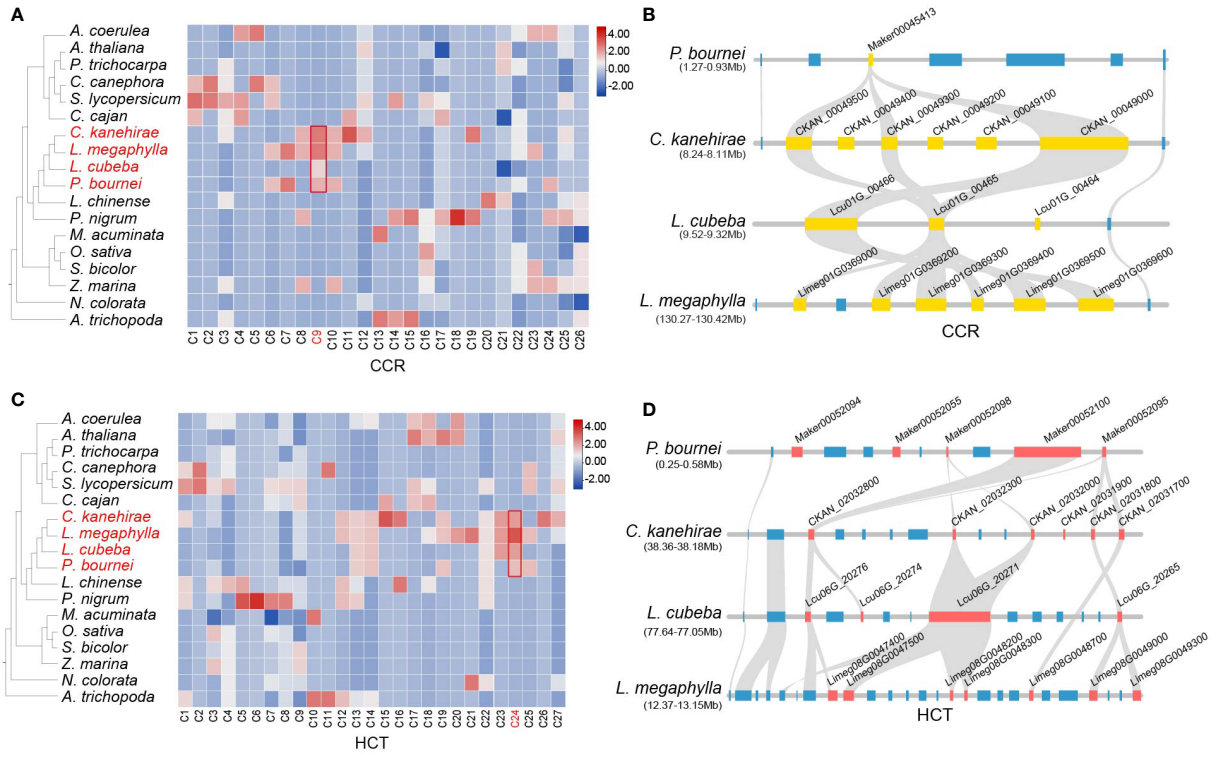


FIGURE 8
Lauraceae-specific *CCR* and *HCT* genes in lignin biosynthesis pathway. **(A)** Heatmap of 26 microsynteny clusters identified to be related with *CCR* gene family, one of which specific to Lauraceae were highlighted in a red square. Color in the heatmap was determined by the gene number found in each cluster for each species. **(B)** The syntenic block containing of *CCR* gene family within the Lauraceae-specific microsynteny gene cluster (C9) identified in **(A)**. Here this syntenic block was compared among *P. bournei*, *C. kanehirae*, *L. cubeba* and *L. megaphylla*. Yellow squares represent the *CCR* genes and blue ones represent other genes on the syntenic block. **(C)** Heatmap of 27 microsynteny clusters identified to be related with *HCT* gene familiy, one of which specific to Lauraceae were highlighted in a red square. Color in the heatmap was determined by the gene number found in each cluster for each species. **(D)** The syntenic block containing of *HCT* gene family inside the Lauraceae-specific microsynteny gene cluster (C24) in **(C)**. Here this syntenic block was compared among *P. bournei*, *C. kanehirae*, *L. cubeba* and *L. megaphylla*. Red squares represents the *HCT* genes and blue ones represents other genes on the syntenic block.

that has antibacterial and insecticidal effects, and may also play a role in improving WDR (Okon et al., 2020). Our comparative analyses demonstrated that the *OMT*, *CYP*, and *BBE* gene families involved BIA biosynthesis showed specific expansion in Lauraceae. Most members of these gene families originated from TD/PD duplications, which greatly enriched the enzymatic genes of the BIA biosynthesis pathway. These data indicate the significant value of TD/PD duplications in BIA biosynthesis. In the *OMT* gene family, a total of four Lauraceae-specific microsyntenic clusters were identified, including genes of the *4OMT*, *6OMT*, and *CoOMT* subfamilies. Again, TD and PD duplications were associated with significant expansion of the *CoOMT* gene family in *L. megaphylla*, which may have contributed to the accumulation of palmatine, thereby further improving WDR.

In addition to the Lauraceae-specific gene microsyntenic clusters uncovered for the biosynthesis of bioactive compounds related to WDR, we also found conserved TFBS clusters in the promoter regions of genes in these conserved clusters. These conserved TFBS clusters suggest conserved transcriptional regulation of secondary metabolite biosynthesis efficiency, which may lead to the high WDR trait shared among many Lauraceae woods. In the *OMT* gene family, the Lauraceae-specific promoter motifs were mainly TFBSs of bHLH, MYB, ERF and WRKY TFs. In the *CYP* gene family, the conserved promoter motifs were generally TFBSs of bHLH, MYB, ERF, and NAC TFs, all of which have been reported to be involved in the regulation of BIA biosynthesis (Yamada et al., 2011; Deng et al., 2018). In addition, we found that B3, GATA, Trihelix, and C2H2 TFs may bind these Lauraceae-specific TFBS clusters. However, their involvement in the regulation of alkaloid biosynthesis requires further evaluation. Compared with other species, the unique characteristics of Lauraceae species in BIA biosynthesis suggest that isoquinoline alkaloids may play a large proportion of roles in the decay resistance of Lauraceae.

There are diverse metabolic branches downstream of the general phenylpropanoid biosynthesis. Of these branches, we investigated the lignin and flavonoid pathways in the present study. The *C4H* and *4CL* genes of the general phenylpropanoid pathway, *F3H* and *F3′5′H* of the flavonoid pathway, and all gene families of the lignin pathway have undergone significant TD/PD duplication in Lauraceae. C4H is the second key enzyme in the general phenylpropanoid biosynthesis pathway, and belongs to the CYP73A subfamily. C4H directly affects the biosynthesis and yield of flavonoids and lignin in plants (Ryan et al., 2002; Millar et al., 2007). Lauraceae-specific genes were found in the *C4H* gene family.

In addition to carrying motifs in the coding and promoter regions that were different from other species, these *C4H* genes also had unique TFBS clusters specific to Lauraceae. Such TFBSs in the clusters are adjacent to each other, including binding sites of bHLH, C2H2, ERF, and MYB TFs, which all have important regulatory functions in phenylpropanoid biosynthesis (Ma and Constabel, 2019; Yadav et al., 2020; Meng et al., 2021). Moreover, TD/PD events also occurred in the Lauraceae-specific genes of the *C4H* gene family, which greatly increased their coding space, and further contributed to the WDR of Lauraceae species. In addition, Lauraceae-specific TFBS clusters were also found in the promoter regions of genes encoding PAL and 4CL. PAL is a rate-limiting enzyme that catalyzes the first step in the phyenylpropanoid biosynthesis pathway. Thus, it plays an important role in phenylpropanoid biosynthesis (Zhao et al., 2021a). 4CL, the third enzyme in the general phenylpropanoid biosynthesis pathway, participates in monolignol biosynthesis through the production of p-coumaroyl-CoA, a precursor for the biosynthesis of lignin, flavonoid compounds, and plant defense compounds (isoflavonoids). Therefore, compared with other plant groups, general phenylpropanoid biosynthesis genes in Lauraceae are highly unique, which affects the biosynthesis of flavonoids and lignin and may improve the natural durability of Lauraceae wood. Studies found that functional disruption of *CCR* and *HCT* genes affects lignin content (Thévenin et al., 2011; Wang et al., 2015). Although microsyntenic clusters were notable in the *CCR* and *HCT* gene families, no unique motifs were found among the protein sequences and promoter regions of homologous genes. We suspected that these Lauraceae-specific genes may have arisen more recently and have not yet diverged significantly from the original genes, in addition, these genes also showed significant TD/PD expansion.

In summary, we investigated the WDR of Lauraceae species by identifying microsynteny clusters among different angiosperm lineages. The Lauraceae-specific biosynthetic genes related to WDR, the conserved motifs of the encoding proteins, the unique and conserved gene expansion and TFBS clusters may play a vital role in increasing and regulating WDR, which may be the main reason for the super decay resistance of Lauraceae. The present genome resources and investigation lay the foundation for molecular breeding or genetic engineering of Lauraceae, and provide key resources for further exploration of the naturally durable wood of Lauraceae species.

Materials and methods

Plant material

A healthy, fruitful, mature *L. megaphylla* individual was selected and used for whole genome sequencing. This individual was collected from naturally regenerated forest at the National Tree Breeding Station for Nanmu in Zhuxi, Forest Farm of Zhuxi County, Hubei, China. For RNA sequencing, flower buds, stems, buds, and leaves were sampled from healthy trees in the same location, with three replicates per tissue. Tissues were immediately flash frozen and stored at -80 °C for subsequent nucleic acid extractions.

Genome sequencing

For Nanopore sequencing, PromethION libraries were prepared and sequenced on a Nanopore PromethION platform. For Illumina sequencing, 150-bp paired-end (PE) libraries were prepared for sequencing on an Illumina HiSeq X Ten platform. The Hi-C library prepared with the MboI restriction enzyme was sequenced in an Illumina HiSeq X Ten to generate 1488.194 million reads (~223 Gb, roughly 170x coverage of the assembled genome) from 150-bp PE reads. For RNA sequencing, four tissues (flower buds, stems, buds, and leaves) were used to construct mRNA sequencing libraries, after which 150-bp PE sequencing was performed in an Illumina HiSeq X Ten. RNA sequencing produced 996.020 million raw reads (~145 Gb).

More details regarding genome sequencing are available in Note S2.

De novo genome assembly and quality control

De novo genome assembly involved three steps: primary assembly, Hi-C scaffolding, and polishing. First, we used SMARTdenovo (see "URLs" section), WTDBG (version 2.1) (Ruan and Li, 2020), and Canu (version 1.7) (Koren et al., 2017) to generate four of the primary assemblies from ONT long reads. Then, one primary assembly (v0.3, with reasonably sized assembly, fewest contigs, and highest contig N50) was chosen as the optimal assembly, and further polished with three rounds of pilon (see "URLs" section) with clean Illumina reads to generate assembly v1.0. Based on Hi-C data and assembly v1.0, primary scaffolds were produced with 3D-DNA (version 180922) (see "URLs" section). These scaffolds were inspected and manually corrected using Juicebox (version 1.8) (see "URLs" section) and re-scaffolded by 3D-DNA. Afterwards, we optimized the new scaffolds with gap closing using LR_Gapcloser (version 1.1) (see "URLs" section) followed by four rounds of pilon polishing.

Benchmarking Universal Single Copy Orthologs (BUSCO) and LTR Assembly Index (LAI) were used to assess genome completeness and continuity. To evaluate the completeness of the assembly and uniformity of the sequencing, 178 Gb of ONT reads, 160 Gb of clean Illumina reads, and 90 Gb of RNA sequencing reads were aligned to the assembly genome using BWA-MEM (see "URLs" section), minimap2 (Li, 2018), and HiSat2 (version 2.1.0) (see "URLs" section), respectively.

More details of genome assembly are available in Note S3.

Genome annotation

Protein-coding genes were predicted using the MAKER2 pipeline (Holt and Yandell, 2011) including *ab initio*, homolog

proteins, and EST-based prediction methods. We annotated non-coding RNAs (ncRNAs) with several databases and software including tRNAscan-SE (version 1.3.1) (Lowe and Eddy, 1997), RNAMMER (version 1.2) (Lagesen et al., 2007), Rfam database (version 9.1) (see "URLs" section), and BLASTN (version 2.2.28+).

Functions of predicted genes were annotated using sequence similarity searches by BLAT (version 36) (Kent, 2002) with 30% identity and 1e-05 E-value cutoff, as well as domain similarity annotations using InterProScan (version 5.27-66.0) (see "URLs" section). The completeness of genome annotation was assessed using BUSCO. Centurion (Varoquaux et al., 2015) was used to infer the location of all centromeres in the genome based on corrected Hi-C data.

Repeated elements were annotated using RepeatModeler (version 1.0.10) (see "URLs" section) and RepeatMasker (version 4.0.7, rmblast-2.2.28) (see "URLs" section) with homology-based and *de novo* approaches. In addition, we examined classification, age distribution, birth, and death of LTR-RTs.

More details of genome annotation are available in Note S4 and S5.

Gene family and phylogenetic inference

To determine the phylogenetic relationships among Magnoliids, we used Orthofinder (version 2.3.1) (Emms and Kelly, 2019) to identify gene families from 6 eudicots including *Aquilegia coerulea* (Filiault et al., 2018), *Populus trichocarpa* (Tuskan et al., 2006), *Arabidopsis thaliana* (Michael et al., 2018), *Coffea canephora* (Denoeud et al., 2014), *Solanum lycopersicum* (Sato et al., 2012) and *Cajanus cajan* (Varshney et al., 2012), 4 monocots including *Zostera marina* (Olsen et al., 2016), *Sorghum bicolor* (Deschamps et al., 2018), *Musa acuminata* (D'Hont et al., 2012) *and Oryza sativa* (Ouyang et al., 2006), 6 Magnoliids including *Piper nigrum* (Hu et al., 2019), *Liriodendron chinense* (Chen et al., 2019), *Persea americana* (Rendon-Anaya et al., 2019), *Cinnamomum kanehirae* (Chaw et al., 2019a), *Litsea cubeba* (Chen et al., 2020b) and *Lindera megaphylla* and 2 outgroup species including *Amborella trichopoda* (Albert et al., 2013) and *Nymphaea colorata* (Zhang et al., 2020b). A total of 34,888 orthogroups, including 112 orthologous single-copy gene families and 885 low-copy orthologs with minimum of 83.3% of species having single-copy genes in any orthogroup. Amino acid sequence alignment was performed on these low-copy genes using MUSCLE (version 3.8.31) (Edgar, 2004).

Phylogenetic trees were constructed using concatenation-, coalescent-, and microsynteny-based approaches (Zhao et al., 2021c). For the concatenation-based approach, the maximum likelihood tree was constructed based on concatenated low-copy amino acid sequences with IQ-TREE (version 1.6.7) (Nguyen et al., 2014), employing the best-fit model (-m JTT+F+R5) with ultrafast bootstrapping (-bb 1000). For the coalescent-based approach, gene trees of 855 low-copy gene families were inferred by IQ-TREE. Next, we removed low bootstrap support branches (less than 50%) using the Newick utilities. Then, gene trees were used to construct species trees with ASTRAL-pro. Quartet support of each node was estimated for this coalescent tree. Finally, the microsynteny-based method included two steps. First, after an all-by-all protein alignment of the whole genome was performed using DIAMOND (Buchfink et al., 2015), pairwise synteny blocks were identified using MCScanX (Wang et al., 2012). Then, microsyntenic clusters were detected using Infomap (see "URLs" section).

The maximum likelihood (ML) phylogenetic tree was generated with IQ-TREE (version 1.6.7), using the Mk+R+FO model and ultrafast bootstrapping (-bb 1000). The ML tree constructed using the coalescent-based approach was used as an input tree to estimate divergence time with the MCMCTree program in the PAML package (version 4.9h) (Yang, 2007). Dating was calibrated according to the TimeTree web service (http://www.timetree.org/) by placing soft bounds at four split nodes as constraints for calibrating tree age: (1) the *A. trichopoda* node (173-199 Mya), (2) *L. chinense* (117-130 Mya), (3) *O. sativa-S. bicolor* (42-52 Mya), and (4) *P. trichocarpa-A. thaliana* (98-177 Mya). Expansion and contraction of gene families were inferred with CAFÉ (version 4.1) (De Bie et al., 2006).

Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) enrichment analyses were performed using the R package clusterProfiler (version 3.6.0) (Yu et al., 2012).

Additional details are available in Note S6.

Analysis of microsyntenic clusters

The microsyntenic clusters were identified with a computational pipeline previously setup (Zhao et al., 2021c). Key steps in the process are as follows. After an all-vs-all reciprocal sequence similarity search for all annotated genomes using DIAMOND (Buchfink et al., 2015), pairwise synteny block detection was performed using MCScanX (Wang et al., 2012). Then the synteny network was clustered using the Infomap algorithm (see "URLs" section). After that, a synteny cluster matrix was obtained, and the number of each species in each cluster was noted, in which the rows and columns correspond to the various species and clusters, respectively. The matrix was then converted into a binary matrix for phylogenetic inference, where 1 denoted the presence of a specific cluster for the species and 0 denoted its absence. This matrix was analyzed using the cor function in R tools to obtain the correlation coefficient between species. The correlation matrix was then plotted and visualized using the R package corrplot (Wei and Simko, 2017). The states of synteny clusters of 16 species were visualized using the UpSetR package (Conway et al., 2017). Clusters shared among Magnoliids-eudicots, Magnoliids-monocots, and eudicots-monocots were further visualized in Cytoscape (Shannon et al., 2003). To select representative clusters of Magnoliids, eudicots, and monocots, the following criteria were set: for Magnoliids, microsyntenic clusters present in four or more species were reserved; for eudicots, microsyntenic clusters present in four or more species were reserved; for monocots, microsyntenic clusters present in three or more species were reserved.

Genome duplication

We examined genome-wide gene duplications in *L. megaphylla*, *C. kanehirae*, and *L. cubeba* using DupGen_finder (Qiao et al., 2019) with default parameters. The duplicated genes were annotated into five different gene duplication models, including whole-genome duplication (WGD), tandem duplication (TD), proximal duplication (less than 10 gene distance on the same chromosome: PD), transposed duplications (TRD), or dispersed duplications (DSD).

Secondary metabolite biosynthesis pathways

Protein sequences from sequenced *Lindera* genomes were processed with the Ensemble Enzyme Prediction Pipeline (E2P2) package (version 3.1) (see "URLs" section) to identify putative enzymes. Based on these enzymatic annotations, we then constructed a metabolic pathway database by querying the Plant Metabolic Network (see "URLs" section). The derived pathway database was then validated using SAVI (version 3.1) (Schlapfer et al., 2017) to remove any false positives and redundant pathways, such as non-plant pathway variants, as well as pathways already included in larger pathways. Gene family trees were constructed using IQ-TREE (version 1.6.7) with 1,000 bootstrap replicates. The sequences spanning 2 kb upstream of genes were used to identify transcription factor binding sites (TFBS) in promoters. Putative TF binding sites for suspected promoter sequences were predicted by PlantRegMap (Tian et al., 2019) with q-value ≤ 0.05.

URLs

SMARTdenovo [https://github.com/ruanjue/smartdenovo];
Pilon [http://github.com/broadinstitute/pilon];
3D-DNA (version 180922) [https://github.com/theaidenlab/3d-dna];
Juicebox (version 1.8) [https://github.com/aidenlab/Juicebox];
LR_Gapcloser (version 1.1) [https://github.com/CAFS-bioinformatics/LR_Gapcloser];
BWA-MEM [https://github.com/lh3/bwa];
HiSat2 (version 2.1.0) [https://github.com/infphilo/hisat2];
RepeatMasker [http://www.repeatmasker.org];
RepeatModeler [http://www.repeatmasker.org];
Rfam database (version 9.1) [http://eggnogdb.embl.de];
InterProScan (version 5.27-66.0) [http://www.ebi.ac.uk/InterProScan];
Infomap algorithm [https://github.com/mapequation/infomap];
Timetree web service (http://www.timetree.org/);
PMN Ensemble Enzyme Prediction Pipeline (E2P2, version 3.1) (https://gitlab.com/rhee-lab/E2P2);
Plant Metabolic Network (https://www.plantcyc.org).

Data availability statement

The data presented in the study are deposited both in the NCBI repository with the accession number SRP382804 (https://www.ncbi.nlm.nih.gov/), and in the Genome Warehouse in National Genomics Data Center with the accession number GWHBKHA00000000 (https://ngdc.cncb.ac.cn/gwh).

Author contributions

J-XL conceived and designed the study; X-CT, J-FG, X-MY, T-LS, SN, S-WZ, Y-TB, Z-CL, and LK prepared the materials and performed related analysis; G-JS provided the specimens; X-CT, J-FG, and J-FM wrote the manuscript; J-FM involved in structuring and polishing the manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

We acknowledge support from the Key Program of the National Natural Science Foundation of China (32030010).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1122549/full#supplementary-material

References

Albert, V. A., Barbazuk, W. B., Depamphilis, C. W., Der, J. P., Leebens-Mack, J., Ma, H., et al. (2013). The *Amborella* genome and the evolution of flowering plants. *Science* 342, 1241089. doi: 10.1126/science.1241089

Anouhe, J.-B. S., Niamké, F. B., Faustin, M., Virieux, D., Pirat, J.-L., Adima, A. A., et al. (2018). The role of extractives in the natural durability of the heartwood of *Dicorynia guianensis* amsh: new insights in antioxydant and antifungal properties. *Ann. For. Sci.* 75, 1–10. doi: 10.1007/s13595-018-0691-0

Bowers, J. E., Chapman, B. A., Rong, J., Li, C., Chai, X., and Tu, P.. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438. doi: 10.1038/nature01521

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Cao, Y., Xuan, B., Peng, B., Wang, H.-Y., Lin, C.-Y. I., Wu, C.-S., et al. (2016). The genus *Lindera*: a source of structurally diverse molecules having pharmacological significance. *Phytochem. Rev.* 15, 869–906. doi: 10.1007/s11101-015-9432-2

Chaw, S.-M., Liu, Y.-C., Wu, Y.-W., Zhao, C., Wang, P., Xue, L., et al. (2019a). Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nat. Plants* 5, 63–73. doi: 10.1038/s41477-018-0337-0

Chen, J., Hao, Z., Guang, X., Gao, M., Wang, J.-Y., Liu, K.-W., et al. (2019). *Liriodendron* genome sheds light on angiosperm phylogeny and species–pair differentiation. *Nat. Plants* 5, 18–25. doi: 10.1038/s41477-018-0323-6

Chen, Y.-C., Li, Z., Zhao, Y.-X., Jiang, Y. T., Liu, X. D., Liao, X. Y., et al. (2020b). The *Litsea* genome and the evolution of the laurel family. *Nat. Commun.* 11, 1675. doi: 10.1038/s41467-020-15493-5

Chen, S. P., Sun, W. H., Xiong, Y. F., and Chen, C.-F.. (2020a). The *Phoebe* genome sheds light on the evolution of magnoliids. *Horticulture Res.* 7, 146. doi: 10.1038/s41438-020-00368-z

Chou, C.-J., Lin, L.-C., Chen, K.-T., and Hahn, M. W.. (1994). Northalifoline, a new isoquinolone alkaloid from the pedicels of *Lindera megaphylla*. *J. Natural Products* 57, 689–694. doi: 10.1021/np50108a001

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an r package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938–2940. doi: 10.1093/bioinformatics/btx364

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W.. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271. doi: 10.1093/bioinformatics/btl097

Deng, X., Zhao, L., Fang, T., Xiong, Y., Ogutu, C., Yang, D., et al. (2018). Investigation of benzylisoquinoline alkaloid biosynthetic pathway and its transcriptional regulation in lotus. *Horticulture Res.* 5, 29. doi: 10.1038/s41438-018-0035-0

Denoeud, F., Carretero-paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., et al. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* 345, 1181–1184. doi: 10.1126/science.1255274

Deschamps, S., Zhang, Y., Llaca, V., Ye, L., Sanyal, A., King, M., et al. (2018). A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nat. Commun.* 9, 4844. doi: 10.1038/s41467-018-07271-1

Dewey, C. N. (2011). Positional orthology: putting genomic evolutionary relationships into context. *Briefings Bioinf.* 12, 401–412. doi: 10.1093/bib/bbr040

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., et al. (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* 488, 213–217. doi: 10.1038/nature11241

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Ekeuku, S. O., Pang, K.-L., and Chin, K.-Y. (2020). Palmatine as an agent against metabolic syndrome and its related complications: a review. *Drug Design Dev. Ther.* 14, 4963–4974. doi: 10.2147/DDDT.S280520

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 1–14. doi: 10.1186/s13059-019-1832-y

Filiault, D. L., Ballerini, E. S., Mandáková, T., Aköz, G., Derieg, N. J., Schmutz, J., et al. (2018). The aquilegia genome provides insight into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history. *Elife* 7, e36426. doi: 10.7554/eLife.36426.050

Hagel, J. M., and Facchini, P. J. (2013). Benzylisoquinoline alkaloid metabolism: a century of discovery and a brave new world. *Plant Cell Physiol.* 54, 647–672. doi: 10.1093/pcp/pct020

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinf.* 12, 491. doi: 10.1186/1471-2105-12-491

Hu, L., Xu, Z., Wang, M., Fan, R., Yuan, D., Wu, B., et al. (2019). The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. *Nat. Commun.* 10, 4702. doi: 10.1038/s41467-019-12607-6

Ikezawa, N., Tanaka, M., Nagayoshi, M., Shinkyo, R., Sakaki, T., Inouye, K., et al. (2003). Molecular cloning and characterization of CYP719, a methylenedioxy bridge-forming enzyme that belongs to a novel P450 family, from cultured *Coptis japonica* cells*. *J. Biol. Chem.* 278, 38557–38565. doi: 10.1074/jbc.M302470200

Imenshahidi, M., and Hosseinzadeh, H. (2020). *Berberine neuroprotection and antioxidant activity. oxidative stress and dietary antioxidants in neurological diseases*. Eds. C. R. Martin and V. R. Preedy (London: Academic Press), 199–216.

Inui, T., Kawano, N., Shitan, N., et al. (2012). Improvement of benzylisoquinoline alkaloid productivity by overexpression of 3′-hydroxy-N-methylcoclaurine 4′-o-methyltransferase in transgenic *Coptis japonica* plants. *Biol. Pharm. Bull.* 35, 650–659. doi: 10.1248/bpb.35.650

Isman, M. (2002). Insect antifeedants. *Pesticide Outlook* 13, 152–157. doi: 10.1039/b206507j

Jagels, R., Visscher, G., and Wheeler, E. (2005). An Eocene high arctic angiosperm wood. *IAWA J.* 26, 387–392. doi: 10.1163/22941932-02603009

Jiao, L., Lu, Y., Zhang, M., Chen, Y., Wang, Z., Guo, Y., et al. (2022). Ancient plastid genomes solve the tree species mystery of the imperial wood "Nanmu" in the forbidden city, the largest existing wooden palace complex in the world. *PLANTS PEOPLE PLANET* 4, 696–709. doi: 10.1002/ppp3.10311

Kawaguchi, H., Kim, M., Ishida, M., Ahn, Y.-J., Yamamoto, T., Yamaoka, R., et al. (1989). Several antifeedants from *Phellodendron amurense* against *Reticulitermes speratus*. *Agric. Biol. Chem.* 53, 2635–2640. doi: 10.1080/00021369.1989.10869702

Kent, W. J. (2002). BLAT-the BLAST-like alignment tool. *Genome Res.* 12, 656–664. doi: 10.1101/gr.229202

Koren, S., Walenz, B., Berlin, K., Miller, J., Bergman, N., Phillippy, A., et al. (2017). Canu: scalable and accurate long-read assembly *via* adaptive *K*-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H.-H., Rognes, T., Ussery, D. W., et al. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35, 3100–3108. doi: 10.1093/nar/gkm160

Li, S. (2014). Transcriptional control of flavonoid biosynthesis: fine-tuning of the MYB-bHLH-WD40 (MBW) complex. *Plant Signaling Behav.* 9, e27522. doi: 10.4161/psb.27522

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, S., and Zachgo, S. (2013). TCP 3 interacts with R2R3-MYB proteins, promotes flavonoid biosynthesis and negatively regulates the auxin response in *Arabidopsis thaliana*. *Plant J.* 76, 901–913. doi: 10.1111/tpj.12348

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964. doi: 10.1093/nar/25.5.955

Lv, Q., Qiu, J., Liu, J., Li, Z., Zhang, W., Wang, Q., et al. (2020). The *Chimonanthus salicifolius* genome provides insight into magnoliid evolution and flavonoid biosynthesis. *Plant J.* 103, 1910–1923. doi: 10.1111/tpj.14874

Ma, D., and Constabel, C. P. (2019). MYB repressors as regulators of phenylpropanoid metabolism in plants. *Trends Plant Sci.* 24, 275–289. doi: 10.1016/j.tplants.2018.12.003

Ma, R., Xiao, Y., Lv, Z., Tan, H., Chen, R., Li, Q., et al. (2017). AP2/ERF transcription factor, Ii049, positively regulates lignan biosynthesis in *Isatis indigotica* through activating salicylic acid signaling and lignan/lignin pathway genes. *Front. Plant Sci.* 8, 1361. doi: 10.3389/fpls.2017.01361

Meng, X., Wang, Y., Li, J., Jiao, N., Zhang, X., Zhang, Y., et al. (2021). RNA Sequencing reveals phenylpropanoid biosynthesis genes and transcription factors for *Hevea brasiliensis* reaction wood formation. *Front. Genet.* 12, 763841–763841. doi: 10.3389/fgene.2021.763841

Michael, T. P., Jupe, F., Bemm, F., Motley, S. T., Sandoval, J. P., Lanz, C., et al. (2018). High contiguity arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nat. Commun.* 9, 541. doi: 10.1038/s41467-018-03016-2

Millar, D. J., Long, M., Donovan, G., Fraser, P. D., Boudet, A.-M., Danoun, S., et al. (2007). Introduction of sense constructs of cinnamate 4-hydroxylase (CYP73A24) in transgenic tomato plants shows opposite effects on flux into stem lignin and fruit flavonoids. *Phytochemistry* 68, 1497–1509. doi: 10.1016/j.phytochem.2007.03.018

Mondal, S. K., and Roy, S. (2018). Genome-wide sequential, evolutionary, organizational and expression analyses of phenylpropanoid biosynthesis associated MYB domain transcription factors in *Arabidopsis*. *J. Biomolecular Structure Dynamics* 36, 1577–1601. doi: 10.1080/07391102.2017.1329099

Mounguengui, S., Saha Tchinda, J.-B., Ndikontar, M., Dumarçay, S., Attéké, C., Perrin, D., et al. (2016). Total phenolic and lignin contents, phytochemical screening, antioxidant and fungal inhibition properties of the heartwood extractives of ten Congo basin tree species. *Ann. For. Sci.* 73, 287–296. doi: 10.1007/s13595-015-0514-5

Nascimento, M. D., Santana, A., Maranhão, C., Oliveira, L., and Bieber, L.. (2013). Phenolic extractives and natural resistance of wood. *Biodegradation-Life Sci.* 801, 349–370. doi: 10.5772/56358

Nguyen, T.-D., and Dang, T.-T. T. (2021). Cytochrome P450 enzymes as key drivers of alkaloid chemical diversification in plants. *Front. Plant Sci.* 12, 682181. doi: 10.3389/fpls.2021.682181

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q.. (2014). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Okon, E., Kukula-Koch, W., Jarzab, A., Halasa, M., Stepulak, A., Wawruszak, A., et al. (2020). Advances in chemistry and bioactivity of magnoflorine and magnoflorine-containing extracts. *Int. J. Mol. Sci.* 21, 1330. doi: 10.3390/ijms21041330

Olsen, J. L., Rouzé, P., Verhelst, B., Lin, Y.-C., Bayer, T., Collen, J., et al. (2016). The genome of the seagrass zostera marina reveals angiosperm adaptation to the sea. *Nature* 530, 331–335. doi: 10.1038/nature16548

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Res.* 46, e126. doi: 10.1093/nar/gky730

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., et al. (2006). The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Res.* 35, D883–D887. doi: 10.1093/nar/gkl976

Park, I.-K., Lee, H.-S., Lee, S.-G., Park, J.-D., and Ahn, Y.-J.. (2000). Antifeeding activity of isoquinoline alkaloids identified in *Coptis japonica* roots against *Hyphantria cunea* (Lepidoptera: Arctiidae) and *Agelastica coerulea* (Coleoptera: Galerucinae). *J. Economic Entomology* 93, 331–335. doi: 10.1603/0022-0493-93.2.331

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., et al. (2019). Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* 20, 1–23. doi: 10.1186/s13059-019-1650-2

Rendon-Anaya, M., Ibarra-Laclette, E., and Mendez-Bravo, A. (2019). The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. *Proc. Natl. Acad. Sci. U.S.A.* 116, 17081–17089. doi: 10.1073/pnas.1822129116

Robin, A. Y., Giustini, C., Graindorge, M., Matringe, M., and Dumas, R.. (2016). Crystal structure of norcoclaurine-6-O-methyltransferase, a key rate-limiting step in the synthesis of benzylisoquinoline alkaloids. *Plant J.* 87, 641–653. doi: 10.1111/tpj.13225

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* 17, 155–158. doi: 10.1038/s41592-019-0669-3

Ryan, K. G., Swinny, E. E., Markham, K. R., and Winefield, C.. (2002). Flavonoid gene expression and UV photoprotection in transgenic and mutant *Petunia* leaves. *Phytochemistry* 59, 23–32. doi: 10.1016/S0031-9422(01)00404-6

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., et al. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485, 635–641. doi: 10.1038/nature11119

Schlapfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., et al. (2017). Genome-wide prediction of metabolic enzymes, pathways and gene clusters in plants. *Plant Physiol.* 173, 2041–2059. doi: 10.1104/pp.16.01942

Schultz, T. P., and Nicholas, D. D. (2000). Naturally durable heartwood: evidence for a proposed dual defensive function of the extractives. *Phytochemistry* 54, 47–52. doi: 10.1016/S0031-9422(99)00622-6

Shang, J., Tian, J., Cheng, H., Yan, Q., Li, L., Jamal, A., et al. (2020a). The chromosome-level wintersweet (Chimonanthus praecox) genome provides insights into floral scent biosynthesis and flowering in winter. *Genome Biol.* 21, 1–28. doi: 10.1186/s13059-020-02088-y

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Ioannidis, P., Kriventseva, E. V., Zdobnov, E. M., et al. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Teng, K., Tan, P., Guo, W., Yue, Y., Fan, X., Wu, J., et al. (2018). Heterologous expression of a novel *Zoysia japonica* C2H2 zinc finger gene, ZjZFN1, improved salt tolerance in *Arabidopsis*. *Front. Plant Sci.* 9, 1159. doi: 10.3389/fpls.2018.01159

The Angiosperm Phylogeny Group, Chase, M. W., Christenhusz, M. J. M., Fay, M. F., Byng, J. W., Judd, W. S., et al. (2016). An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. *Botanical J. Linn. Soc.* 181, 1–20. doi: 10.1111/boj.12385

Thévenin, J., Pollet, B., Letarnec, B., Saulnier, L., Gissot, L., Maia-Grondard, A., et al. (2011). The simultaneous repression of CCR and CAD, two enzymes of the lignin biosynthetic pathway, results in sterility and dwarfism in *Arabidopsis thaliana*. *Mol. Plant* 4, 70–82. doi: 10.1093/mp/ssq045

Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J., and Gao, G.. (2019). PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res.* 48, D1104–D1113. doi: 10.1093/nar/gkz1020

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, populus trichocarpa (Torr. & Gray). *Science* 313, 1596–1604. doi: 10.1126/science.1128691

Vance, C., Kirk, T., and Sherwood, R. (1980). Lignification as a mechanism of disease resistance. *Annu. Rev. Phytopathol.* 18, 259–288. doi: 10.1146/annurev.py.18.090180.001355

Van de Peer, Y. (2004). Computational approaches to unveiling ancient genome duplications. *Nat. Rev. Genet.* 5, 752–763. doi: 10.1038/nrg1449

Varoquaux, N., Liachko, I., Ay, F., Burton, J. N., Shendure, J., Dunham, M. J., et al. (2015). Accurate identification of centromere locations in yeast genomes using Hi-c. *Nucleic Acids Res.* 43, 5331–5339. doi: 10.1093/nar/gkv424

Varshney, R. K., Chen, W., Li, Y., Bharti, A. K., Saxena, R. K., Schlueter, J. A., et al. (2012). Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* 30, 83. doi: 10.1038/nbt.2022

Wang, G.-F., He, Y., Strauch, R., Olukolu, B. A., Nielsen, D., Li, X., et al. (2015). Maize homologs of hydroxycinnamoyltransferase, a key enzyme in lignin biosynthesis, bind the nucleotide binding leucine-rich repeat Rp1 proteins to modulate the defense response. *Plant Physiol.* 169, 2230–2243. doi: 10.1104/pp.15.00703

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49. doi: 10.1093/nar/gkr1293

Wei, T., and Simko, V. (2017). *R package "corrplot": Visualization of a correlation matrix (Version 0.84)*. Available from: http://CRAN.R-project.org/package=corrplot.

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., et al. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci.* 111, E4859–E4868. doi: 10.1073/pnas.1323926111

Xie, J., Qi, J., Huang, X., Zhou, N., and Hu, Y.. (2015). Comparative analysis of modern and ancient buried *Phoebe zhennan* wood: surface color, chemical components, infrared spectroscopy, and essential oil composition. *J. Forestry Res.* 26, 501–507. doi: 10.1007/s11676-015-0034-z

Yadav, V., Wang, Z., Wei, C., Amo, A., Ahmed, B., Yang, X., et al. (2020). Phenylpropanoid pathway engineering: An emerging approach towards plant defense. *Pathogens* 9, 312. doi: 10.3390/pathogens9040312

Yamada, Y., Kokabu, Y., Chaki, K., Yoshimoto, T., Ohgaki, M., Yoshida, S., et al. (2011). Isoquinoline alkaloid biosynthesis is regulated by a unique bHLH-type transcription factor in *Coptis japonica*. *Plant Cell Physiol.* 52, 1131–1141. doi: 10.1093/pcp/pcr062

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yu, G., Wang, L. G., Han, Y., and He, Q. Y.. (2012). clusterProfiler: an r package for comparing biological themes among gene clusters. *OMICS-A J. Integr. Biol.* 16, 284–287. doi: 10.1089/omi.2011.0118

Zeng, L., Zhang, Q., Sun, R., Kong, H., Zhang, N., Ma, H., et al. (2014). Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nat. Commun.* 5, 1–12. doi: 10.1038/ncomms5956

Zhang, L., Chen, F., Zhang, X., Li, Z., Zhao, Y., Lohaus, R., et al. (2020b). The water lily genome and the early evolution of flowering plants. *Nature* 577, 79–84. doi: 10.1038/s41586-019-1852-5

Zhang, C., Scornavacca, C., Molloy, E. K., and Mirarab, S.. (2020a). ASTRAL-pro: quartet-based species-tree inference despite paralogy. *Mol. Biol. Evol.* 37, 3292–3307. doi: 10.1093/molbev/msaa139

Zhao, T., Li, R., Yao, W., Wang, Y., Zhang, C., Li, Y., et al. (2021a). Genome-wide identification and characterisation of phenylalanine ammonia-lyase gene family in grapevine. *J. Hortic. Sci. Biotechnol.* 96, 456–468. doi: 10.1080/14620316.2021.1879685

Zhao, T., and Schranz, M. E. (2019). Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proc. Natl. Acad. Sci. United States America* 116, 2165–2174. doi: 10.1073/pnas.1801757116

Zhao, T., Wu, T., Pei, T., Wang, Z., Yang, H., Jiang, J., et al. (2021b). Overexpression of SlGATA17 promotes drought tolerance in transgenic tomato plants by enhancing activation of the phenylpropanoid biosynthetic pathway. *Front. Plant Sci.* 12, 634888–634888. doi: 10.3389/fpls.2021.634888

Zhao, T., Zwaenepoel, A., Xue, J. Y., Kao, S. M., Li, Z., Schranz, M. E., et al. (2021c). Whole-genome microsynteny-based phylogeny of angiosperms. *Nat. Commun.* 12, 3498. doi: 10.1038/s41467-021-23665-0

Zhou, M., and Memelink, J. (2016). Jasmonate-responsive transcription factors regulating plant secondary metabolism. *Biotechnol. Adv.* 34, 441–449. doi: 10.1016/j.biotechadv.2016.02.004

Zhou, X.-L., Zhang, L.-Q., Yang, L., Huang, F., Wang, Y.-H., Huang, X., et al. (2019). The complete chloroplast genome of cinnamomum pittosporoides reveals its phylogenetic relationship in lauraceae. *Mitochondrial DNA Part B* 4, 3246–3247. doi: 10.1080/23802359.2019.1669503

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 25 April 2023
DOI 10.3389/fpls.2023.1174972

OPEN ACCESS

EDITED BY
Kai-Hua Jia,
Shandong Academy of Agricultural Sciences, China

REVIEWED BY
Jie Gao,
Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences (CAS), China
Nian Wang,
Shandong Agricultural University, China

*CORRESPONDENCE
Hanhan Xia
xiahanhan@zhku.edu.cn

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 27 February 2023
ACCEPTED 22 March 2023
PUBLISHED 25 April 2023

CITATION
Huang W-C, Liao B, Liu H, Liang Y-Y, Chen X-Y, Wang B and Xia H (2023) A chromosome-scale genome assembly of *Castanopsis hystrix* provides new insights into the evolution and adaptation of Fagaceae species.
Front. Plant Sci. 14:1174972.
doi: 10.3389/fpls.2023.1174972

A chromosome-scale genome assembly of *Castanopsis hystrix* provides new insights into the evolution and adaptation of Fagaceae species

Wei-Cheng Huang[1,2,3†], Borong Liao[1†], Hui Liu[2,3], Yi-Ye Liang[2,3], Xue-Yan Chen[2,3], Baosheng Wang[2,3] and Hanhan Xia[1*]

[1]College of Horticulture and Landscape Architecture, Zhongkai University of Agriculture and Engineering, Guangzhou, China, [2]Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China, [3]South China National Botanical Garden, Chinese Academy of Sciences (CAS), Guangzhou, China

Fagaceae species dominate forests and shrublands throughout the Northern Hemisphere, and have been used as models to investigate the processes and mechanisms of adaptation and speciation. Compared with the well-studied genus *Quercus*, genomic data is limited for the tropical-subtropical genus *Castanopsis*. *Castanopsis hystrix* is an ecologically and economically valuable species with a wide distribution in the evergreen broad-leaved forests of tropical-subtropical Asia. Here, we present a high-quality chromosome-scale reference genome of *C. hystrix*, obtained using a combination of Illumina and PacBio HiFi reads with Hi-C technology. The assembled genome size is 882.6 Mb with a contig N50 of 40.9 Mb and a BUSCO estimate of 99.5%, which are higher than those of recently published Fagaceae species. Genome annotation identified 37,750 protein-coding genes, of which 97.91% were functionally annotated. Repeat sequences constituted 50.95% of the genome and LTRs were the most abundant repetitive elements. Comparative genomic analysis revealed high genome synteny between *C. hystrix* and other Fagaceae species, despite the long divergence time between them. Considerable gene family expansion and contraction were detected in *Castanopsis* species. These expanded genes were involved in multiple important biological processes and molecular functions, which may have contributed to the adaptation of the genus to a tropical-subtropical climate. In summary, the genome assembly of *C. hystrix* provides important genomic resources for Fagaceae genomic research communities, and improves understanding of the adaptation and evolution of forest trees.

KEYWORDS

***Castanopsis hystrix*, cellulose synthase (CesA) gene, chromosome-scale genome assembly, comparative genomic analysis, gene family**

1 Introduction

The Fagaceae family includes nine genera and roughly 900 species, which dominate forests and shrublands throughout the Northern Hemisphere (Oh and Manos, 2008; Petit et al., 2013). The three largest genera, *Quercus* (about 450 species), *Lithocarpus* (about 300 species), and *Castanopsis* (about 120 species) rapidly diverged after the Cretaceous-Paleogene boundary (K-Pg) (Zhou et al., 2022b) and currently occupy various habitats (Petit et al., 2013; Cannon et al., 2018). *Quercus* species are the dominant tree species of temperate forests in Eurasia and North American, while *Castanopsis* and *Lithocarpus* are mainly found in the tropical-subtropical evergreen forests of East and Southeast Asia (Petit et al., 2013; Cannon et al., 2018). Fagaceae species have been widely used as models of ecological and evolutionary genomic studies for the investigation of the processes and mechanisms of adaptation and speciation (Petit et al., 2013; Cavender-Bares, 2019; Kremer and Hipp, 2020). To date, more than 10 genomes of *Quercus* species have been assembled (Table 1), and the genomes of a dozen to one hundred individual oaks, such as those of *Q. acutissima* (Fu et al., 2022; Yuan et al., 2023), *Q. dentata* (Zhou et al., 2022a), *Q. petraea* (Leroy et al., 2020) and *Q. variabilis* (Liang et al., 2022) have been re-sequenced. By contrast, there is only a limited amount of genomic data available for the genus *Castanopsis*, and only one genome assembly (*C. tibetana*) is available for this genus (Sun et al., 2022). Molecular markers have been used to investigate the genetic diversity and evolutionary history of *Castanopsis* species (Shi et al., 2011; Li et al., 2014; Sun et al., 2014; Sun et al., 2016; Jiang et al., 2020; Li et al., 2022). However, our knowledge of the evolution of those species is incomplete or possibly biased due to a lack of sufficient genomic data. The availability of whole genome-wide data would provide an unprecedented opportunity for acquiring a deeper understanding of the adaptation and evolution of the genus *Castanopsis*, and would expand Fagaceae genome resources for comparative analysis.

Castanopsis hystrix (2n=2x=24) is one of the most important and dominant species of the tropical-subtropical evergreen forests of Asia (Li, 1996). In China, *C. hystrix* is naturally distributed in mixed and secondary forests, and its distribution extends from Nanling Mountain to Hainan Island and from Taiwan to south Tibet (Huang et al., 1999). *C. hystrix* is an ecologically and economically valuable species, and its forests play critical roles in water and soil conservation, disaster prevention, biodiversity, and the global carbon budget (Huang et al., 2015; You et al., 2018; Liang et al., 2019; Zhang et al., 2019a). *C. hystrix* is also a source of well-textured heartwood, which is widely used in furniture, construction, and shipbuilding, and it also produces seeds that can be used to extract tanning agents and starch (Chen et al., 1993; Chang et al., 1995). Due to the overexploitation of natural forests, the once widespread *C. hystrix* populations have been greatly diminished and fragmented (Zhao et al., 2020). High-quality genomic data are essential for assessing the patterns of genetic diversity, tracking the evolutionary history, and developing effective and efficient conservation strategies for this plant species. To date, only plastid and nuclear SSR markers have been used to investigate differences in the genetic diversity and divergence of *C. hystrix* (Li et al., 2007; Li et al., 2022); however, information on its nuclear genome is still unavailable.

In this study, we assembled and annotated the first chromosome-scale high-quality genome of *C. hystrix* by integrating PacBio HiFi long-reads, Illumina short-reads, RNAseq, and Hi-C sequencing data. We performed comparative genomic analysis to explore the evolution of genes, gene families, and genomes of *C. hystrix* and related Fagaceae species. Our study provides new insights into the genome evolution of Fagaceae tree species and provides essential genomic resources for germplasm conservation and genetic improvement of *C. hystrix*.

2 Material and methods

2.1 Plant sampling and genome sequencing

Fresh leaves were collected from an adult *C. hystrix* tree growing in Guangdong Fenghuangshan Forest Park (23.22° N, 113.39° E) and immediately frozen in liquid nitrogen until further use. Total genomic DNA was isolated from leave tissues using a DNeasy Plant MiniKit (Qiagen, Germany). The DNA quality and concentration were assessed by agarose gel electrophoresis and the Qubit Fluorometer (Thermo Fisher Scientific, USA). To obtain whole genome sequencing data, three DNA libraries were constructed and sequenced. First, an Illumina library with insert size of ~350 bp was sequenced on an Illumina NovaSeq 6000 platform with 150 bp paired-end reads. Second, a 20 kb HiFi library was prepared using the SMRTbell Express Template Preparation kit 2.0 (Pacific Biosciences, USA), and then sequenced on the Pacbio Sequel II platform to produce long-reads. Finally, a Hi-C sequencing library was constructed and sequenced on an Illumina NovaSeq 6000 platform (paired-end 150 bp).

Leaves at three different development stages (bud, immature, and mature) were collected from the same tree used for genome sequencing. Total RNA was extracted from leave samples using an RNAprep Pure Plus Kit (Tiangen, China), and the quality of RNA was evaluated using a Nanodrop spectrophotometer (Thermo Fisher Scientific, USA) and an Agilent 5400 (Agilent Technologies, USA). Total RNAs isolated from different leave tissues were mixed in equal amounts. A synthesized complementary DNA (cDNA) library was sequenced on an Illumina NovaSeq 6000 platform (paired-end 150 bp).

2.2 Genome survey and *de novo* assembly

To predict genomic characteristics, k-mer analysis was performed based on Illumina paired-end reads. The 17 bp *K*-mers were counted using Jellyfish v2.2.7 (Marcais and Kingsford, 2011), and genome size, heterozygosity, and repetitive element content were predicted based on the k-mer count distribution using GenomeScope v2.0 (Vurture et al., 2017).

The *de novo* assembly of *C. hystrix* genome was conducted in three steps by integrating Illumina short-reads, PacBio HiFi long-reads, and Hi-C sequencing data. First, the PacBio HiFi reads were

TABLE 1 Comparisons of genome assembly quality among 12 Fagaceae species.

Species	Sequencing platform	Genome size (Mb)	Percentage of scaffolds anchored to pseud-chromosome	Contig N50 (Mb)	Number of contigs	Scaffold N50 (Mb)	Number of scaffolds	BUSCOs (%)	No. of protein-coding genes	Average gene length (bp)	Percentage of repetitive sequences	Reference
Castanopsis hystrix	Illumina, Pacbio, Hi-C	882.69	98.07%	40.95	211	75.63	172	99.50%	37,750	4,819	50.95%	This study
Castanopsis tibetana	Illumina, ONT, Hi-C	878.64	98.67%	3.33	477	76.69	37	92.95%	40,937	4,857	54.30%	Sun et al. (2022)
Castanea crenata	Illumina, ONT, Hi-C	718.30	99.72%	6.36	206	NA	NA	97.60%	46,744	3,880	58.78%	Wang et al. (2022a)
Castanea mollissima	Illumina, PacBio, Hi-C	688.98	99.75%	2.83	671	57.34	112	92.44%	33,638	NA	53.24%	Wang et al. (2020)
Quercus acutissima	Illumina, PacBio, 10x Genomics	758.00	99.00%	1.44	770	2.89	388	90.50%	31,490	5,145	48.00%	Fu et al. (2022)
Quercus gilva	Illumina, PacBio, Hi-C	889.71	96.54%	28.32	773	70.35	515	98.60%	36,442	3,724	57.57%	Zhou et al. (2022c)
Quercus lobata	Illumina, PacBio, Hi-C	847.00	96.00%	1.90	NA	75.00	2014	95.00%	39,373	6,575	54.40%	Sork et al. (2022)
Quercus mongolica	Illumina, PacBio, Hi-C	809.84	95.65%	2.64	645	66.74	330	94.45%	36,553	6,085	53.75%	Ai et al. (2022)
Quercus robur	Illumina, Roche 454	789.35	96.00%	0.07	22,615	1.35	1409	91%	25,808	2,907	54.30%	Plomion et al. (2018)
Quercus suber	Illumina	953.00	Scaffold-level	0.08	36,760	0.50	23,344	95%	79,752	NA	51.60%	Ramos et al. (2018)
Quercus variabilis	DNBSEQ, PacBio, Hi-C	796.30	98.80	26.04	327	64.86	245	98%	32,466	5,272	67.60%	Han et al. (2022)
Fagus sylvatica	Illumina, PacBio, Hi-C	540.30	99.09%	0.14	6,650	46.56	167	97.40%	63,736	3,919	59.09%	Mishra et al. (2022)

NA, not reported in original paper.

10.3389/fpls.2023.1174972

error-corrected using NextDenovo v2.4.0 (https://github.com/Nextomics/NextDenovo), and were then initially assembled using Hifiasm v0.15.4 (Cheng et al., 2022). Second, the draft assembly was polished using NextPolish v1.3.1 (Hu et al., 2020), and redundant contigs were filtered using Redundans pipeline (Pryszcz and Gabaldón, 2016). Finally, contigs were linked to 12 pseudo-chromosomes of *C. hystrix* using ALLHiC (Zhang et al., 2019b) and Juicebox (Durand et al., 2016) based on Hi-C data. The quality of the genome assembly was evaluated using BUSCO (Benchmarking Universal Single-Copy Orthologs) (Seppey et al., 2019).

2.3 Prediction of genes and repetitive elements

The repeat regions, protein-coding genes, and non-coding RNA (ncRNA) were annotated in the *C. hystrix* genome assembly. Tandem repeats were identified using Tandem Repeats Finder v4.09 (Price et al., 2005), and dispersed repeats were identified by integrating *de novo* and homology-based methods. Briefly, *de novo* prediction was performed using LTR_FINDER v1.0.6 (Xu and Wang, 2007), LTR_retriever v2.9.0 (Ou and Jiang, 2018), RepeatScout v1.0.5 (Price et al., 2005), and RepeatModeler v2.0.1 (Flynn et al., 2020). The homology-based approach was conducted using Repeatmasker v4.1.0 (Chen, 2004). The *C. hystrix* assembly was searched against the RepBase library (Jurka et al., 2005) to identify sequences that are similar to known repetitive elements.

To annotate protein-coding genes, we conducted *de novo*, homology-based and RNA-Seq-assisted predictions on the repeat-masked *C. hystrix* genome. For *de novo* gene annotation, coding regions of genes were predicted using Augustus v3.2.3 (Stanke et al., 2006), Geneid v1.4 (Blanco et al., 2007), Genescan v1.0 (Burge and Karlin, 1997), GlimmerHMM v3.04 (Majoros et al., 2004), and SNAP (Aylor et al., 2006). For homology-based prediction, protein sequences of *Castanea mollissima* (Wang et al., 2020), *Castanopsis tibetana* (Sun et al., 2022), *Fagus sylvatica* (Mishra et al., 2018), *Quercus lobata* (Sork et al., 2016), *Quercus robur* (Plomion et al., 2018), and *Quercus suber* (Ramos et al., 2018) were downloaded from Genbank and aligned with the *C. hystrix* genome using TblastN v2.2.26 (Altschul et al., 1990). By comparing the homologous genome sequences to the matched proteins, gene models were constructed using GeneWise v2.4.1 (Birney et al., 2004). For RNA-Seq-based auxiliary prediction, a *C. hystrix* transcriptome was assembled using Trinity v2.1.1 (Grabherr et al., 2011) and aligned to the *C. hystrix* genome assembly using Hisat v2.0.4 (Kim et al., 2015). After that, gene models were predicted using PASA v2.0.2 (Keilwagen et al., 2016). Gene models predicted by the three methods were integrated using EvidenceModeler v1.1.1 (Haas et al., 2008), resulting in a non-redundant gene set. The ncRNAs, including rRNAs, micro RNAs (miRNAs), and small nuclear RNAs (snRNAs) were identified by searching the genome assembly against the Rfam database (Griffiths-Jones et al., 2003) with default parameters using Infernal v1.1 (Nawrocki and Eddy, 2013). tRNAs were predicted using the program tRNAscan-se v2.0 (Chan et al., 2021).

To infer gene functions, protein sequences were compared with those in Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), non-redundant (NR), Gene Ontology (GO) (Ashburner et al., 2000), SwissProt (Boeckmann et al., 2003), InterPro (Hunter et al., 2009), and protein family (Pfam) (Finn et al., 2014) databases using Blastp (E-value cutoff of $1e^{-5}$). The motifs and domains were characterized using InterProScan v5.31 (Zdobnov and Apweiler, 2001) by searching against public databases, including ProDom, PRINTS, Pfam, SMRT, PANTHER, and PROSITE.

2.4 Gene family evolution analyses

To track the gene family evolution, we analyzed the protein sequences of *C. hystrix* generated in this study together with those of 10 other species representing major lineages of Fagaceae and eudicots. Proteins of these species were downloaded from public databases. These species included *C. tibetana* (https://db.cngb.org; Accession number: CNA0019678), *C. mollissima* (https://ngdc.cncb.ac.cn; Accession number: GWHANWH00000000), and *Oryza sativa* (https://phytozome-next.jgi.doe.gov/info/Osativa_v7_0). Other seven species were downloaded from National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/), including *Fagus sylvatica* (GCA_907173295.1), *Juglans regia* (GCF_001411555.2), *Malus domestica* (GCA_002114115.1), *Prunus persica* (GCA_000346465.2), *Populus trichocarpa* (GCA_000002775.5), *Quercus robur* (GCA_932294415.1), *Vitis vinifera* (GCA_000003745.3). We identified orthologous genes using OrthoFinder v2.5.4 (Emms and Kelly, 2019), and then aligned gene coding regions using the package ParaAT v2.0 (Zhang et al., 2012). Single gene alignments were concatenated using seqkit v1.3 (Shen et al., 2016), and poorly aligned regions were excluded using Trimal v1.4 (Capella-Gutierrez et al., 2009). Then, a maximum likelihood (ML) tree was constructed based on the alignment of orthologous genes using IQ-TREE v2.1.2 (Nguyen et al., 2015), and dated using MCMCTree in the PAML v4.9j package (Yang, 2007). Two fossil calibrations were used to constrain the age of nodes. The first split within the Fagaceae family (genus *Fagus* vs. the rest of the genera) was constrained to 82–81 million years ago (Mya) (Grímsson et al., 2016), and the divergence time between genera *Castanopsis* and *Castanea* was restricted to 52.2 Mya (Wilf et al., 2019). Based on the dated phylogenetic tree, the expansion and contraction of gene families were inferred using CAFÉ v4.2.1 (De Bie et al., 2006).

2.5 Genome synteny and whole genome duplication analyses

To investigate the syntenic relationship between *C. hystrix* and relative species, proteins of *C. mollissima* (Wang et al., 2020) and *C. tibetana* (Sun et al., 2022) were downloaded from Genbank and compared with the genome of *C. hystrix* using Blastp (E-value cutoff of $1e^{-5}$). Collinear blocks were inferred using MCScanX (Wang et al., 2012) and visualized in JCVI v1.2.20 (Tang et al., 2008). The times of whole genome duplication (WGD) events were inferred

from the synonymous substitution rates (*Ks*) between paralogous and orthologous gene pairs. The *Ks* of gene pairs was calculated using the Nei-Gojobori algorithm as implemented in MCScanX.

2.6 Long terminal repeat retrotransposons analysis

To investigate the evolution of LTRs in *C. hystrix* and relative species, we identified LTRs in four Fagaceae species (*C. mollissima*, *C. tibetana*, *Q. robur*, and *F. sylvatica*) following the same procedure used for *C. hystrix* (see above). For full-length LTRs, the reverse transcriptase (RT) domains were identified using TEsorter v1.4.5 (Zhang et al., 2022), and were then aligned using MAFFT v7.475 (Katoh et al., 2002) with default parameters. The phylogenetic trees of LTRs were constructed based on the alignment of RT domains using FastTree v2.1.10 (Price et al., 2009). To estimate the insertion times (*T*) of full-length LTRs, the Kimura two-parameter distance (*K*) of each LTR-RT pair was calculated and converted to the insertion time using the formula $T = K/2\mu$, where the substitution rate (μ) was estimated using the baseml program in the PAML package.

2.7 Evolutionary analysis of the CesA gene family

Hard and well-textured heartwood are typical features of *C. hystrix* trees (Watanabe et al., 2014). Cell wall and lignin metabolic pathway genes are essential for wood formation. The cellulose synthase (CesA) gene family is involved in primary cell wall formation and cellulose synthase is considered the most important enzyme in the synthesis of cellulose microfibrils in plant cells (Kumar and Turner, 2015; Wang et al., 2022b). Hence, we conducted genome-wide characterization of the CesA family in *C. hystrix* and three relative Fagaceae species (*C. mollissima*, *C. tibetana* and *Q. robur*). The CesA genes in each species were identified using two methods. First, CesA protein sequences of *A. thaliana* (Persson et al., 2007) and *O. sativa* (Hazen et al., 2002) were blasted against the genomes of *C. hystrix*, *C. mollissima*, *C. tibetana*, and *Q. robur*, and homologous genes with an E-value cutoff of $1e^{-10}$ were identified. Second, two DNA-binding domains (PF03552 and PF00535) from Pfam (https://pfam.xfam.org/) were searched against protein sequences of Fagaceae species using HMMER v3.3.2 (Finn et al., 2011). The unions identified by both methods were considered to be common elements. To verify the reliability of the intersected results, we analyzed the completeness of CesA gene domains using Pfam and the conserved domain database (CDD, https://www.ncbi.nlm.nih.gov/cdd/). Then, the theoretical isoelectric points (PI) and molecular weights of CesA proteins were analyzed on the ExPASy website (https://web.expasy.org/compute_pi/).

For phylogenetic analysis, the amino acid sequences of each CesA member were aligned using MUSCLE v3.8 (Edgar, 2004), and phylogenetic trees were constructed using IQ-TREE with 1000 bootstraps and online visualization using iTOL (https://itol.embl.de/) (Letunic and Bork, 2019). To investigate in detail the classification of protein motifs, Multiple Em for Motif Elicitation (MEME) (http://memesuite.org/) was used to annotate the conserved motifs in these proteins. The maximum number of motifs was set to 10 and the motif width was set 10 to100 in MEME analysis. Blastp and MCScanX were used to identify syntenic blocks and duplication events with default parameters and visualization using TBtools (Chen et al., 2020).

3 Results

3.1 Genome assembly and assessment

The *C. hystrix* genome was assembled by using integrated multiple sequencing and assembly technologies. Whole genome sequencing resulted in 52.92 Gb of Illumina short-reads (~59×), 28.14 Gb of PacBio HiFi long-reads (~31×), and 141.12 Gb of Hi-C data (~160×). An initial genome survey using k-mer analysis estimated that the genome size of *C. hystrix* is about 897.51 Mb and that it has a high level of heterozygosity of 1.26% and a repeat content of 57.38% (Table S1). Illumina short-reads, PacBio HiFi long-reads, and Hi-C sequencing data revealed that the assembled *C. hystrix* genome is 882.69 Mb, including 211 contigs and 172 scaffolds (Table 1). The contig N50 and scaffold N50 length are 40.95 Mb and 75.63 Mb, respectively. In total, 865.64 Mb (98.07%) of assembled sequences were mounted on 12 pseudo-chromosomes ranging from 51.51 Mb to 103.15 Mb (Figure 1A, Table 1). The heat map of Hi-C interactions shows that the genome assembly is intact and robust (Figure 1B).

The high accuracy and completeness of the *C. hystrix* genome assembly was supported by three analyses. First, joint analysis of GC content and sequencing depth revealed no obvious deviation in quality across the genome, suggesting the high quality of genome sequencing and assembly (Figures 1, S1). Second, approximately 97.66% of cleaned PacBio HiFi long-reads were successfully mapped to the genome, and more than 99% of the genome assembly had a coverage >10× (Table S2), suggesting that the genome assembly was accurate and complete. Finally, BUSCO analyses revealed that 99.5% of universal single-copy orthologs were present in the genome assembly (Table 1), indicating the high integrity of the genome assembly.

3.2 Genome annotation

A total of 449.72 Mb (50.95%) of the *C. hystrix* genome was annotated as repetitive sequences (Tables 1, S3). The most abundant repetitive elements were LTRs (374.50 Mb), followed by tandem repeats (47.64 Mb), long interspersed nuclear elements (LINEs; 18.08 Mb), DNA transposons (12.90 Mb), and short interspersed nuclear elements (SINEs; 16,791 bp) (Table S3).

By integrating *de novo*, homology-based, and RNA-Seq-assisted predictions, a total of 37,750 protein-coding genes were predicted in the *C. hystrix* genome (Tables 1, S4). The average lengths of coding sequences (CDSs), exons and introns are 1,067 bp, 244 bp and 1,112

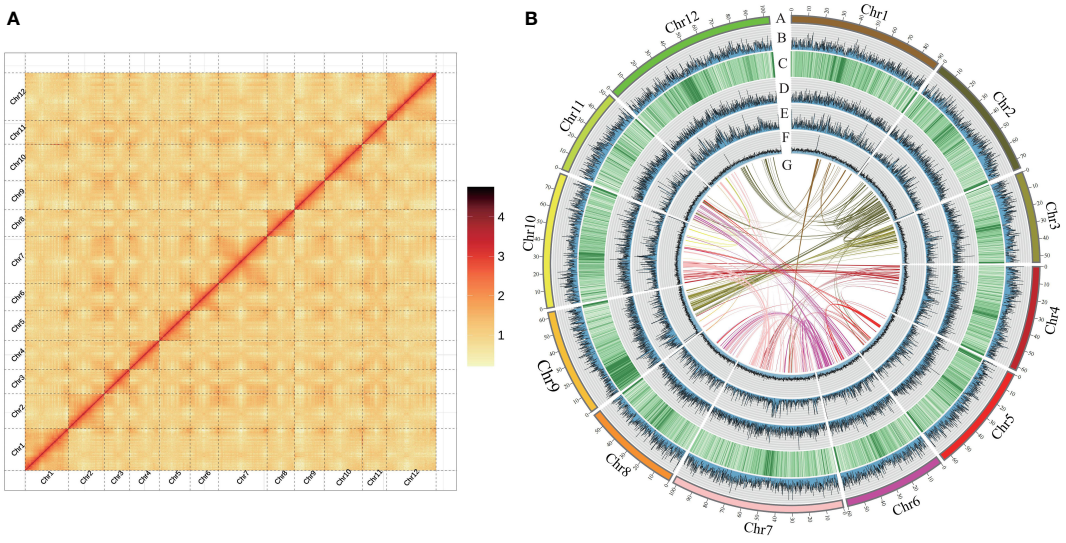


FIGURE 1
Features of *Castanopsis hystrix* genome. **(A)** Genome-wide analysis of chromatin interactions in the *C hystrix* genome based on Hi-C data. **(B)** The Synteny and distribution of genomic features. (A) The 12 pseudochromosomes; (B) gene density; (C–E) the density of total repeat sequences, Gypsy LTR-RTs, and Copia LTR-RTs; (F) histogram of GC content; (G) intragenomic collinearity. (B–F) were drawn in 100 kb overlapping sliding windows.

bp, respectively (Table S4). By comparing the predicted gene set with six public databases, 36,962 (97.91%) of the total predicted genes were functionally annotated (Table S5). Non-coding RNA annotation identified 922 miRNAs, 741 tRNAs, 8,971 rRNAs, and 665 snRNAs in *C. hystrix* (Table S6).

3.3 Gene family evolution in *C. hystrix*

To explore the evolutionary history of the *C. hystrix* gene family, we clustered 36,448 (96.6%) annotated genes into 19,143 gene families. Among these, 12,573 gene families were shared with those of four other studied Fagaceae species (Figure 2A), and 299 families (1,043 genes) were unique to *C. hystrix*. Functional enrichment analysis showed that unique genes of *C. hystrix* were significantly enriched in 10 KEGG pathways and 115 GO terms, including Fatty acid biosynthesis, Porphyrin and chlorophyll metabolism, malate transport, and polynucleotide adenylyltransferase activity (Table S7; Figure S2).

A phylogenetic tree constructed using 556 single-copy orthologs among *C. hystrix* and other 10 angiosperms revealed that two *Castanopsis* species (*C. hystrix* and *C. tibetana*) were grouped together, and these two species are sister to a *Castanea* species (*C. mollissima*) (Figure 2B). Calibration of the phylogenetic tree using two Fagaceae fossil records showed that the divergence time between *C. hystrix* and *C. tibetana* is 30.4 Mya (95% HPD: 19.6–40.2 Mya) (Figures 2B, S3). The close phylogenetic relationships between *Castanopsis* and *Castanea* species were supported by the high genome synteny and colinearity (Figure 2C).

Based on the clustered gene families and dated phylogenetic tree, CAFÉ analyses detected 2283 expanded gene families and 2505 contracted gene families in *C. hystrix* (Figure 2B; Tables S8). Among these, 202 expanded and 62 contracted gene families were statistically significant ($P < 0.01$; Table S8). The 202 expanded gene families were enriched in 7 KEGG pathways and 36 GO terms, such as "Arginine and proline metabolism", "Phenylalanine metabolism", "Fatty acid degradation", and "Trehalose biosynthetic process" (Table S9; Figure S2). The 62 contracted gene families were primarily enriched in KEGG pathway processes "Sesquiterpenoid and triterpenoid biosynthesis", "Plant-pathogen interaction", and "MAPK signaling" (Table S9). A search of *C. hystrix* expanded genes families against PlantTFDB (http://planttfdb.gao-lab.org/) revealed that 29 genes were categorized into four transcription factors (TFs) families (FAR1, B3, bHLH, and NAC). Among these, 23 genes belong to the FAR1 family, and the other six genes belong to B3 (one gene), bHLH (two genes), and NAC (three genes) families (Table S10). We also found that 17 and 16 gene families significantly expanded and contracted, respectively, in the most common ancestor of *C. hystrix* and *C. tibetana*. Functional enrichment analysis revealed that the 17 expanded gene families were overrepresented in 11 KEGG pathways and 8 GO terms, including "Fatty acid degradation", "Plant-pathogen interaction" and "RNA-DNA hybrid ribonuclease activity" (Table S9). The 16 contracted gene families were enriched in six KEGG pathways and four GO terms (Table S9).

3.4 WGD in *C. hystrix*

Comparative genomic analyses were performed to discern the number of WGD events in *C. hystrix*. A total of 65 syntenic blocks (2,442 collinear genes) with sizes ranging from 11 to 48 gene pairs were detected in *C. hystrix*, accounting for 6.47% of the total gene set. The number of collinear genes in *C. hystrix* was close to those of other Fagaceae species (2484–2673 genes; 6.53%–7.71% of the total gene set) but lower than that in *V. vinifera* (3297 genes; 12.85% of

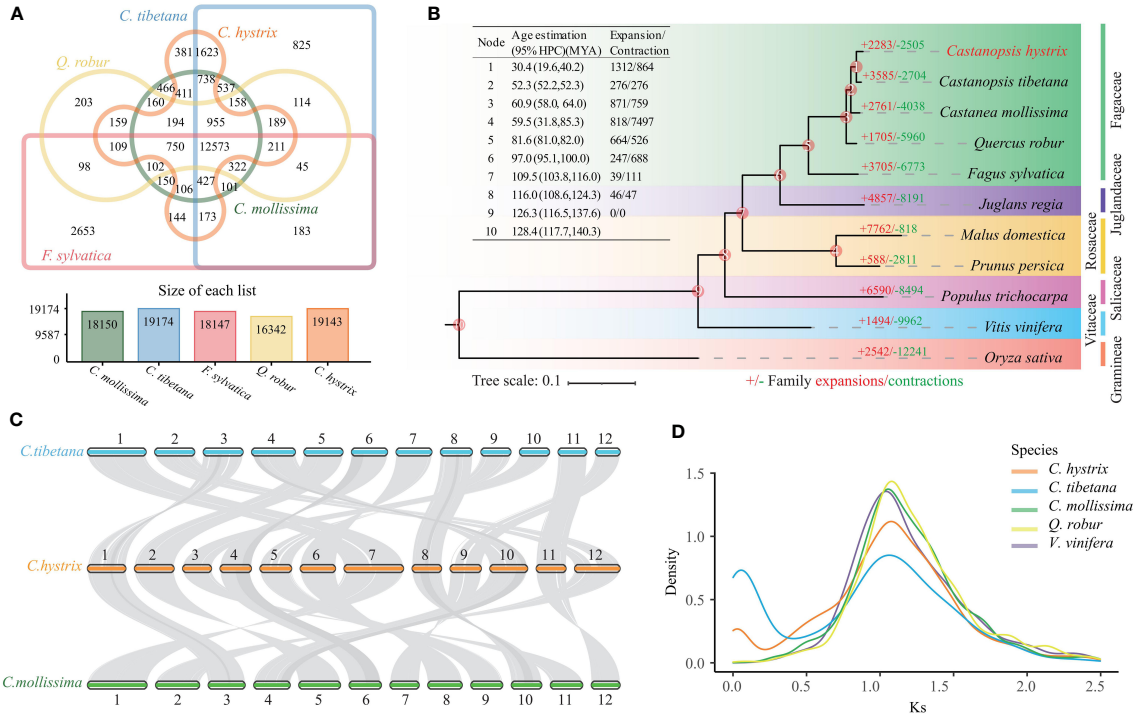


FIGURE 2
Genomic evolutionary and comparative genomic analyses. **(A)** Shared and unique gene families in *C hystrix*, *C tibetana*, *C mollissima*, *Q. robur*, and *F sylvatica*. **(B)** Phylogenomic tree and expansion and contraction of gene families among *C hystrix* and 10 other species. Numbers in red (+) and green (–) show the number of expanded and contracted gene families, respectively. **(C)** The synteny blocks between *C hystrix*, *C tibetana*, and *C mollissima*. Syntenic blocks were connected by grey lines. **(D)** The synonymous substitution rates (Ks) distributions of paralogous genes.

the total gene set) (Table S11). The *Ks* values of paralogous and orthologous gene pairs showed that all four Fagaceae species and *V. vinifera* shared a *Ks* peak of approximately 1.08 units (Figure 2D), most likely representing the triplication event (γ) shared by all eudicots (Murat et al., 2015). Synteny analysis revealed a 1:1 syntenic depth ratio for *C. hystrix* vs. Fagaceae species and a 2:2 syntenic depth ratio for *C. hystrix* vs. *V. vinifera* (Figure S4). These results suggested that no independent WGD events have occurred in *C. hystrix* and other Fagaceae species.

3.5 Expansion of LTRs in *C. hystrix*

Copia and Gypsy are the two most abundant LTR super families in *C. hystrix* and three other Fagaceae species. In *C. hystrix*, Copia- and Gypsy-type LTRs accounted for 37.49% and 38.04% of LTRs, respectively (Figure 3A; Table S12). The content of Copia- and Gypsy-type LTRs was slightly different among Fagaceae species (Figure 3A; Table S12), indicating independent expansion or elimination of repetitive elements. Phylogenetic analyses using RT domains of LTRs revealed that Copia-type elements were clustered into seven major groups, with Ale-type repeats forming the largest group (N = 355) followed by Angela (N = 320), SIRE (N = 235), Tork (N = 46), TAR (N = 29), Ikeros (N = 28), and Ivana (N = 21; Figure 3B). The Gypsy-type elements were grouped into six clades, and the OTA group accounted for 91% (1,658) of Gypsy members (Figure 3B). Full analyses with all Gypsy and Copia elements from the five Fagaceae species showed that the lineages of Copia and Gypsy were grouped according to their respective tribes, indicating different evolutionary relationships among LTR families (Figure S5). To further explore the details of LTR expansion, we estimated the insertion time of full-length LTRs. In *C. hystrix*, the insertion time peaks of both Copia- and Gypsy-type LTRs were found approximately at 2 Mya, while a more ancient amplification peak was found around 8 Mya in *C. tibetana* (Figure 3C). In other Fagaceae species, a significant burst of LTRs was detected at 1–3 Mya, but the extent of expansion varied among species and was also different between Copia- and Gypsy-type LTRs (Figure 3C).

3.6 Evolution of the CesA gene family

Genome-wide characterization of the CesA family in *C. hystrix* identified 34 CesA-like genes (Figure 4A; Table S13). Phylogenetic analysis suggested that these genes could be divided into seven subfamilies (CesA, CslA–CslH) (Figure 4C; Table S13). Genes from the same subfamily showed similar protein domains and motif compositions, supporting their phylogenetic relationships (Figures 4D, S6). Similar numbers of CesA-like genes were found in three closely related Fagaceae species (41, 46, and 45 genes in *C. tibetana*, *C. mollissima*, and *Q. robur*, respectively) and two distinct related species, *A. thaliana* (40 genes) and *O. sativa* (45 genes)

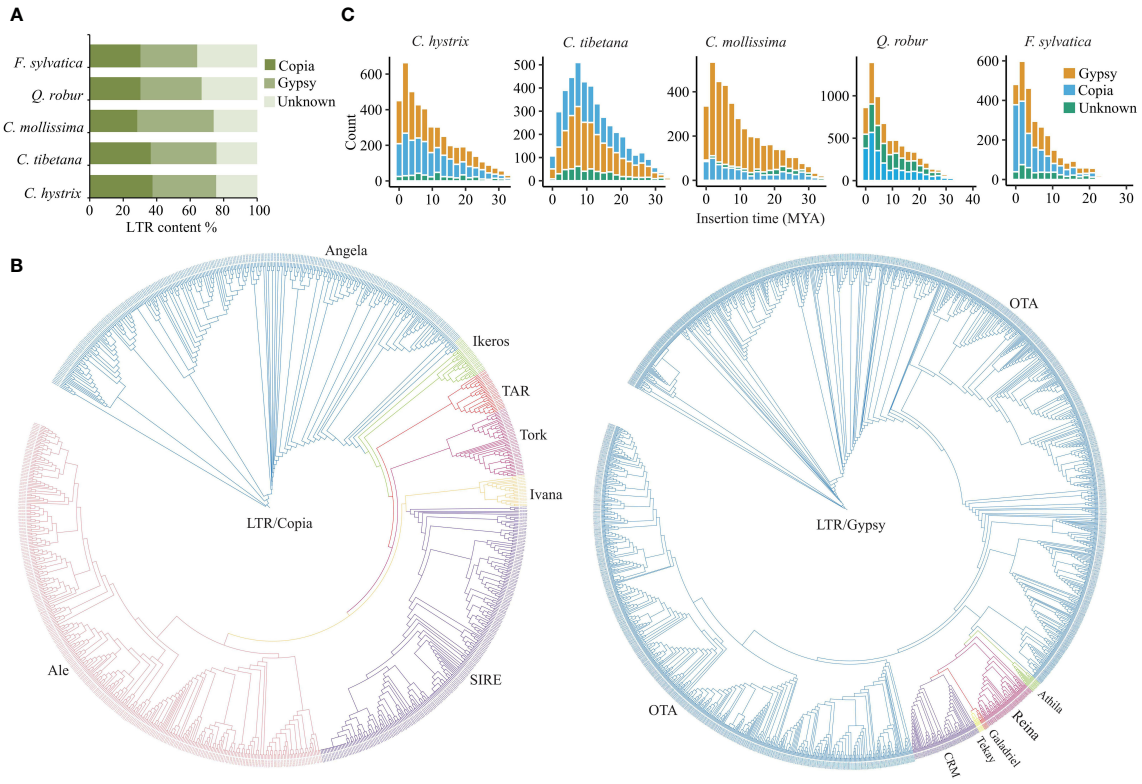


FIGURE 3
The features of LTR expansion in the Fagaceae genomes. **(A)** Comparison of LTR contents in *C hystrix* and 4 other species. **(B)** Neighbor-joining trees of Copia and Gypsy LTRs from *C hystrix*. **(C)** Insertion time estimates of full-length LTRs in five Fagaceae species.

(Figure 4A; Table S14). However, Fagacea species showed different CesA subfamily content to that of *A. thaliana* and *O. sativa* (Figure 4A). For example, the number of CslE and CslG genes in Fagaceae species (6–12 and 4–9, respectively) was much higher than in *A. thaliana* (one and three, respectively) and *O. sativa* (three and none, respectively) (Figure 4A). Nine CslA genes were identified in *A. thaliana* and *O. sativa*, but only three CslA members were found in Fagaceae species (Figure 4A). In addition, the collinearity of CesA-like gene between *C. hystrix* and other Fagaceae species was clearly higher than those for *C. hystrix* vs. *A. thaliana* and *O. sativa* (Figures 4B, S7). An analysis of the distribution of CesA-like genes across the genome of *C. hystrix* revealed tandem duplication of 10 CesA genes (Figure S8).

4 Discussion

In this study, we generated a high-quality chromosome-scale assembly of *C. hystrix*. The assembled genome was approximately 882.6 Mb, of which more than 98% of the sequences were anchored to 12 pseudo-chromosomes ranging from 51.5 to 103.2 Mb in size. The contig N50 of the *C. hystrix* genome assembly was 40.95 Mb, which is higher than those of recently published Fagaceae species, such as *C. tibetana* (3.32 Mb) (Sun et al., 2022), *C. mollissima* (2.83 Mb) (Wang et al., 2020), *Castanea crenata* (6.36 Mb) (Wang et al., 2022a), *Quercus gilva* (28.32 Mb) (Zhou et al., 2022c), *Q. lobata* (1.90 Mb) (Sork et al., 2022), *Quercus variabilis* (26.04 Mb) (Han et al., 2022), and *F. sylvatica* (0.14 Mb) (Mishra et al., 2022). Genome assembly integrity, as assessed by BUSCO, reached 99.5% for *C. hystrix*, surpassing that of previously assembled Fagaceae genomes (90.5%–98.6%; Table 1). The high quality of the genome assembly can be mainly attributed to the successful implementation of new sequencing technologies, a statistical algorithm, and analytical approaches. Although gap-free T2T genomes are available in model species (Naish et al., 2021; Song et al., 2021), *de novo* genome assembly is still challenging for forest trees because of their large and complex genomes. Our genome assembly of *C. hystrix* is one of the most high-quality genomes of Fagaceae species ever reported.

Based on comparative genome analysis, we found high genome synteny between *C. hystrix* and *C. tibetana* and *C. mollissima*, although these species diverged more than 30 million years ago (Zhou et al., 2022b). We also found that *C. hystrix* and other investigated Fagaceae species did not experience WGD after the triplication event (γ) (Murat et al., 2015). These results are consistent with the previous hypothesis that ploidy level and genome structure are conserved among Fagaceae species, which may have facilitated the adaptive introgression between species (Chen et al., 2014; Cannon and Petit, 2020). Transposable elements (TEs) account for large parts of plant genomes, where they play an important role in evolution (Bennetzen and Wang, 2014; Akakpo et al., 2020). The proportion of the repetitive elements in the *C.*

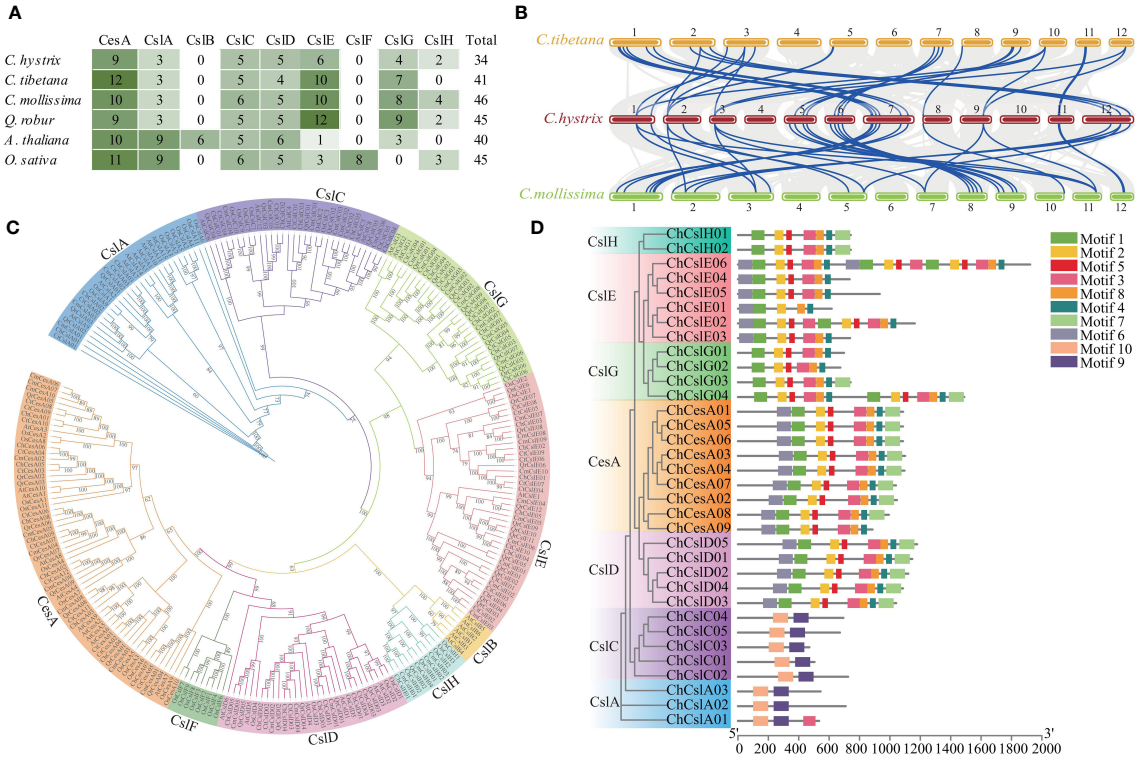


FIGURE 4
Identification and Evolution of CesA family in Fagaceae. **(A)** The heatmap shown a comparison of the numbers of CesA genes among four Fagaceae plants, *Arabidopsis thaliana*, and *Oryza sativa*. **(B)** Synteny analysis of CesA genes between *C hystrix*, *C tibetana*, and *C mollissima*. The blue lines highlight the syntenic CesA gene pairs. **(C)** Phylogenetic tree of CesA gene families in four Fagaceae plants, *A thaliana*, and *O. sativa*. **(D)** Phylogenetic relationships and architecture of the conserved protein motifs in 34 CesA genes from *C hystrix*.

hystrix genome was 50.95%, similar to that reported for other Fagaceae, such as *C. tibetana* (54.30%) (Sun et al., 2022), *C. mollissima* (53.24%) (Wang et al., 2020), *Q. mongolica* (53.75%) (Ai et al., 2022), and *Q. variabilis* (26.04 Mb) (Han et al., 2022). Evolutionary analyses of LTRs showed that *C. hystrix* and relative Fagaceae species experienced a recent large-scale LTR burst, but the time and extent of LTR expansion varied between species and between LTR families, which may have influenced the structure and function of genomes and contributed to the adaptation and evolution of Fagaceae species.

Whole genome annotation and analysis revealed considerable gene family expansion and contraction in *C. hystrix* and relative species. These expanded and contracted gene families were involved in multiple important biological processes and molecular functions, providing valuable information for understanding the genetic basis of adaptation, evolution, and speciation in Fagaceae. For example, 17 gene families expanded in the most recent ancestor of *C. tibetana* and *C. hystrix*, and 202 gene families independently expanded in *C. hystrix*. Functional enrichment analysis suggested that the 17 expanded gene families were highly overrepresented in stress and defense-associated pathways, such as plant–pathogen interaction and Fatty acid degradation (Kindl, 1993; Goepfert and Poirier, 2007; Dodds and Rathjen, 2010; Chhajed et al., 2020). Fatty acid degradation is essential for seed development, seed germination, and post-germinative growth before the establishment of photosynthesis (Kindl, 1993; Goepfert and Poirier, 2007). In addition, expanded gene families in *C. hystrix* were enriched in the biological processes "Phenylpropanoids", which influences plant responses to biotic and abiotic stimuli (La Camera et al., 2004; Vogt, 2010), and "Arginine and proline metabolism", which plays key roles in nitrogen distribution and recycling in plants (Slocum, 2005; Rennenberg et al., 2010). Several expanded genes in *C. hystrix* are also members of the transcription factor family FAR1, which modulates phyA signaling (Lin et al., 2007) and regulates the balance between growth and defense under shade conditions (Liu et al., 2019). Therefore, the gene family expansions might have facilitated the adaptation of the genus *Castanopsis* to a tropical-subtropical climate, after they had diverged from their deciduous counterparts in cool-temperate areas. Furthermore, CslE/CslG genes of the CesA family exhibited expansion and tandem duplication in Fagaceae species. CesA genes are involved in the biosynthesis of various polysaccharide polymers, in particular hemicelluloses (Richmond and Somerville, 2000; Lerouxel et al., 2006). A recent study suggested that the expansion of the CesA family might have contributed to the formation of the high-density timbers that are characteristic of Dipterocarpaceae species (Wang et al., 2022b). Thus, we suspect that CesA gene expansion might be related to the development of the high-density woods of Fagaceae species. Taken together, these considerations suggest that gene family expansions might have played critical roles in the

genetic, morphological, and physiological innovations of Fagaceae species.

In conclusion, we obtained the first chromosome-scale genome assembly of *C. hystrix* using a combination of multiple sequencing and assembly approaches. Genome-wide characterization and evolutionary analysis provided novel insights into the genome evolution and key regulatory pathways of wood formation in Fagaceae species. The *C. hystrix* genome assembly contains both high-quality reference sequences and important functional genes, which expands the genome resources for Fagaceae species and opens the possibility of conducting comparative and functional genomic studies of forest tree species.

Data availability statement

The data presented in the study are deposited in the National Genomics Data Center (NGDC) database, BioProject accession number PRJCA015225.

Author contributions

HX designed this study. BW and Y-YL collected samples. W-CH, BL, HL, and X-YC analyzed the data. HX, W-CH, and BL wrote the paper. All authors read and approved the final manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (no. 32001244).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1174972/full#supplementary-material

References

Ai, W., Liu, Y., Mei, M., Zhang, X., Tan, E., Liu, H., et al. (2022). A chromosome-scale genome assembly of the Mongolian oak (*Quercus mongolica*). *Mol. Ecol. Resour.* 22, 2396–2410. doi: 10.1111/1755-0998.13616

Akakpo, R., Carpentier, M. C., Ie, H. Y., and Panaud, O. (2020). The impact of transposable elements on the structure, evolution and function of the rice genome. *New Phytol.* 226, 44–49. doi: 10.1111/nph.16356

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Aylor, D. L., Price, E. W., and Carbone, I. (2006). SNAP: Combine and map modules for multilocus population genetic analysis. *Bioinformatics* 22, 1399–1401. doi: 10.1093/bioinformatics/btl136

Bennetzen, J. L., and Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.* 65, 505–530. doi: 10.1146/annurev-arplant-050213-035811

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and genomewise. *Genome Res.* 14, 988–995. doi: 10.1038/sdata.2018.69

Blanco, E., Parra, G., and Guigó, R. (2007). Using geneid to identify genes. *Curr. Protoc. Bioinf.* 18, 4–3. doi: 10.1002/0471250953.bi0403s18

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31, 365–370. doi: 10.1093/nar/gkg095

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94. doi: 10.1006/jmbi.1997.0951

Cannon, C. H., Brendel, O., Deng, M., Hipp, A. L., Kremer, A., Kua, C. S., et al. (2018). Gaining a global perspective on fagaceae genomic diversification and adaptation. *New Phytol.* 218, 894–897. doi: 10.1111/nph.16091

Cannon, C. H., and Petit, R. J. (2020). The oak syngameon: more than the sum of its parts. *New Phytol.* 226, 978–983. doi: 10.1111/nph.16091

Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009). TrimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Cavender-Bares, J. (2019). Diversification, adaptation, and community assembly of the American oaks (*Quercus*), a model clade for integrating ecology and evolution. *New Phytol.* 221, 669–692. doi: 10.1111/nph.15450

Chan, P. P., Lin, B. Y., Mak, A. J., and Lowe, T. M.. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* 49, 9077–9096. doi: 10.1093/bioinformatics/btt509

Chang, C., Lin, M., Lee, S., Liu, K. C. C., Hsu, F. L., and Lin, J. Y. (1995). Differential inhibition of reverse transcriptase and cellular DNA polymerase-α activities by lignans isolated from Chinese herbs, *Phyllanthus myrtifolius* moon, and tannins from *Lonicera japonica* thunb and *Castanopsis hystrix*. *Antiviral Res.* 27, 367–374. doi: 10.1016/0166-3542(95)00020-M

Chen, N. (2004). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinf.* 5, 4–10. doi: 10.1002/0471250953.bi0410s05

Chen, S. C., Cannon, C. H., Kua, C. S., Liu, J. J., and Galbraith, D. W. (2014). Genome size variation in the fagaceae and its implications for trees. *Tree Genet. Genomes* 10, 977–988. doi: 10.1007/s11295-014-0736-y

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chen, H., Tanaka, T., Nonaka, G., Fujioka, T., and Mihashi, K. (1993). Hydrolysable tannins based on a triterpenoid glycoside core, from *Castanopsis hystrix*. *Phytochemistry* 32, 1457–1460. doi: 10.1016/0031-9422(93)85159-O

Cheng, H., Jarvis, E. D., Fedrigo, O., Koepfli, K. P., Urban, L., Gemmell, N., et al. (2022). Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* 40, 1332–1335. doi: 10.1038/s41587-022-01261-x

Chhajed, S., Mostafa, I., He, Y., Abou-Hashem, M., El-Domiaty, M., and Chen, S. (2020). Glucosinolate biosynthesis and the glucosinolate-myrosinase system in plant defense. *Agronomy* 10, 1786. doi: 10.3390/agronomy10111786

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: A computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271. doi: 10.1093/bioinformatics/btl097

Dodds, P. N., and Rathjen, J. P. (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. *Nat. Rev. Genet.* 11, 539–548. doi: 10.1038/nrg2812

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., et al. (2016). Juicebox provides a visualization system for Hi-c contact maps with unlimited zoom. *Cell Syst.* 3, 99–101. doi: 10.1016/j.cels.2015.07.012

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Emms, D. M., and Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238. doi: 10.1186/s13059-019-1832-y

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi: 10.1093/nar/gkr367

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., et al. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* 117, 9451–9457. doi: 10.1073/pnas.1921046117

Fu, R., Zhu, Y., Liu, Y., Feng, Y., Lu, R. S., Li, Y., et al. (2022). Genome-wide analyses of introgression between two sympatric Asian oak species. *Nat. Ecol. Evol.* 6, 924–935. doi: 10.1038/s41559-022-01754-7

Goepfert, S., and Poirier, Y. (2007). β-oxidation in fatty acid degradation and beyond. *Curr. Opin. Plant Biol.* 10, 245–251. doi: 10.1016/j.pbi.2007.04.007

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R. (2003). Rfam: an RNA family database. *Nucleic Acids Res.* 31, 439–441. doi: 10.1093/nar/gkg006

Grímsson, F., Grimm, G. W., Zetter, R., and Denk, T. (2016). Cretaceous And paleogene fagaceae from north America and Greenland: evidence for a late Cretaceous split between fagus and the remaining fagaceae. *Acta Palaeobotanica* 56, 247–305. doi: 10.1515/acpa-2016-0016

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9, 1–22. doi: 10.1186/gb-2008-9-1-r7

Han, B., Wang, L., Xian, Y., Xie, X. M., Li, W. Q., Zhao, Y., et al. (2022). A chromosome-level genome assembly of the Chinese cork oak (*Quercus variabilis*). *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1001583

Hazen, S. P., Scott-Craig, J. S., and Walton, J. D. (2002). Cellulose synthase-like genes of rice. *Plant Physiol.* 128, 336–340. doi: 10.1104/pp.010875

Hu, J., Fan, J., Sun, Z., and Liu, S. (2020). NextPolish: A fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* 36, 2253–2255. doi: 10.1093/bioinformatics/btz891

Huang, C., Zhang, Y., and Bruce, B. (1999). "Fagaceae," in *Flora of China*, Eds. Z. Y. Wu, P. H. Raven and D. Y. Hong (Beijing, China: Science Press and Missouri Botanical Garden Press) 4, 314–400. Available at: http://flora.huh.harvard.edu/china/mss/volume04/FAGACEAE.published.pdf.

Huang, W., Zhou, G., Deng, X., Liu, J., Duan, H., Zhang, D., et al. (2015). Nitrogen and phosphorus productivities of five subtropical tree species in response to elevated CO2 and n addition. *Eur. J. For. Res.* 134, 845–856. doi: 10.1007/s10342-015-0894-y

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res.* 37, D211–D215. doi: 10.1093/nar/gkn785

Jiang, K., Xie, H., Liu, T., Liu, C., and Huang, S. (2020). Genetic diversity and population structure in *Castanopsis fissa* revealed by analyses of sequence-related amplified polymorphism (SRAP) markers. *Tree Genet. Genomes* 16, 52. doi: 10.1007/s11295-020-01442-2

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic Genome Res.* 110, 462–467. doi: 10.1159/000084979

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Grau, J., and Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* 44, e89–e89. doi: 10.1093/nar/gkw092

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Kindl, H. (1993). Fatty acid degradation in plant peroxisomes: function and biosynthesis of the enzymes involved. *Biochimie* 75, 225–230. doi: 10.1016/0300-9084(93)90080-C

Kremer, A., and Hipp, A. L. (2020). Oaks: an evolutionary success story. *New Phytol.* 226, 987–1011. doi: 10.1111/nph.16274

Kumar, M., and Turner, S. (2015). Plant cellulose synthesis: CESA proteins crossing kingdoms. *Phytochemistry* 112, 91–99. doi: 10.1016/j.phytochem.2014.07.009

La Camera, S., Gouzerh, G., Dhondt, S., Hoffmann, L., Fritig, B., Legrand, M., et al. (2004). Metabolic reprogramming in plant innate immunity: the contributions of phenylpropanoid and oxylipin pathways. *Immunol. Rev.* 198, 267–284. doi: 10.1111/j.0105-2896.2004.0129.x

Lerouxel, O., Cavalier, D. M., Liepman, A. H., and Keegstra, K. (2006). Biosynthesis of plant cell wall polysaccharides - a complex process. *Curr. Opin. Plant Biol.* 9, 621–630. doi: 10.1016/j.pbi.2006.09.009

Leroy, T., Louvet, J. M., Lalanne, C., Le Provost, G., Labadie, K., Aury, J. M., et al. (2020). Adaptive introgression as a driver of local adaptation to climate in European white oaks. *New Phytol.* 226, 1171–1182. doi: 10.1111/nph.16095

Letunic, I., and Bork, P. (2019). Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259. doi: 10.1093/nar/gkz239

Li, J. Q. (1996). The origin and distribution of the family fageceae. *Acta Phytotaxon. Sin.* 34, 376–396. Available at: https://www.jse.ac.cn/EN/Y1996/V34/I4/376.

Li, J., Ge, X., Cao, H., and Ye, W. H. (2007). Chloroplast DNA diversity in *Castanopsis hystrix* populations in south China. *For. Ecol. Manage.* 243, 94–101. doi: 10.1016/j.foreco.2007.02.012

Li, C., Sun, Y., Huang, H. W., and Cannon, C. H. (2014). Footprints of divergent selection in natural populations of *Castanopsis fargesii* (Fagaceae). *Heredity* 113, 533–541. doi: 10.1038/hdy.2014.58

Li, N., Yang, Y., Xu, F., Chen, X., Wei, R., Li, Z., et al. (2022). Genetic diversity and population structure analysis of *Castanopsis hystrix* and construction of a core collection using phenotypic traits and molecular markers. *Genes* 13, 2383. doi: 10.3390/genes13122383

Liang, X., He, P., Liu, H., Zhu, S., Uyehara, I. K., Hou, H., et al. (2019). Precipitation has dominant influences on the variation of plant hydraulics of the native *Castanopsis fargesii* (Fagaceae) in subtropical China. *Agric. For. Meteorol.* 271, 83–91. doi: 10.1016/j.agrformet.2019.02.043

Liang, Y. Y., Shi, Y., Yuan, S., Zhou, B. F., Chen, X. Y., An, Q. Q., et al. (2022). Linked selection shapes the landscape of genomic variation in three oak species. *New Phytol.* 233, 555–568. doi: 10.1111/nph.17793

Lin, R., Ding, L., Casola, C., Ripoll, D. R., Feschotte, C., and Wang, H. (2007). Transposase-derived transcription factors regulate light signaling in *Arabidopsis*. *Science* 318, 1302–1305. doi: 10.1126/science.1146281

Liu, Y., Wei, H., Ma, M., Li, Q., Kong, D., Sun, J., et al. (2019). *Arabidopsis* FHY3 and FAR1 regulate the balance between growth and defense responses under shade conditions. *Plant Cell* 31, 2089–2106. doi: 10.1105/tpc.18.00991

Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20, 2878–2879. doi: 10.1093/bioinformatics/bth315

Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Mishra, B., Gupta, D. K., Pfenninger, M., Hickler, T., Langer, E., Nam, B., et al. (2018). A reference genome of the European beech (*Fagus sylvatica* l.). *Gigascience* 7, 1–8. doi: 10.1093/gigascience/giy063

Mishra, B., Ulaszewski, B., Meger, J., Aury, J. M., Bodénès, C., Lesur-Kupin, I., et al. (2022). A chromosome-level genome assembly of the European beech (*Fagus sylvatica*) reveals anomalies for organelle DNA integration, repeat content and distribution of SNPs. *Front. Genet.* 12 2748. doi: 10.3389/fgene.2021.691058

Murat, F., Zhang, R., Guizard, S., Gavranovic, H., Flores, R., Steinbach, D., et al. (2015). Karyotype and gene order evolution from reconstructed extinct ancestors highlight contrasts in genome plasticity of modern rosid crops. *Genome Biol. Evol.* 7, 735–749. doi: 10.1093/gbe/evv014

Naish, M., Alonge, M., Wlodzimierz, P., Tock, A. J., Abramson, B. W., Schmücker, A., et al. (2021). The genetic and epigenetic landscape of the *Arabidopsis centromeres*. *Science* 374, eabi7489. doi: 10.1126/science.abi7489

Nawrocki, E. P., and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29, 2933–2935. doi: 10.1093/bioinformatics/btt509

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Oh, S. H., and Manos, P. S. (2008). Molecular phylogenetics and cupule evolution in fagaceae as inferred from nuclear *CRABS CLAW* sequences. *Taxon* 57, 434–451. doi: 10.2307/25066014

Ou, S., and Jiang, N. (2018). LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 176, 1410–1422. doi: 10.1104/pp.17.01310

Persson, S., Paredez, A., Carroll, A., Palsdottir, H., Doblin, M., Poindexter, P., et al. (2007). Genetic evidence for three unique components in primary cell-wall cellulose

synthase complexes in *Arabidopsis*. *Proc. Natl. Acad. Sci.* 104, 15566–15571. doi: 10.1073/pnas.0706592104

Petit, R. J., Carlson, J., Curtu, A. L., Loustau, M. L., Plomion, C., González-Rodríguez, A., et al. (2013). Fagaceae trees as models to integrate ecology, evolution and genomics. *New Phytol.* 197, 369–371. doi: 10.1111/nph.12089

Plomion, C., Aury, J. M., Amselem, J., Leroy, T., Murat, F., Duplessis, S., et al. (2018). Oak genome reveals facets of long lifespan. *Nat. Plants* 4, 440–452. doi: 10.1038/s41477-018-0172-3

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* 26, 1641–1650. doi: 10.1093/molbev/msp077

Price, A. L., Jones, N. C., and Pevzner, P. A. (2005). *De novo* identification of repeat families in large genomes. *Bioinformatics* 21, i351–i358. doi: 10.1093/bioinformatics/bti1018

Pryszcz, L. P., and Gabaldón, T. (2016). Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* 44, e113–e113. doi: 10.1093/nar/gkw294

Ramos, A. M., Usié, A., Barbosa, P., Barros, P. M., Capote, T., Chaves, I., et al. (2018). The draft genome sequence of cork oak. *Sci. Data* 5, 1–12. doi: 10.1038/sdata.2018.69

Rennenberg, H., Wildhagen, H., and Ehlting, B. (2010). Nitrogen nutrition of poplar trees. *Plant Biol.* 12, 275–291. doi: 10.1111/j.1438-8677.2009.00309.x

Richmond, T. A., and Somerville, C. R. (2000). The cellulose synthase superfamily. *Plant Physiol.* 124, 495–498. doi: 10.1104/pp.124.2.495

Seppey, M., Manni, M., and Zdobnov, E. M. (2019). BUSCO: assessing genome assembly and annotation completeness. *Gene prediction: Methods Protoc.* 1962, 227–245. doi: 10.1007/978-1-4939-9173-0_14

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* 11, e163962. doi: 10.1371/journal.pone.0163962

Shi, M. M., Michalski, S. G., Chen, X. Y., Chen, X. Y., and Durka, W. (2011). Isolation by elevation: genetic structure at neutral and putatively non-neutral loci in a dominant tree of subtropical forests, *Castanopsis eyrei*. *PLoS One* 6, e21302. doi: 10.1371/journal.pone.0021302

Slocum, R. D. (2005). Genes, enzymes and regulation of arginine biosynthesis in plants. *Plant Physiol. Biochem.* 43, 729–745. doi: 10.1016/j.plaphy.2005.06.007

Song, J. M., Xie, W. Z., Wang, S., Guo, Y. X., Koo, D. H., Kudrna, D., et al. (2021). Two gap-free reference genomes and a global view of the centromere architecture in rice. *Mol. Plant* 14, 1757–1767. doi: 10.1016/j.molp.2021.06.018

Sork, V. L., Cokus, S. J., Fitz-Gibbon, S. T., Zimin, A. V., Puiu, D., Garcia, J. A., et al. (2022). High-quality genome and methylomes illustrate features underlying evolutionary success of oaks. *Nat. Commun.* 13 2047. doi: 10.1038/s41467-022-29584-y

Sork, V. L., Fitz-Gibbon, S. T., Puiu, D., Crepeau, M., Gugger, P. F., Sherman, R., et al. (2016). First draft assembly and annotation of the genome of a california endemic oak *Quercus lobata* nee (Fagaceae). *G3: Genes Genomes Genet.* 6, 3485–3495. doi: 10.1534/g3.116.030411

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–W439. doi: 10.1093/nar/gkl200

Sun, Y., Guo, J., Zeng, X., Chen, R., Feng, Y., Chen, S., et al. (2022). Chromosome-scale genome assembly of *Castanopsis tibetana* provides a powerful comparative framework to study the evolution and adaptation of fagaceae trees. *Mol. Ecol. Resour.* 22, 1178–1189. doi: 10.1111/1755-0998.13539

Sun, Y., Hu, H., Huang, H., and Vargas-Mendoza, C. F. (2014). Chloroplast diversity and population differentiation of *Castanopsis fargesii* (Fagaceae): a dominant tree species in evergreen broad-leaved forest of subtropical China. *Tree Genet. Genomes* 10, 1531–1539. doi: 10.1007/s11295-014-0776-3

Sun, Y., Surget-Groba, Y., and Gao, S. (2016). Divergence maintained by climatic selection despite recurrent gene flow: a case study of *Castanopsis carlesii* (Fagaceae). *Mol. Ecol.* 25, 4580–4592. doi: 10.1111/mec.13764

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science* 320, 486–488. doi: 10.1126/science.1153917

Vogt, T. (2010). Phenylpropanoid biosynthesis. *Mol. Plant* 3, 2–20. doi: 10.1093/mp/ssp106

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. doi: 10.1093/bioinformatics/btx153

Wang, J., Hong, P., Qiao, Q., Zhu, D., Zhang, L., Lin, K., et al. (2022a). Chromosome-level genome assembly provides new insights into Japanese chestnut (*Castanea crenata*) genomes. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1049253

Wang, S., Liang, H., Wang, H., Li, L., Xu, Y., Liu, Y., et al. (2022b). The chromosome-scale genomes of *Dipterocarpus turbinatus* and *Hopea hainanensis* (Dipterocarpaceae) provide insights into fragrant oleoresin biosynthesis and hardwood formation. *Plant Biotechnol. J.* 20, 538–553. doi: 10.1111/pbi.13735

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49. doi: 10.1093/nar/gkr1293

Wang, J., Tian, S., Sun, X., Cheng, X., Duan, N., Tao, J., et al. (2020). Construction of pseudomolecules for the Chinese chestnut (*Castanea mollissima*) genome. *G3: Genes Genomes Genet.* 10, 3565–3574. doi: 10.1534/g3.120.401532

Watanabe, M., Kitaoka, S., Eguchi, N., Watanabe, Y., Satomura, T., Takagi, K., et al. (2014). Photosynthetic traits and growth of *Quercus mongolica* var. *crispula* sprouts attacked by powdery mildew under free-air CO_2 enrichment. *Eur. J. For. Res.* 133, 725–733. doi: 10.1007/s10342-013-0744-8

Wilf, P., Nixon, K. C., Gandolfo, M. A., and Cúneo, N. R. (2019). Eocene Fagaceae from Patagonia and gondwanan legacy in Asian rainforests. *Science* 364, eaaw5139. doi: 10.1126/science.aaw5139

Xu, Z., and Wang, H. (2007). LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

You, Y., Huang, X., Zhu, H., Liu, S., Liang, H., Wen, Y., et al. (2018). Positive interactions between *Pinus massoniana* and *Castanopsis hystrix* species in the uneven-aged mixed plantations can produce more ecosystem carbon in subtropical China. *For. Ecol. Managemen* 410, 193–200. doi: 10.1016/j.foreco.2017.08.025

Yuan, S., Shi, Y., Zhou, B. F., Liang, Y. Y., Chen, X. Y., An, Q. Q., et al. (2023). Genomic vulnerability to climate change in *Quercus acutissima*, a dominant tree species in East Asian deciduous forests. *Mol. Ecol* 00, 1–17. doi: 10.1111/mec.16843

Zdobnov, E. M., and Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847–848. doi: 10.1093/bioinformatics/17.9.847

Zhang, P., He, Y., Feng, Y., De La Torre, R., Jia, H., Tang, J., et al. (2019a). An analysis of potential investment returns of planted forests in south China. *New Forests* 50, 943–968. doi: 10.1007/s11056-019-09708-x

Zhang, R. G., Li, G. Y., Wang, X. L., Dainat, J., Wang, Z. X., Ou, S., et al. (2022). TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. *Horticulture Res.* 9, uhac017. doi: 10.1093/hr/uhac017

Zhang, Z., Xiao, J., Wu, J., Zhang, H., Liu, G., Wang, X., et al. (2012). ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* 419, 779–781. doi: 10.1016/j.bbrc.2012.02.101

Zhang, X., Zhang, S., Zhao, Q., Ming, R., and Tang, H. (2019b). Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-c data. *Nat. Plants* 5, 833–845. doi: 10.1038/s41477-019-0487-8

Zhao, Z., Liu, Y., Tian, Z. W., Jia, H. Y., Zhao, R. R., and An, N. (2020). Dynamics of seed rain, soil seed bank and seedling regeneration of *Castanopsis hystrix*. *Sci. Silvae Sin.* 56, 37–49. doi: 10.11707/j.1001-7488.20200505

Zhou, X., Liu, N., Jiang, X., Qin, Z., Farooq, T. H., Cao, F., et al. (2022c). A chromosome-scale genome assembly of *Quercus gilva*: Insights into the evolution of *Quercus* section *Cyclobalanopsis* (Fagaceae). *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1012277

Zhou, B. F., Shi, Y., Chen, X. Y., Yuan, S., Liang, Y. Y., and Wang, B. (2022a). Linked selection, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence in *Quercus dentata*. *J. Systematics Evol.* 60, 1344–1357. doi: 10.1111/jse.12817

Zhou, B. F., Yuan, S., Crowl, A. A., Liang, Y. Y., Shi, Y., Chen, X. Y., et al. (2022b). Phylogenomic analyses highlight innovation and introgression in the continental radiations of fagaceae across the northern hemisphere. *Nat. Commun.* 13 1320. doi: 10.1038/s41467-022-28917-1

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 30 May 2023
DOI 10.3389/fpls.2023.1161693

OPEN ACCESS

EDITED BY
Kai-Hua Jia,
Shandong Academy of Agricultural Sciences, China

REVIEWED BY
Yaqiong Wu,
Chinese Academy of Sciences, China
Zihan Zhang,
Chinese Academy of Forestry, China

*CORRESPONDENCE
Li-min Sun
sunlimin06@163.com

†These authors have contributed equally to this work

RECEIVED 08 February 2023
ACCEPTED 25 April 2023
PUBLISHED 30 May 2023

CITATION
Cao Z-y, Su L-n, Zhang Q, Zhang X-y, Kang X-j, Li X-h and Sun L-m (2023) The development and transcriptome regulation of the secondary trunk of *Ginkgo biloba* L.. *Front. Plant Sci.* 14:1161693. doi: 10.3389/fpls.2023.1161693

The development and transcriptome regulation of the secondary trunk of *Ginkgo biloba* L.

Zhong-yun Cao[1†], Li-ning Su[1†], Qian Zhang[2], Xin-yue Zhang[1], Xiao-jing Kang[1], Xin-hui Li[1] and Li-min Sun[1*]

[1]State Forestry and Grassland Administration Key Laboratory of Silviculture in Downstream Areas of the Yellow River, Forestry College of Shandong Agricultural University, Tai'an, Shandong, China, [2]Shandong Academy of Forestry Sciences, Jinan, Shandong, China

Secondary trunk *Ginkgo biloba* is one of the specific germplasms of *G. biloba*. In this study, paraffin sectioning, high-performance liquid chromatography and transcriptome sequencing technology were used to study the development of the secondary trunk of *G. biloba* from the morphological, physiological and molecular levels. The results showed that the secondary trunk of *G. biloba* originated from the latent buds in the stem cortex at the junction of the root and stem of the main trunk. The development process of secondary trunk was divided into 4 periods: the dormancy period of the secondary trunk buds, the differentiation period, the formation period of transport tissue, and the budding period. Transcriptome sequencing was performed by comparing the germination period and elongation growth period of the secondary trunk with the normal parts of the same period where no secondary trunks occurred. Differential genes involved in phytohormone signal transduction, phenylpropane biosynthesis, phenylalanine metabolism, glycolysis and other pathways can regulate not only the inhibition of early dormant buds but also the later development of the secondary trunk. Genes related to IAA synthesis are upregulated and indole-3-acetic acid content is increased, leading to the up-regulated expression of IAA intracellular vector genes. The IAA response gene (SAUR) receives and responds to IAA signals to promote the development of the secondary trunk. Through the enrichment of differential genes and functional annotations, a key regulatory pathway map for the occurrence of the secondary trunk of *G. biloba* was sorted out.

KEYWORDS

***Ginkgo biloba*, secondary trunk, anatomy, endogenous hormones, transcriptome sequencing**

1 Introduction

Ginkgo biloba L. is the only remaining species of the Ginkgoaceae in China during the Quaternary glacial period with high adaptability, longevity and ornamental value (Seward, 1938; Cao, 2007). *G. biloba* is different from other gymnosperm species in that the phenomenon of 'arising branches from the base of a stem' (secondary trunk) is common from young trees to thousand-year-old trees. Del Tredici found in his survey of the West Tianmu Mountain Nature Reserve that 40% of the *G. biloba* in the area could produce sprouts, most of which were connected to a callus-like tumor (basal tree tumor) at the base of the stem. Therefore, it was believed that *G. biloba* could be regenerated by secondary trunk from the basal tree tumors, but the origin of the above-ground sprouts was not mentioned (Tredici, 1992a; Tredici, 1992b).

Xing Shiyan named the 'branches arising from the base of a stem' of *G. biloba* as the 'secondary trunk' for the first time and showed that the secondary trunk of *G. biloba* originated from the latent buds of the stem at the junction of root and stem. And the secondary trunk was essentially different from the root tillers of other tree species. The root system of return-young *G. biloba* does not produce rootstocks like the paulownia and other species. The secondary trunks of *G. biloba* usually grow upright around the main stem, with thick stems, small angles to the main stem, large, thick, multi-lobed leaves, obvious 'return-young' characteristics, growing faster than the main stem (Xing, 1996; Xing, 2013). The mechanism of the secondary trunk is related to the type of stem (branch) differentiation system, and the secondary trunk of *G. biloba* was part of the normal development of individuals in the natural state. With the increase of age, the number of secondary trunks increases greatly after the senescence of the tree top or destruction of the top buds of the secondary trunk, and the secondary trunk can be regenerated on the secondary trunk. As the number of secondary trunks increases, the base of the secondary trunk can form root discs. The secondary trunk retains the characteristics of the main stem and has a low basal rooting rate (Xing, 1996). The current research on the secondary trunk of *G. biloba* has mainly focused on the growth characteristics, distribution, regeneration, and utilization of the secondary trunk, but little has been reported on the origin of the secondary trunk of *G. biloba* at the anatomical level and the intrinsic molecular mechanism affecting the development of the secondary trunk.

In this study, the paraffin sectioning method was used to observe the initiation of the secondary trunk of *G. biloba*. By high-performance liquid chromatography, the endogenous hormones were measured in two different developmental stages of the secondary trunk of *G. biloba* and lateral branches to understand the influence of endogenous hormones on the development of the secondary trunk. At the same time, transcriptome sequencing and bioinformatics analysis were used to screen the differentially expressed genes affecting the formation and development of the secondary trunk and to speculate possible transcriptional regulation relationships. Through the research on the development of the secondary trunk, the occurrence mechanism of the secondary trunk can be understood and it can lay a theoretical foundation for revealing the mystery of the longevity of ancient ginkgo trees.

2 Materials and methods

2.1 Experimental material

The experimental materials for anatomical observation and endogenous hormone determination of the secondary trunk of *G. biloba* were collected from 4 years of seedling *G. biloba* and grafted *G. biloba* in the Forestry Experimental Station (N36°10', E117°10') of the South Campus of Shandong Agricultural University. The seedlings grow robustly and free from pests and diseases. The region is located in the southeast of Tai'an City, Shandong Province. It belongs to a warm temperate zone and semi-humid continental monsoon climate.

The experimental materials for the transcriptome sequencing of the secondary trunks were collected from the 4-year-old *G. biloba* seedlings from the 75# family of Shandong Forestry Germplasm Resource Center. The seedlings grew robustly and free from pests and diseases (Figure 1). The region is located in Zhangqiu District, Jinan City, Shandong Province. It belongs to a warm temperate zone and semi-humid continental monsoon climate.

2.2 Experimental method

2.2.1 Anatomical observation on the origin of the secondary trunk of *G. biloba*

For the 4-year-old seedling *G. biloba*, to observe the morphology of dormant buds of the secondary trunk and the development process of the secondary trunk after breaking the dormancy, the following experiment method was used: the *G. biloba* seedlings with consistent growth were divided into 8 groups of 3 replicates each, with 10 seedlings in each replicate. At the end of February, the first sampling was conducted and 5 seedlings were collected from each replicate from the first group to observe the dormant bud status by paraffin section. In early March, the remaining 7 groups of *G. biloba* were treated with stumping before the leaves were unfolded to break the dormant state of buds and promote the occurrence of the secondary trunk of *G. biloba* After cutting, a set of samples were taken at weekly intervals and brought back to the laboratory to make paraffin sections, and stained with safranin and green counterstaining. The paraffin sections were observed and photographed under a Nikon E200 microscope.

2.2.2 Endogenous hormone determination

The secondary trunks and the lowermost lateral branches of four-year-old grafted *G. biloba* were collected at two different developmental stages (germination and elongation), and three replicates of secondary trunks and lateral branches from each developmental stage were set up for the endogenous hormone determination by high-performance liquid chromatography. The

FIGURE 1
Experimental samples of transcriptome. TF, Secondary trunk at stage of germination; TS, Secondary trunk at stage of elongation; CF、CS, Normal tissue around secondary trunk.

specific method referred to Zhang Ning's extraction and determination method (Zhang, 2019). The hormones determined: zeatin (ZT), auxin (IAA), abscisic acid (ABA), and gibberellin (GA_3).

2.2.3 Transcriptome sequencing and analysis

The 4-year-old *G. biloba* seedlings from the no.75 family of Shandong Forest Germplasm Resource Center were selected. In April 2018, the secondary trunks of the four-year-old *G. biloba* seedlings of the same family at two stages of development (germination stage and elongation growth stage) were collected, and the normal tissues near the secondary trunk were collected as controls. Three replicates were set up for each period and control samples (Table 1). According to the RNA extraction method of Liu Xiaojing (Liu et al, 2018), the total RNA of 12 *G. biloba* samples was extracted to ensure the use of qualified samples for transcriptome sequencing.

After the samples were qualified, 3ug of RNA was taken from each sample as the starting material for library construction. The qualified total RNA samples were enriched into mRNA. The obtained mRNA was broken into short fragments by adding a fragmentation buffer. The fragmented mRNA was then used as a template to synthesize the first strand of cDNA with a six-base random primer. The buffer, DNA polymerase I, RNase H and dNTPs were added to synthesize the second strand of cDNA. The double-stranded cDNA was purified by QiaQuick PCR kit and eluted with EB buffer. The eluted and purified double-stranded cDNA was followed by terminal repair, base A addition and sequencing joint. Finally, different size fragments were selected for PCR amplification to complete the preparation of the library. After the library was constructed, the preliminary quantification was done with Qubit3.0 and the library was diluted to 1ng/ul. Agilent 2100 was used to detect the insert size of the library, and the qRT-PCR method was used to accurately quantify the effective concentration of the library. After the library inspection was qualified, sequencing was carried out by illumine platform.

Raw Data was filtered by removing low-quality sequences and joint pollution to obtain high-quality Clean Data. Mapped Data was get by comparing Clean Data with *G. biloba* reference genome (http://gigadb.org/dataset/100613). DESeq2 was used to analyze differentially expressed genes between the treatment and control groups. Genes with $q<0.05$ and $|\log2 \text{ Fold change}|\geq1$ were selected as significantly differentially expressed genes, and each comparison group was screened to obtain the number of up-regulated and down-regulated genes. All the different Genes were analyzed by GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes). The Sequence data are available in the NCBI Sequence Read Archive (SRA): SRR23730290, SRR23730291, SRR23730292, SRR23730293, SRR23730294, SRR23730288, SRR23730289, SRR23730295, SRR23730296, SRR23730297, SRR23730298, SRR23730299.

The RNA of secondary trunks at two different development stages and normal tissue without the secondary trunk of the same plant at the same location were extracted for reverse transcription, and the expression of differentially expressed genes in the transcriptome was verified. Fifteen randomly selected genes

TABLE 1 Experimental samples of transcriptome.

sample	repetition	germination stage	elongation growth stage
the secondary trunk	repetition 1	TF1	TS1
	repetition 2	TF2	TS2
	repetition 3	TF3	TS3
contrast	repetition 1	CF1	CS1
	repetition 2	CF2	CS2
	repetition 3	CF3	CS3

10.3389/fpls.2023.1161693

related to the development of the secondary trunk of *G. biloba* were quantified by real-time fluorescence.

3 Result

3.1 Anatomic characteristics of the origin of the secondary trunk

Anatomical studies on the secondary trunk of *G. biloba* at different developmental periods of showed that the development of the secondary trunk can be divided into four periods:

Dormancy period of the buds of the secondary trunk: In the natural state, that was, before the cutting treatment, a longitudinal cut at the base of the *G. biloba* stem showed a group of active parenchyma cells at the cortex, which had a small volume, large nucleus, deep staining, dense arrangement, and relatively vigorous division ability, and formed a densely distributed region. This part of the cells was the germinal cells of the buds of the secondary trunk, namely the dormant buds of the secondary trunk (Figure 2A). This part of cells differentiated into a primordial cell group composed of 1-2 layers of cells at the top through vertical and horizontal divisions, and a central blast cell area derived from the primordial cell group was below. The rib-shaped meristem area was under the central blast area. The meristem region was a cluster of closely arranged cells in a sub-circular or oval shape, the cells of the cluster were significantly smaller than surrounding parenchyma cells, and the cells were not uniform in size. One side of the cells were smaller, closely arranged, with a large nucleus, dense cytoplasm and deep staining, and the other side of the cells were larger and sub-circular. The cell cluster showed a strong capacity to divide (Figure 2B).

Differentiation period: Series of sections after cutting treatment showed that the dormancy stage of dormant buds was broken and the cell mass began to differentiate further, forming a clear gap above the primitive cell mass, which gradually enlarged. The apical cells in the meristematic region were closely arranged, with dense cytoplasm, deep staining, and vigorous division, forming a slight protrusion in the middle, which was the apical meristem. At the same time, the peripheral meristem areas on both sides of the apical meristem also showed strong splitting ability, but the differentiation of leaf primordium was not obvious (Figures 2C, D).

The formation period of the transport tissue: With the continuous development of the adventitious buds and the enlargement of the apical gap, the apical meristem formed more obvious protrusions. The surrounding meristem maintained a high frequency of cell division during the development of adventitious buds and formed protrusions after periclinal division, resulting in two leaf primordium. At this time, obvious vascular tissue could be seen at the bottom of adventitious buds, indicating that adventitious buds began to connect with the vascular bundle of the stem, forming a continuous vascular system to ensure that the trunk provided sufficient nutrients for the development of the buds of the secondary trunk (Figures 2E, F). The leaf primordium continued to grow to form young leaves and the secretory cavity began to exist in the young leaves. The surrounding meristem developed further to create a new leaf primordium on the inner side of the young leaf, at which point a continuous vascular system had been formed (Figure 2G).

Germination period: Adventitious buds continued to differentiate and produced new young leaves, with a certain number of secretory cavities in the young leaves, bud scales and buds. Since the tangential division speed of the rib-like meristem was greater than that of the periclinal division, the cells proliferated continuously in the longitudinal direction, so that the adventitious buds broke through the bark of the main trunk and developed into

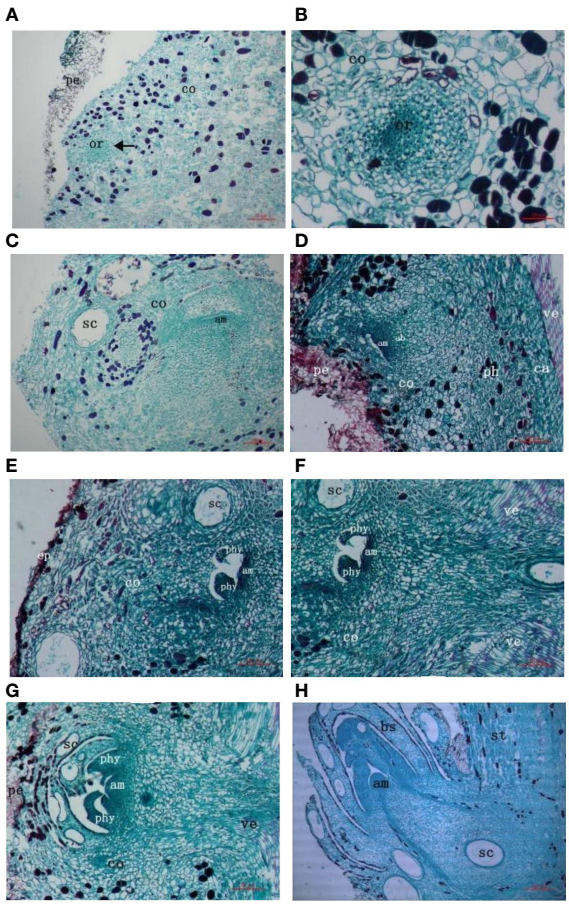


FIGURE 2
Anatomy of the origin of *G. biloba* secondary trunk. **(A)** The originating cells of secondary trunk form at the cortex. Arrows indicate originating cells. **(B)** The originating cells of secondary trunk differentiate into cell mass. **(C, D)** The cell mass further divides and differentiates to form meristematic regions with cytological division of shoot end. **(E, F)** The meristem region develops into leaf primordium and adventitious bud begins to connect with the vascular system of stem. **(G)** The adventitious buds continue to develop and produce new leaf primordial. The adventitious buds are connected to the vascular system of the stem. **(H)** The buds of secondary trunk form and break bark to grow. Ruler is 10μm. pe, periderm; or, originating cells of secondary trunk, co, cortex; sc, secretory cavity; am, apical meristem; ab, adventitious bud; ph, phloem; ca, cambium; ve, vessel; ep, epidermis; phy, phyllopodium; st, stem; bs, bud scale.

the secondary trunk under suitable light, temperature and moisture conditions (Figure 2H).

3.2 Hormone changes in the growth process of the secondary trunk and lateral branches

During the germination and elongation growth period, the content of each hormone in the secondary trunk was significantly higher than that in the lateral branches. During the germination period, the content of IAA in the secondary trunk (1.52μg·g^{-1}) and the lateral branches (1.47μg·g^{-1}) remained at a high level, and the content of IAA in the secondary trunk was significantly higher than that in the lateral branches ($t < 0.05$) at the same development stage. In addition, the content of ZA and GA_3 in the secondary trunk (0.98μg·g^{-1}、 1.26μg·g^{-1}) was significantly different from that in the lateral branch (0.66μg·g^{-1}、 0.63μg·g^{-1}) ($t < 0.01$). And the content of the two hormones in the secondary trunk was also higher than that in the lateral branch.

The content of various hormones in the secondary trunk also showed significant differences at different developmental stages ($t<0.05$). At the elongation stage (1.64μg·g^{-1}、 1.58μg·g^{-1}), the content of IAA and ZA in the secondary trunk increased significantly ($t<0.05$) compared with that at the germination stage (1.52μg·g^{-1}、 0.98μg·g^{-1}), while the content of GA_3 decreased significantly ($t < 0.01$). The same expression trend was also observed in the lateral branches.

Among the common endogenous hormones in plants, ABA is a growing-inhibiting hormone, while IAA, ZA and GA_3 are growing-promoting hormones. IAA and ZA synergistically affect the formation and development of buds, and ZA is the determinant of bud germination and growth. Therefore, IAA/ZA and (IAA+ZA +GA_3)/ABA are the two more referential ratios. The results showed that the ratios of IAA+ZA+GA_3/ABA were greater than 1 for both the secondary trunk and lateral branches at the germination stage, and the ratio of IAA+ZA+GA_3/ABA in the secondary trunk was significantly greater than that of the lateral branch ($t < 0.01$), and the ratio of IAA/ZA was significantly smaller than that of the lateral branch ($t < 0.01$). During the elongation growth period, the ratios of IAA+ZA+GA_3/ABA were greater than 1 for both the secondary trunk and lateral branches, and the ratio of IAA+ZA+GA_3/ABA in the secondary trunk was greater than that of the lateral branches, while the ratio of IAA/ZA was smaller than that of the lateral branches. The ratio of (IAA+ZA+GA_3)/ABA decreased during the elongation period compared to the germination period and the growth trend slowed down, while the ratio of IAA+ZA+GA_3/ABA increased slightly in the lateral branches (Table 2; Figure 3).

3.3 Transcriptome sequencing

3.3.1 Quality evaluation of sequencing libraries and comparison results with reference genomes

In this study, transcriptome sequencing was performed on 12 *G. biloba* samples (including secondary trunk and control of two periods, three biological replicates respectively). The results showed that the Clean Reads Rate in each sample was above 96%, and after filtering, the Clean Q30 Bases Rate was all above 92.99%. The percentage of gene sequences of each sample aligned to reference genome sequence was above 88%. The results of gene sequence alignment showed that all samples met the sequencing requirements, providing quality assurance for subsequent bioinformatics analysis (Table 3; Figure S1). The clustering results of differentially expressed genes showed that the differentially expressed genes could be roughly divided into four categories, namely, the secondary trunk (TF) and control (CF, the part without the secondary trunk) at the germination stage as well as the secondary trunk (TS) and control (CS, the part without the secondary trunk) at the elongation growth stage. All repetition groups were clustered into the same category, with good repeatability within the group (Figure S2A).

3.3.2 Differential expression analysis

The statistical results of differentially expressed genes showed that a total of 74 significantly different genes were screened out in the TF vs. CF comparison group (3 up-regulated and 71 down-regulated expressions). A total of 104 significantly different genes (76 up-regulated and 28 down-regulated expressions) were screened out in the TF vs. TS comparison group. A total of 2,151 significantly different genes (1368 up-regulated and 783 down-regulated expressions) were screened out in the TS vs. CS comparison group (Table 4; Figures S2B, C).

3.3.3 GO function analysis of differentially expressed genes

All DEGs were annotated in the KEGG database, NR database, Swissprot database, KOG database, GO database, Pfam database,

TABLE 2 Hormonal changes at different developmental stages in secondary trunk and lateral branch of *G. biloba*.

Plant organs	IAA μg·g^{-1}	ZA μg·g^{-1}	GA_3 μg·g^{-1}	ABA μg·g^{-1}	IAA/ZA	(IAA+ZA +GA_3)/ABA
F1	1.52 ± 0.0088	0.98 ± 0.0000	1.26 ± 0.0176	1.61 ± 0.0058	1.54 ± 0.0033	2.33 ± 0.0067
Z1	1.47 ± 0.0120	0.66 ± 0.0033	0.63 ± 0.0153	1.55 ± 0.0067	2.23 ± 0.0088	1.78 ± 0.0100
F2	1.64 ± 0.0232	1.56 ± 0.0451	0.59 ± 0.0384	1.76 ± 0.0265	1.05 ± 0.0260	2.15 ± 0.0333
Z2	1.52 ± 0.0548	1.14 ± 0.0867	0.39 ± 0.0296	1.63 ± 0.0273	1.35 ± 0.1100	1.88 ± 0.0960

F1 and Z1 represent the secondary trunk and lateral branch of *G. biloba* during the germination stage. F2 and Z2 represent the secondary trunk and lateral branch of *G. biloba* during the elongation stage.

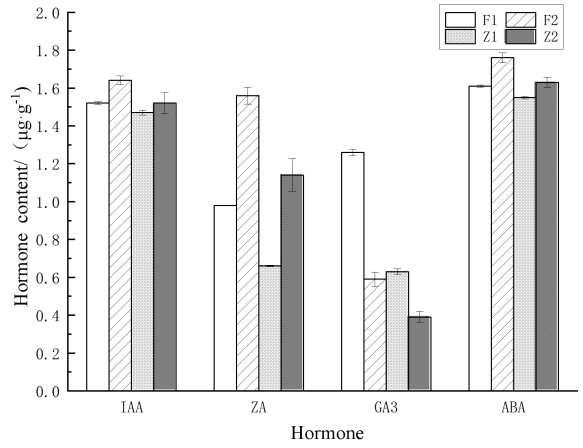


FIGURE 3
Hormonal changes at different developmental stages in the secondary trunk and the lateral branch of *G. biloba*.

etc. The functional enrichment analysis of the annotated DEGs in the GO database could reflect the cell state of the bud of the secondary trunk and the control group at different stages. GO analysis was performed on the differential genes produced at two stages of secondary trunk development compared with the control group. It was found that during the process of secondary trunk development, the differential genes in the two stages were involved in the metabolic process, cell process, single organism process, cell part, membrane part, catalytic activity and protein binding in biological processes. It was speculated that the above aspects were closely related to the germination and growth of the buds of the secondary trunk (Figures S3–S5).

3.3.4 KEGG pathway analysis

All differentially expressed genes were enriched into 126 KEGG pathways. In all comparison groups, pathways enriched to the top 15 in the number of differential genes were shown in Table 5. 42 differentially expressed genes were enriched in plant hormone signal transduction pathways (32 up-regulated and 10 down-regulated genes) (Figure S6; Table 5).

TABLE 3 The quality of sequencing library and comparison of cDNA library of sample and reference genome of *G. biloba*.

Sample	Clean Reads Number	Clean Reads Rate(%)	Clean Q30 Bases Rate(%)	Mapped Reads	Mapping Rate
CF1	41,474,404	97.48	93.36	37, 628, 351	0.9073
CF2	43,072,218	96.97	93.4	38, 660, 133	0.8976
CF3	47,253,398	97.61	93.74	42, 477, 701	0.8989
CS1	45, 053, 288	97.74	93.23	40, 667, 095	0.9026
CS2	47,512,890	97.66	94.05	42, 994, 997	0.9049
CS3	44,086,324	97.31	93.81	39, 409, 256	0.8939
TF1	42, 085, 858	97.48	94.33	38, 460, 000	0.9138
TF2	44, 025, 658	97.65	93.46	39, 373, 354	0.8943
TF3	38, 561, 556	97.64	93.67	34, 432, 465	0.8929
TS1	44, 387, 138	97.58	93.13	40, 114, 174	0.9037
TS2	47, 149, 178	97.83	92.99	41, 689, 486	0.8842
TS3	47, 419, 422	97.78	93.99	42, 503, 427	0.8963

TF, The repeated groups of secondary trunk at the germination stage; TS, The repeated groups of secondary trunk at the elongation stage; CF, The repeated groups of control samples around secondary trunk during germination; CS, The repeated groups of the control samples around secondary trunk during the elongation period. (1) Clean Reads Number: The total number of filtered high-quality sequences. (2) Clean Reads Rate (%): The percentage of the number high-quality sequences obtained after filtering to the number of raw sequences. (3) Clean Q30 Bases Rate (%): After filtration, the percentage of bases whose mass value is greater than 30 in the total sequence. (4) Mapped Reads: The number of sequences of each sample aligned to reference genomic. (5) Mapping Rate: The percentage of each sample sequence aligned to the reference genome sequence.

TABLE 4 Number of differentially expressed genes in each comparison group of *G. biloba*.

Comparison group	Up	Down	Total
TF vs. CF	3	71	74
TF vs. TS	76	28	104
TS vs. CS	1, 368	783	2, 151

(1) Up, Up-regulated genes; (2) Down, Down-regulated genes; (3) Total, Sum of differential genes.

3.3.5 The regulation of the germination and development of the secondary trunk of *G. biloba*

Based on the results of endogenous hormone determination and transcriptomic data analysis, the molecular regulatory pathways related to the germination and development of the secondary trunk of *G. biloba* were plotted. The results showed that the germination and development of secondary trunk were closely related to plant

TABLE 5 Pathways enriched to the top 15 in the number of differential genes.

Pathway name	Ko	Number of up-regular DEGs	Number of down-regular DEGs	Total number
Plant hormone signal transduction	ko04075	32	10	42
Cutin, suberine and wax biosynthesis	ko00073	40	1	41
Phenylpropanoid biosynthesis	ko00940	35	4	39
Phenylalanine metabolism	ko00360	30	4	34
Plant-pathogen interaction	ko04626	23	8	31
Biosynthesis of amino acids	ko01230	22	4	26
Carbon metabolism	ko01200	22	3	25
Amino sugar and nucleotide sugar metabolism	ko00520	16	8	24
Flavonoid biosynthesis	ko00941	18	4	22
Glycerophospholipid metabolism	ko00564	18	1	19
Purine metabolism	ko00230	16	3	19
Glycolysis/Gluconeogenesis	ko00010	17	2	19
Pyruvate metabolism	ko00620	15	3	18
Starch and sucrose metabolism	ko00500	13	3	16
Glycerolipid metabolism	ko00561	15	0	15

hormone signal transduction, phenylpropane biosynthesis, phenylalanine metabolism, glycolysis and other metabolic pathways, and there were complex mutual regulatory relationships, among which the synthesis and regulation of IAA and GA_3 played an important role in the development of the secondary trunk (Figures 4A, B).

3.3.6 Verification of differential gene expression

In this study, 15 differentially expressed genes were randomly selected for qPCR verification. The results showed that the expression trend of these 15 differentially expressed genes was generally consistent with the transcriptome sequencing data. *Gb36988*, *Gb33056*, *Gb10062*, *Gb29808*, *Gb04557*, *Gb04566*, *Gb24946*, *Gb14994*, *Gb25911*, *Gb01010*,

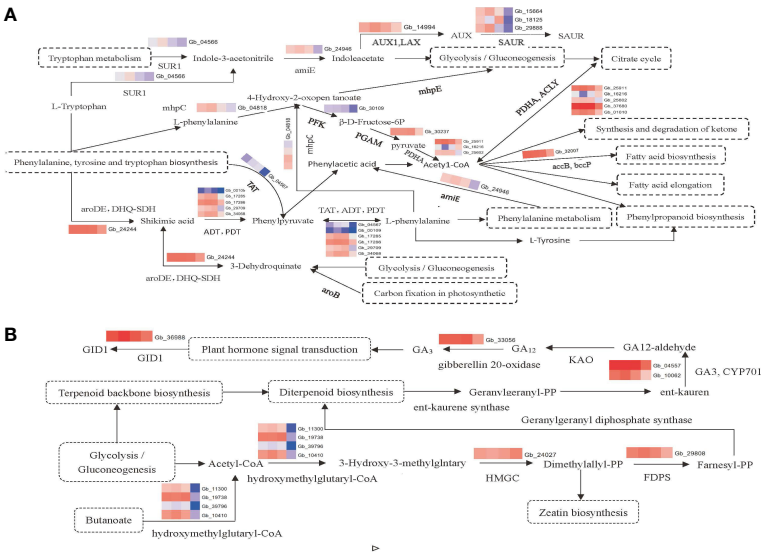


FIGURE 4
(A) Pathway map of differential gene expression related synthesis and regulation of IAA. **(B)** Pathway map of differential gene expression related synthesis and regulation of gibberellin.

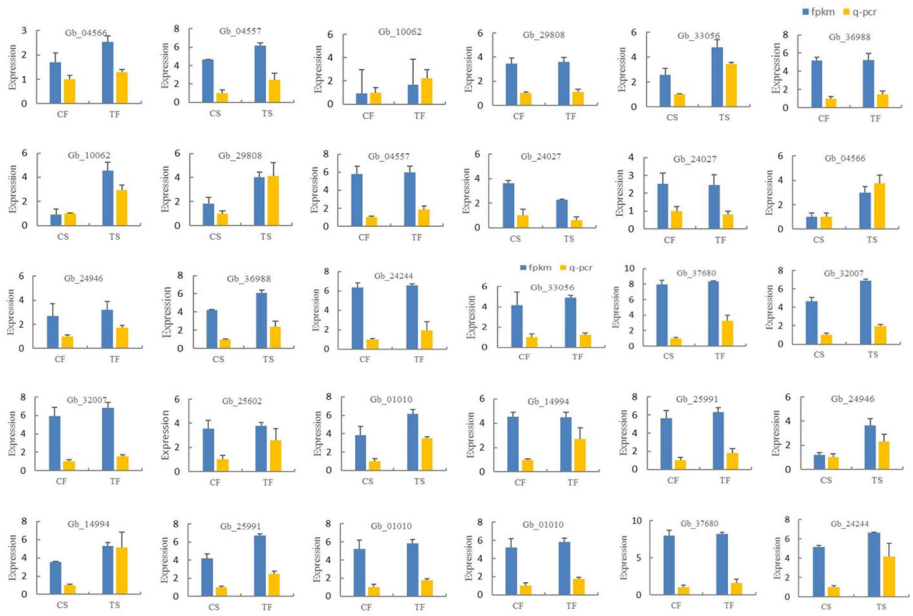


FIGURE 5
Expression verification of differentially expressed genes in different samples of *G. biloba*.

Gb25602, *Gb32007*, *Gb37680*, and *Gb24244* were up-regulated during the two development stages of the secondary trunk and down-regulated in the control group. *Gb24027* was down-regulated in the two developmental stages of the secondary trunk and up-regulated in the control group. The results showed that the transcriptome sequencing data was accurate (Figure 5).

4 Discussion

The origin of the secondary trunk of *G. biloba* has been controversial (Fujii, 1985; Tredici, 1992b). Some studies suggested that the secondary trunk of *G. biloba* was produced by the chichi which was another special organ of *G. biloba*. Latent adventitious buds at the apex of rooted chichi could germinate into upgrowing branches under certain conditions and then developed into the secondary trunk (Xing, 1996; Fu, 2014). The chichi could produce roots and branches after it touched the ground (Fujii, 1985; Tredici, 1992a). It was also believed that the secondary trunk of *G. biloba* originated from the latent buds in the cortex of the stem at the junction of the root and stem. The secondary trunk was 'shoots from the stem' rather than shoots from the roots, part of the ontogenetic development of *G. biloba*, but no clear anatomical evidence has been found (Xing, 1996). This study confirmed the result and, in conjunction with previous studies, suggested that the origin of the secondary trunk should be divided into two categories. One type is produced by adventitious buds in the chichi, which can produce adventitious roots to provide nutrients for the secondary trunk, with a high germination rate. The other type is directly developed from the latent buds that exist in the cortex of the stem at the junction of the root and stem, which relies on the roots of the main stem to transport nutrients for development. The secondary trunk itself can not produce roots, retaining the characteristics of the main stem. The germination rate of the secondary trunk in the natural state is low and a large amount of germination does not occur until external stimulation. The secondary trunk of *G. biloba* may belong to autologous inhibitory dormancy, which allows plants to devote resources to controlling plant structure and reproductive growth, and allows regeneration when shoots are damaged (Horvath et al., 2003). The occurrence of the secondary trunk may be related to the age of *G. biloba*. With the increase of age, the secondary trunk will increase significantly after the senescence of the tops of the tree (Xing, 1996; Men et al., 2021). The secondary trunk is a reproductive strategy of *G. biloba* to sustain a life system when the tree reaches its growth limit or to resist bad environments and is the key to the preservation of *G. biloba* (Xiang et al., 2000; Lin and Zhang, 2004).

The secondary trunks are produced by breaking the dormancy of the dormant buds within the cortex at the junction of the root and stem of *G. biloba*, and are "branches from the stem", but the secondary trunks are different from normal lateral branches in many aspects. The secondary trunk has obvious characteristics of "return to juvenile", with an upright shape and a significantly smaller angle to the main stem than that of the lateral branches. Usually, the growth rate of the secondary trunk is significantly higher than that of the main stem, which can eliminate the position effect (Xing and Miao, 1996). This study found that the development of the secondary trunk was closely related to the regulation of endogenous hormones. At the germination stage, the content of each hormone in the secondary trunk was higher than that in the lateral branches, and the content of ZA and GA_3 was significantly higher than that in the lateral branches. ZA could directly promote the growth of the lateral buds and played an important role in the germination stage of the secondary trunk (Lin and Zhang, 2004; Müller and Leyser, 2011).

GA_3 might play a negative regulatory role in regulating the growth of lateral buds in *Arabidopsis* (Silverstone et al., 1997), rice (Qi et al., 2011) and hybrid aspen (Mauriat et al., 2011). However, for perennial *G. biloba*, the content of GA_3 was significantly increased at the early developmental stage of the secondary trunk, and with the elongation growth, the content of IAA and ZA increased significantly and the GA_3 content decreased significantly but was still higher than that of normal lateral branches. This is because the secondary trunk buds are in a dormancy state at the early stage of germination and the higher content of GA_3 is conducive to breaking the dormancy and promoting germination. When the secondary trunk began to grow high, the GA_3 gradually decreased and the content of each hormone was still higher than that of the lateral branches, indicating that the growth rate of the secondary trunk is higher than that of the lateral branches. ABA is an inhibitor of dormancy, and GA_3 can counteract the inhibitory effect of ABA (Faust et al, 1997). GA_3 can also promote the growth of lateral buds of papaya and mulberry (Ni, 2015). It indicates that GA_3 is not unique to *G. biloba* in promoting the development of secondary trunk buds. Although GA_3 has a negative regulatory effect on lateral buds in a few plants, GA_3 promotes the growth of lateral buds for many woody plants (Ni, 2015).

The development of buds is usually closely related to the ratio of IAA to ZA in the primary tissue (Beveridge et al., 2003). At the germination stage, the (IAA+ZA+GA_3)/ABA value of the secondary trunk and lateral branches were greater than 1, and the value of the secondary trunk was significantly greater than that of the lateral branches. It indicates that they are all growing and the growth trend of the secondary trunk is more obvious than that of lateral branches. The smaller IAA/ZA value is more favorable to the germination of the secondary trunk. In the elongation growth period, the (IAA+ZA+GA_3)/ABA value of the secondary trunk decreased compared with the germination period and the growth trend slowed down, while the ratio of lateral branches increased slightly, which was due to the difference in the time between the secondary trunk and lateral branches entering the rapid growth period. The secondary trunk responds quickly to internal and external environmental signals, while the lateral branches respond relatively slowly.

Through transcriptome analysis, the GO enrichment analysis of the screened differentially expressed genes revealed that the differentially expressed genes related to the secondary trunk development were mainly enriched in the metabolic process, cell process, cell part, membrane part, catalytic activity and protein binding in the biological processes. It was generally consistent with the research results of Yang Lili about the release of natural dormancy of grape winter buds (Yang et al., 2020).

The secondary trunk is a unique organ of *G. biloba*. The dormancy of secondary trunk buds is controlled by the environment and genes. Combined with the endogenous hormone determination and sequencing analysis, it was found that endogenous hormones may be an important internal cause for the development of the secondary trunk. The development of the secondary trunk is closely related to several metabolic pathways and has complex interrelationships, especially with the synthesis and metabolism of hormones. The genes related to phenylalanine metabolism and phenylpropane biosynthesis were significantly enriched at the early developmental stage of adventitious bud in *Arabidopsis*, and phenylalanine significantly inhibited the development of adventitious bud in *Arabidopsis* (Wang et al., 2013).

However, these two pathways were also significantly enriched in this study. Several genes related to the biosynthetic pathways of phenylalanine, tyrosine and tryptophan were enriched in pathways related to the synthesis and regulation of IAA. For example, genes related to L-phenylalanine synthesis, such as *Gb24244*(*aroDE*, *DHQ-SDH*), *Gb00109*(*ADT*, *PDT*), *Gb17285*(*ADT*, *PDT*), *Gb17286*(*ADT*, *PDT*), *Gb29709*(*ADT*, *PDT*), *Gb34068*(*ADT*, *PDT*), *Gb04567*(*TAT*), were up-regulated in the secondary trunk group. The synthesis of shikimic acid could be promoted by *DHQ-SDH*, and then L-phenylalanine could be synthesized by tyrosine transaminase (*TAT*) and aromatic acid dehydrase (*ADT*, *PDT*), which inhibited the germination of the secondary trunk at the initial stage of germination. Phenylalanine was involved in phenylpropane biosynthesis and phenylalanine metabolism. In addition, the tryptophan metabolism-related genes *Gb04566*(*SUR1*)and *Gb24946*(*amiE*), and IAA signal response-related genes *Gb14994*(*AUX1*), *Gb15664*(*SAUR*), and *Gb29888*(*SAUR*) in plant hormone signal transduction pathway were also enriched, all up-regulated in the secondary trunk group. Tryptophan is the substrate of IAA synthesis. When *G. biloba* is stimulated by the certain internal and external environment, the up-regulated expression of *Gb04566* (*SUR1*) promotes the accumulation of indole-3-acetonitrile, which increases the content of indole-3-acetic acid and upregulates the expression of IAA inflow vector gene in the tryptophan metabolic pathway. Finally, the IAA response gene SAUR receives and responds to the IAA signal to stimulate the development of the secondary trunk. In addition, the biosynthetic pathway of phenylalanine, tyrosine, and tryptophan can also produce L-tryptophan, which facilitates the accumulation of indole-3-acetonitrile and thus the synthesis of indole-3-acetic acid. At the same time, indole-3-acetic acid can indirectly promote the glycolysis pathway, which uses starch and sucrose metabolism as substrates to provide energy for the development of the secondary trunk.

The biosynthesis pathways of phenylalanine, tyrosine and tryptophan are important pathways linking tryptophan and phenylalanine, promoting the synthesis of acetyl-CoA through the intermediate products such as phenylpyruvate, phenylacetic acid, phenylalanine, and 4-hydroxy-2-oxy valerate. Acetyl-CoA is involved in phenylalanine metabolism, ketone body biosynthesis, phenylpropane biosynthesis, fatty acid synthesis and fatty acid prolongation. In addition, the carbon fixation pathway in glycolysis and photosynthetic organisms can also produce phenylalanine through the shikimic acid pathway (Sun, 2019).

GA_3 can break the dormancy of buds and promote secondary trunk germination (Mäkilä et al., 2023), and the same results were obtained in the transcriptome. Genes related to gibberellin synthesis including *Gb33056*(gibberellin 20-oxidase), *Gb04557*(*GA_3*, *CYP701*) and *Gb10062*(*GA_3*, *CYP701*), and genes related to plant hormone signal transduction including *Gb36988*(*GID1*)and genes promoting diterpenoid biosynthesis including *Gb11300*, *Gb19738*, *Gb39796*, *Gb10410*, *Gb24027*(*HMGCR*) and *Gb29808*(*FDPS*) were enriched in the diterpenoid biosynthesis pathway. The terpene skeleton biosynthesis pathway is based on glycolysis to promote diterpene biosynthesis. Firstly, acetyl-CoA is synthesized, and the synthesis of isoamyl diphosphate is promoted by hydroxy methyl glutaryl-CoA

synthase and hydroxy methyl glutaryl-CoA reductase (*HMGCR*). And then Farnesyl pyrophosphate synthase (*FDPS*) which is the key enzyme related to diterpenoids synthesis promotes the synthesis of farnesyl pyrophosphate and further promotes the diterpenoid biosynthetic pathway. In the diterpene biosynthesis pathway, the terpene skeleton biosynthesis is used as the substrate to synthesize geranylgeranyl pyrophosphate which is the substrate of gibberellin to further synthesized kaurene. Through the up-regulated expression of *Gb04557* and *Gb10062*, kaurene oxidase (GA_3, *CYP701*) promotes the synthesis of GA_{12} aldehyde and further synthesizes various gibberellins, including GA_3. The accumulation of gibberellin can promote the phytohormone signal transduction pathway to up-regulate the expression of gibberellin receptor *GID1* (*Gb36988*). And it can break the dormancy of secondary trunk buds in the cortex of the *G. biloba* stem and promote the germination and growth of the secondary trunk. In addition, isoamyl diphosphate, an intermediate product in the terpene skeleton biosynthetic pathway, can also promote zeatin biosynthesis and may play a role in the development of the secondary trunk.

The secondary trunk may be a reproductive strategy of *G. biloba* to sustain the life system when *G. biloba* reached its growth limit or to resist the adverse environment. It is a natural clonal multi-generation reproduction phenomenon. *G. biloba* can continuously produce new secondary trunks to make the mother tree that has reached its life limit rejuvenate and improve adaptability. Reproduction and renewal through secondary trunks is the key to the survival of *G. biloba* after catastrophes (Xiang et al., 2000; Fu et al., 2013).

5 Conclusions

The development process of secondary trunk of *G. biloba* can be divided into four stages, namely, dormancy stage, differentiation stage, conduction tissue formation stage, and germination stage. The contents of IAA, ZA, and GA_3 in the budding stage of secondary trunk development were higher than those in the lateral branches, breaking dormancy of dormant buds and promoting secondary trunk germination. After the germination stage, the content of GA_3 significantly decreased, while the content of IAA and ZA significantly increased, and the secondary trunk entered the elongation growth stage. In secondary trunk, the ratio of (IAA+ZA+GA_3)/ABA is greater than that of lateral branches, while the ratio of IAA/ZA is extremely smaller than that of lateral branches, so the growth speed of secondary trunk is higher than that of lateral branches. Through transcriptome sequencing, the differentially expressed genes were screened out, and the key regulatory pathways for the occurrence and development of secondary trunk was sorted out. Phenylalanine metabolism and phenylpropane biosynthesis may inhibit the germination of early secondary trunk buds. After being stimulated by the certain internal and external environment, *G. biloba* synthesizes gibberellin using terpene skeleton biosynthesis as the substrate, which can promote the plant hormones signal transduction pathway and upregulate the expression of gibberellin receptor (*GID1*) to break the dormancy state of secondary trunk buds. At the same time, genes related to IAA synthesis are upregulated and indole-3-acetic acid content is increased, leading to the up-regulated expression of IAA intracellular vector genes. The IAA response gene (*SAUR*) receives and responds to IAA signals to promote the development of the secondary trunk.

Data availability statement

The data presented in the study are deposited in the NCBI Sequence Read Archive (SRA), accession numbers SRR23730290, SRR23730291, SRR23730292, SRR23730293, SRR23730294, SRR23730288, SRR23730289, SRR23730295, SRR23730296, SRR23730297, SRR23730298, SRR23730299.

Author contributions

L-MS designed the research; Z-YC and L-MS carried out the experiments; Z-YC and L-NS analyzed the data and wrote the manuscript; L-NS revised the manuscript. Conceptualization, L-MS. Writing—original draft preparation, Z-YC and L-NS. Writing—review and editing, L-MS. Data curation, X-YZ and X-JK. Experiments preformation, Z-YC, L-NS, QZ, and X-LH. Supervision, funding acquisition, and project administration, L-MS. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the Subject of Key R & D Plan of Shandong Province (Major Scientific and Technological Innovation Project) "Mining and Accurate Identification of Forest Tree Germplasm Resources (2021LZGC023)". The Youth Fund of Natural Science Foundation of Shandong Province (ZR2020QC067). Introduction and training plan of young creative talents in universities of Shandong Province: Research group of forest tree biotechnology.

Acknowledgments

Thanks to Metware Biotechnology Co., Ltd. (Wuhan, China) for the testing services.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1161693/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Q30 quality control chart.

SUPPLEMENTARY FIGURE 2
(A) Cluster Map of DEGs. **(B)** Statistics of differentially expressed genes in each comparison group of *G. biloba*. **(C)** Volcano map of DEGs in each comparison group of *G. biloba*. The abscissa is the expression multiple, and the ordinate is the significance degree of the change of expression quantity (P <0.05). Orange indicates genes that are significantly up-regulated in differential genes. Green indicates genes that are significantly down-regulated, and gray indicates genes that are not significantly different.

SUPPLEMENTARY FIGURE 3
Distribution of q-values for enriched GO entries in each comparison group of *G. biloba*. Take the enriched GO entries in all samples for analysis, the ordinate is the GO entry and the abscissa is the name of different comparison groups. Different colors represent different degrees of enrichment.

SUPPLEMENTARY FIGURE 4
Histogram of GO statistics of differentially expressed genes in each comparison group of *G. biloba*. The abscissa is each major category under GO, which represents various biological processes, cell components and molecular functions. The left ordinate is the proportion of this category, and the right ordinate is the specific number of genes in this category. Different colors represent different groups (differentially expressed genes are up-regulated and down-regulated).

SUPPLEMENTARY FIGURE 5
Q-value enrichment chart of GO items in each comparison group of *G. biloba*. GO items enriched in all comparison groups were taken for analysis, and Q-value enrichment analysis was conducted for the three categories respectively. The ordinate is the secondary entry of GO and the abscissa is the degree of enrichment. Each point represents the degree of enrichment of the GO entry. The closer the color is to red, the higher the degree of enrichment. The size of each point indicates the number of genes enriched in the GO entry. The larger the point, the more genes are enriched in the GO entry, and vice versa.

SUPPLEMENTARY FIGURE 6
Distribution of q-values of enrichment pathways in each comparison group of *G. biloba*.

References

Beveridge, C. A., Weller, J. L., Singer, S. R., and Hofer, J. M. I. (2003). Axillary meristem development. budding relationships between networks controlling flowering, branching, and photoperiod responsiveness. *Plant Physiol.* 131, 927–934. doi: 10.1104/pp.102.017525

Cao, F. (2007). *Ginkgo biloba in China* (Beijing: China Forestry Press).

Faust, M., Erez, A., Rowland, L. J., Wang, S. Y., and Norman, H. A. (1997). Bud dormancy in perennial fruit trees: physiological basis for dormancy induction, maintenance, and release. *Hort Sci.* 32, 623–629. doi: 10.21273/HORTSCI.32.4.623

Fu, Z. (2014). *Study on the ontogeny of ginkgo biloba rooted chichi* (Shandong Province (Taian: Shandong Agricultural University).

Fu, Z., Xing, S., Li, Z., Liu, L., Ren, J., and Liu, Y. (2013). Growth characteristics of secondary trunk *Ginkgo biloba* in shengsheng garden in linyi city. *J. Plant Genet. Resour.* 14, 764–770. doi: 10.13430/j.cnki.jpgr.2013.04.031

Fujii, K. (1985). On the nature and origin of so-called "chichi"(nipple) of *Ginkgo biloba* l. *Bot. Mag(Tokyo).* 9, 444–450. doi: 10.15281/jplantres1887.9.440

Horvath, D. P., Anderson, J. V., Chao, W. S., and Foley, M. E. (2003). Knowing when to grow: signals regulating bud dormancy. *Trends Plant Sci.* 8, 534–540. doi: 10.1016/j.tplants.2003.09.013

Lin, X., and Zhang, D. (2004). The origin analysis of *Ginkgo biloba* population in tianmu mountain. *Forestry Sci.* 2, 28–31. doi: 10.11707/j.1001-7488.20040205

Liu, X., Sun, L., Wu, Q., Men, X., Yao, L., Xing, S., et al. (2018). Transcriptome profile analysis reveals the ontogenesis of rooted chichi in *Ginkgo biloba* l. *Gene* 669, 8–14. doi: 10.1016/j.gene.2018.05.066

Mäkilä, R., Wybouw, B., Smetana, O., Vainio, L., Solé-Gil, A., Lyu, M, et al. (2023). Gibberellins promote polar auxin transport to regulate stem cell fate decisions in cambium. *Nat. Plants*, 1-14. doi: 10.1038/s41477-023-01360-w

Mauriat, M., Sandberg, L. G., and Moritz, T. (2011). Proper gibberellin localization in vascular tissue is required to control auxin-dependent leaf development and bud outgrowth in hybrid aspen. *Plant J.* 67, 805–816. doi: 10.1111/j.1365-313X.2011.04635.x

Men, X., Sun, L., Li, Y., Li, W., and Xing, S. (2021). Multi-omics analysis reveals the ontogenesis of basal chichi in *Ginkgo biloba* l. *Genomics* 113, 2317–2326. doi: 10.1016/j.ygeno.2021.05.027

Müller, D., and Leyser, O. (2011). Auxin, cytokinin and the control of shoot branching. *Ann. Bot.* 107, 1203–1212. doi: 10.1093/aob/mcr069

Ni, J. (2015). *Regulation of gibberellin on growth of lateral buds of jatropha curcas* (Anhui Province (Hefei: China Science and Technology University).

Qi, W., Sun, F., Wang, Q., Chen, M., Huang, Y., Feng, Y., et al. (2011). Rice ethylene-response AP2/ERF factor OsEATB restricts internode elongation by down-regulating a gibberellin biosynthetic gene. *Plant Physiol.* 157, 216–228. doi: 10.1104/pp.111.179945

Seward, A. C. (1938). The story of the maidenhair tree. *Sci. Prog.* 32, 420–440. doi: 10.2307/43412201

Silverstone, A. L., Mak, P. Y. A., Martinez, E. C., and Sun, T. (1997). The new RGA locus encodes a negative regulator of gibberellin response in *Arabidopsis thaliana*. *Genetics* 146, 1087–1099. doi: 10.1093/genetics/146.3.1087

Sun, L. (2019). *Regulation mechanism analysis of the transcription, protein and metabolic of the flavonoids synthesis of the leaves of ginkgo biloba* (Shandong Province (Taian: Shandong Agricultural University).

Tredici, P. D. (1992a). Natural regeneration of *Ginkgo biloba* from downward growing cotyledonary buds(basal chichi). *Am. J. Bot.* 79, 522–530. doi: 10.1002/j.1537-2197.1992.tb14588.x

Tredici, P. D. (1992b). Where the wild ginkgos grow. *Arnoldia* 52, 2–11. doi: 10.2307/44944861

Wang, X., Yang, Z., Zhang, S., Li, H., and Li, S. (2013). Digital gene expression profile analysis of early adventitious buds in *Arabidopsi*. *Chin. J. Bioengineering* 29, 189–202. doi: 10.13345/j.cjb.2013.02.012

Xiang, Y., Xiang, B., Zhao, M., and Wang, Z. (2000). Investigation report on natural forest and *Ginkgo biloba* populations in West tianmu mountain in zhejiang province. *Guizhou Sci.*, 77–92. doi: CNKI:SUN:GZKX.0.2000-Z1-011

Xing, S. (1996). Tree tumor of *Ginkgo biloba*. *Plant J.* 3, 29–30. doi: CNKI:SUN:LYKT.0.1996-02-000

Xing, S. (2013). *Chinese Ginkgo biloba germplasm resources* (Beijing: China Forestry Press).

Xing, S., and Miao, Q. (1996). Study on the biological characteristics of secondary trunk. *Forestry Sci. Technol. Newsletter.* 2, 6–9.

Yang, L., Chen, Z., Zhao, Y., Niu, Z., Niu, S., Shi, X., et al. (2020). The regulation of plant signal transduction on the release of dormancy of 'Xiahei' grape winter buds. *Mol. Plant Breed.* 18, 1438–1446. doi: 10.13271/j.mpb.018.001438

Zhang, N. (2019). *The study on the relationship of floral bud differentiation and endogenous hormones of xanthoceras sorbifolia bunge* (Beijing: Beijing Forestry University).

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 08 June 2023
DOI 10.3389/fpls.2023.1161257

OPEN ACCESS

EDITED BY
Kai-Hua Jia,
Shandong Academy of Agricultural Sciences, China

REVIEWED BY
Shang-Qian Xie,
University of Idaho, United States
Zhenqiao Song,
Shandong Agricultural University, China

*CORRESPONDENCE
Shengkui Zhang
zsk8920@foxmail.com
Bingmiao Gao
hy0207086@hainmc.edu.cn

RECEIVED 08 February 2023
ACCEPTED 17 May 2023
PUBLISHED 08 June 2023

CITATION
Pan K, Dai S, Tian J, Zhang J, Liu J, Li M, Li S, Zhang S and Gao B (2023) Chromosome-level genome and multi-omics analyses provide insights into the geo-herbalism properties of *Alpinia oxyphylla*.
Front. Plant Sci. 14:1161257.
doi: 10.3389/fpls.2023.1161257

Chromosome-level genome and multi-omics analyses provide insights into the geo-herbalism properties of *Alpinia oxyphylla*

Kun Pan[1], Shuiping Dai[1], Jianping Tian[1], Junqing Zhang[1,2], Jiaqi Liu[1], Ming Li[1], Shanshan Li[1], Shengkui Zhang[3]* and Bingmiao Gao[1,2]*

[1]Hainan Provincial Key Laboratory for Research and Development of Tropical Herbs, Haikou Key Laboratory of Li Nationality Medicine, Hainan Quality Monitoring and Technology Service Center for Chinese Materia MedicaRaw Materials, School of Pharmacy, Hainan Medical University, Haikou, Hainan, China, [2]Academician Workstation of Hainan Province and The Specific Research Fund of The Innovation Platform for Academicians of Hainan Province, Haikou, Hainan, China, [3]School of Bioengineering, Qilu University of Technology (Shandong Academy of Sciences), Jinan, Shandong, China

Introduction: *Alpinia oxyphylla* Miquel (*A. oxyphylla*), one of the "Four Famous South Medicines" in China, is an essential understory cash crop that is planted widely in the Hainan, Guangdong, Guangxi, and Fujian provinces. Particularly, *A. oxyphylla* from Hainan province is highly valued as the best national product for geo-herbalism and is an important indicator of traditional Chinese medicine efficacy. However, the molecular mechanism underlying the formation of its quality remains unspecified.

Methods: To this end, we employed a multi-omics approach to investigate the authentic quality formation of *A. oxyphylla*.

Results: In this study, we present a high-quality chromosome-level genome assembly of *A. oxyphylla*, with contig N50 of 76.96 Mb and a size of approximately 2.08Gb. A total of 38,178 genes were annotated, and the long terminal repeats were found to have a high frequency of 61.70%. Phylogenetic analysis demonstrated a recent whole-genome duplication event (WGD), which occurred before *A. oxyphylla's* divergence from W. villosa (~14 Mya) and is shared by other species from the Zingiberaceae family (Ks, ~0.3; 4DTv, ~0.125). Further, 17 regions from four provinces were comprehensively assessed for their metabolite content, and the quality of these four regions varied significantly. Finally, genomic, metabolic, and transcriptomic analyses undertaken on these regions revealed that the content of nootkatone in Hainan was significantly different from that in other provinces.

Discussion: Overall, our findings provide novel insights into germplasm conservation, geo-herbalism evaluation, and functional genomic research for the medicinal plant *A. oxyphylla*.

KEYWORDS

***Alpinia oxyphylla*, genome, metabolomics, geo-herbalism, transcriptomics, nootkatone, valenene synthase**

Introduction

Zingiberaceae is a large, fragrant pantropical family consisting of 1,600 species divided among about 50 genera (Christenhusz and Byng, 2016). The plants are distributed throughout tropical Africa, Asia, and the Americas (Ra Mans et al., 2019). Zingiberaceae plants contain many bioactive terpenoids, flavonoids, and polyphenols that are economically important as traditional medicines, spices, and cosmetics. *Alpinia oxyphylla* Miquel (*A. oxyphylla*) is one of the type species in the genus *Alpinia* (with more than 230 species) and has been approved by China Food and Drug Administration as a medicine and food homology species. Additionally, *A. oxyphylla* has been used as a medicinal and edible plant for hundreds of years. Its fruit when dried or baked with salt is referred to as Fructus Alpiniae Oxyphyllae (FAO), and the Chinese medicine name is "Yi Zhi, Yi Zhi Ren"—in traditional Chinese medicine (TCM), it is one of the "Four Famous South Medicines". FAO is commonly used as a medicine for warming the kidney and spleen, securing essence and arresting polyuria, and stopping diarrhea and saliva in TCM. Classified Materia Medica (1097 A.D.-1108 A.D.) and the Compendium of Materia Medica (1552 A.D.-1578 A.D.) document the use of FAO either alone or in combination with other herbal medicines. Various pharmacological properties of FAO have been reported, such as anti-inflammatory (Yu et al., 2020), anti-oxidant (Thapa et al., 2021), anti-diarrheal (Zhang et al., 2013), anti-Alzheimer's disease (Li et al., 2021b), promoting neuronal regeneration and resisting neurodegenerative diseases (He et al., 2018), and anti-diuretic and diuretic (Li et al., 2016). The unique medicinal and flavor characteristics of *A. oxyphylla* are associated with a variety of metabolites, including rich terpenoids (Chen et al., 2014), diarylheptanoids (Chen et al., 2014), and diarrhea (Zhang et al., 2013). However, there have been some reports on the identification and functional characterization of genes related to the biosynthesis of flavonoids (Yuan et al., 2021) and terpenoids (Yang et al., 2022) in Zingiberaceae. There are multiple kinds of terpenoids, diarylheptanes, and flavonoids in *A. oxyphylla*, and the biosynthetic pathway remains largely unexplored.

A. oxyphylla is a kind of herbaceous plant, which thrives in tropical rain forest and evergreen broad-leaved forests. It is commonly cultivated under rubber forest, pine forest, and eucalyptus forest in the Hainan, Guangdong, Guangxi, and Fujian provinces. Our study on various *A. oxyphylla* cultivation regions reveals that the main components vary depending on the region. Biological properties of medicinal plants are highly influenced by the environment; thus, the chemical composition and content of medicinal plants are dependent on their environment (Ma et al., 2010; Ma et al., 2018; Li et al., 2020b). Traditional Chinese medicinal philosophy only recognizes and values geo-herbs as authentic medicines, and only these are considered safe and of high quality (Chen et al., 2017).

Our previous study conducted an investigation on the wild and cultivated populations of *A. oxyphylla* from various geographical locations and discovered that the individual genetic diversity of *A. oxyphylla* is significantly high (Wang et al., 2012). Despite observing instances of inbreeding and gene flow (Nm=1.453) in *A. oxyphylla* populations, the phylogenetic relationship among certain accessions remains relatively distant due to the notable genetic differentiation of the wild population compared to that of the cultivated population. Additionally, the grouping of all accessions almost completely aligns with their geographical origin, demonstrating the evident regional differentiation of this species (Zou et al., 2013). Furthermore, our research delved into the transcriptome and metabolome of different tissues and fruit development stages of *A. oxyphylla* and identified differentially expressed genes associated with the biosynthesis of flavonoids (Yuan et al., 2021) and terpenoids (Pan et al., 2022), highlighting the primary medicinal component, nootkatone, which primarily significantly accumulates in seeds during the late stage of development. Unfortunately, the lack of genomic data has obstructed a proper attribution of these compounds to specific genes in the biosynthetic pathway. However, with the recent completion of genome sequencing in several Zingiberaceae plants, employing metabonomics, transcription, and genome association methods to analyze the main active components and related biosynthetic pathways has become crucial. Thus, obtaining a comprehensive understanding of the genetic structure of *A. oxyphylla* is pivotal in order to furnish a basis for geo-herbalism evaluation in different planting areas.

Currently, numerous species have seen completed molecular markers, gene mining and cloning, and the functional identification of important agronomic traits, marking the onset of the post-genome era. However, research on Zingiberaceae is still in its nascent stages, with only a few genomes having been sequenced, including those of *Amomum tsao-ko* (Li et al., 2022), *Zingiber officinale* (Cheng et al., 2021b; Li et al., 2021a), *Wurfbainia villosa* (Yang et al., 2022), *Curcuma alismatifolia* (Liao et al., 2022), *Curcuma longa* (Chakraborty et al., 2021), and *Alpinia nigra* (Ranavat et al., 2021). It is worth noting that the genomes of *Curcuma longa* and *Alpinia nigra* are only a sketch, meaning that the whole-genome information of genus *Alpinia* has yet to be revealed or reported. Obtaining a high-quality genome will provide sufficient data for solving phenotypic and genetic variations, advancing studies on the molecular basis of characteristic metabolites in *A. oxyphylla*, and guiding breeding strategies aimed at improving characteristic components.

The present study assembled a high-quality genome, transcriptome, and metabolism of *A. oxyphylla*. It has also revealed that nootkatone can be employed as an indicator to identify the *A. oxyphylla*'s geo-herbalism. The expansion of the valencene synthase gene family is regarded as being responsible for the regional variation of nootkatone content. These findings offer insights into molecular breeding and functional gene identification related to important traits of *A. oxyphylla*. Furthermore, the high-quality reference genome of *A. oxyphylla* presented in this study provides a valuable resource for exploring the evolution, speciation, and geo-herbalism of other species in the Zingiberaceae family.

Materials and methods

Plant materials

The plant materials were collected from a cultivar of "changyuanguo" from Danzhou (19°51′N, 109°50′E), Hainan Province, China, which is considered the authentic production

area of *A. oxyphylla* (Figure 1A). Its leaves were used for genomic sequencing, and its roots were used for flow cytometry and karyomorphological analysis. The fruit utilized for transcriptome and metabolome analysis consisted of an oval-shaped variety and a more commonly found and higher yielding type. A total of three and six biological replicates, respectively, were gathered from Baoting, Danzhou, Naning, and Zhangpu during May of 2019, as outlined in Supplementary Table 1. Tissue from the fruit was uniformly collected at 55-65 days after fruiting, promptly frozen in liquid nitrogen, and stored at -80°C.

Somatic chromosome numbers and karyomorphological analysis

We used living tissue samples from the root tips of *A. oxyphylla* and promoted cell cycle synchronization as needed. The cell mitosis was fixed in the metaphase phase, and then the cellulose enzyme process was used to obtain mid-mitotic division phase. Finally, we performed 4',6-diamidino-2-phenylindole (DAPI) staining and capture under a fluorescence microscope (Leica DM2500) to avoid light microscopy.

FIGURE 1
Morphology of *A. oxyphylla* and overview of *A. oxyphylla* gnome. **(A)** Morphological characteristics of *A. oxyphylla*. (a), Plant; (b), inflorescence; (c), fruits; (d), dried fruit; (e) seed and pericarp. **(B)** Circos plot of *A. oxyphylla* genome assembly. Elements are arranged in the following scheme (from inner to outer): (a) GC content; (b) gene density shown as the distribution densities from high (red) to low (green); (c) repeat sequence densities; (d) non-coding RNA (ncRNA) density; (e) tandem repeat density; (f) the density of other repeats except tandem repeats. The window size is 1Mb.

Twenty scattered cells with good morphology were selected from each material for chromosome counting, and five metaphone cells with clear chromosome morphology and no overlap were used for karyotype analysis. After measuring its length, the homologous pairing was carried out according to the morphological characteristics of chromosomes and the analysis of the measured data, and the karyotype map was arranged (Supplementary Table 2).

Genome size and heterozygosity estimation

After nuclear extraction and staining, the nucleus of *A. oxyphylla* was prepared to for measurement by flow cytometry. *Manihot esculenta* and *Solanum lycopersicum*, with their known genome size, were selected as internal reference species. CFlow Plus 1.0.264.15 software was used for data collection.

To determine the genome characteristics of *A. oxyphylla*, K-mer (k = 17), analysis was performed on the Illumina Hiseq platform with the insert size of 350bp. SOAPnuke (V1.6.5) was used to filter raw sequencing data. The software "kmer_freq_stat", independently developed by Biomarker Technologies Co., Ltd, was used to calculated the depth distribution map of each K-mer, and the heterozygosity rate was calculated according to Marcais and Kingsford (2011).

Library construction and sequencing

Young leaf samples were collected in May 2020, and their total genomic DNA was extracted using the DNAsecure Plant Kit (TIANGEN). The quality of isolated genomic DNA was verified by electrophoresis and the Qubit dsDNA hs assay kit in Qubit ®3.0 Flurometer (Life technologies, AC, USA); 0.3μg DNA per sample was used for library generation. Fragments (350 bp) were generated and used to construct a sequencing library. At last, 150-bp paired-end reads were used to sequence on the Truseq Nano DNA HT Sample Prep Kit (Illumine HiSeq X-Ten platform, USA).

The circular consensus sequencing (CCS) approach was selected for single-molecule real-time (SMRT) long-read sequencing. Five 20-kb insert libraries were prepared using SMRTbell Express Template Prep Kit 2.0e, and a total of nine SMRT cells with 80.10 Gb of sequence data (54-fold coverage of the genome) were obtained and sequenced on the PacBio Sequel II platform.

For Hi-C sequencing, we used 1% formaldehyde solution in an MS buffer (50 mM NaCl; 10 mM potassium phosphate, pH 7.0; 0.1M sucrose) to fix fresh leaves at room temperature for 30 min in a vacuum. After fixation, the leaves were incubated under a vacuum in the MC buffer and then resuspended in a nuclei isolation buffer and filtered. Chromatin extraction and DNA were digested by the HindIII restriction enzyme (NEB), and then they were labeled with biotin on the DNA ends and incubated. Proteinase K was added to reverse cross-linking before ligation. After removing the unligated ends, the purified DNA was sheared to a size of 300-500 bp fragments, and we repaired the DNA ends. Then, the separated DNA fragments were labeled by biotin with Dynabeads® M-280 Streptavidin (Life Technologies). We used the Illumina Hiseq X Ten sequencer to control the Hi-C libraries quality and sequence them.

In addition, the total RNA of the same *A. oxyphylla* individual was extracted from seven tissues (root, stem, leaf, fruit, seed, pericarp, and suction bud) with three biological replicates, and its RNA was extracted using a RNeasy plant mini kit (Qiagen). The RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA) was employed to assess its integrity.

Genome assembly and quality assessment

The *A. oxyphylla* genome was assembled as follows: firstly, after quality controlling of the raw Hi-C data using HI-C-PRO (version 2.8.0) (Servant et al., 2015), contigs were assembled from CCS clean reads with default parameters using Hifiasm (V 0.12) (Cheng et al., 2021a). Secondly, the high-quality paired-end Hi-C reads were first mapped to the reference *A. oxyphylla* genome (GRCm38/mm10) using the Burrows-Wheeler Aligner (BWA) software (Li and Durbin, 2009). We converted the alignment files to BAM files using SAMtools (Li and Durbin, 2009), and then we improved the alignment results; only uniquely alignable pairs reads (mapping quality >20) were selected for further analysis, and we filtered out low-quality sequences using FASTP (version 0.12.6) (Chen et al., 2018a; Chen et al., 2018b). The present study involved a manual inspection of segments that displayed conflicting associations with information obtained from the raw scaffold. Subsequently, a chromosome-level assembly was generated from the draft contig-level assembly by utilizing the LACHESIS34 software, which employs the ligating adjacent chromatin enables scaffolding *in situ* method (Burton et al., 2013).

The accuracy and completeness of the genome assembly were evaluated using several methods. Firstly, the Hi-C interaction heatmap was employed to determine the organization of the genome. Secondly, the presence of contamination in the sequencing data was assessed using GC depth scatter plots and GC content. Thirdly, the alignment of the genome sequencing to the assembled genome was conducted to assess the coverage. Finally, two core eukaryotic gene datasets were utilized to assess the completeness of the genome: Benchmarking Universal Single-Copy Orthologs (BUSCO), accessible at http://busco.ezlab.org/, and Core Eukaryotic Genes Mapping Approach (CEGMA), accessible at http://korflab.ucdavis.edu/datasets/cegma/.

Genome annotation

The tandem repeat sequence was predicted ab initio using TRF (http://tandem.bu.edu/trf/trf.html). Repeat regions were extracted *via* homolog prediction by employing the Repbase (Jurka et al., 2005) database and the RepeatMasker (http://www.repeatmasker.org/) software, along with its in-house scripts (RepeatProtein Mask), using default settings. To build a *de novo* repetitive elements database, LTR_FINDER (http://tlife.fudan.edu.cn/ltr_finder/), RepeatScout (http://www.repeatmasker.org/), and RepeatModeler

(http://www.repeatmasker.org/Repeat Modeler.html) were utilized with default parameters. The resultant TE library consisted of all repeat sequences longer than 100bp and with gaps "N" less than 5%. The Repbase and *de novo* TE libraries constituted a custom library supplied to RepeatMasker for identifying DNA-level repeats. Protein-coding genes were predicted using *de novo* gene prediction, homolog prediction, and RNA-seq-based prediction. For the former, ab initio-based gene prediction was performed using Augustus (v3.2.3), Geneid (v1.4), Genescan (v1.0), GlimmerHMM (v3.04), and SNAP (2013-11-29). Homologous protein sequence data were obtained from Ensembl/NCBI/others, and TblastN (v2.2.26; E-value ≤ 1e-5) was used to align the protein sequences to the genome. (Birney et al., 2004) software was then used to predict the gene structure contained in each protein region *via* accurate spliced alignments with the homologous genome sequences. Then, RNA-Seq reads from different *A. oxyphylla* tissues were mapped to the assembled genome by utilizing TopHat (v2.0.11) (Trapnell et al., 2009). Furthermore, GeneMarkS-T (v5.1) 48 was employed to predict genes based on the assembled transcripts (Tang et al., 2015). Finally, the EVM software (v1.1.1) was utilized to combine gene models from the above approaches (Haas et al., 2008).

Gene function predictions were determined by aligning protein sequences to Swiss-Prot *via* Blastp, using a threshold of E-value ≤ 1e-5 for the best match. Motifs and domains were annotated with InterProScan (v4.8) (Mulder and Apweiler, 2007) by querying against a range of publicly available databases, such as ProDom, PRINTS, Pfam, SMRT, PANTHER, and PROSITE. The corresponding InterPro entry for each gene was used to assign gene ontology (GO) IDs. Additionally, we mapped each gene set to a KEGG pathway and identified the best match for each gene.

Phylogenetic tree construction and evolution rate estimation

To determine the phylogenetic relationships between *A. oxyphylla* and other closely related species, we utilized protein sequences from a set of 832 single-copy ortholog genes. These sequences were aligned using the mafft (v7.205) program designed by (Katoh et al. (2009) and subsequently curated with gblocks (v0.91b). The resulting coding DNA sequences (CDS) alignments were concatenated, guided by the protein alignment, and used to construct a phylogenetic tree with the aid of iqtree (v1.6.11) developed by Nguyen et al. (Nguyen et al., 2015).

Gene family analysis

To cluster families of protein-coding genes, we analyzed proteins from the longest transcripts of each gene from *A. oxyphylla* and other nine closely related species, namely, *Arabidopsis thalian, Sorghum bicolor, Oryza sativa*, *Ananas comosus*, *Musa balbisiana*, *Musa acuminata*, *Zingiber officinale*, *Curcuma alismatifolia*, and *Wurfbainia villosa*. We used the OrthoFinder (v2.5.1) software (Emms and Kelly, 2019) to compare protein-coding sequences within the genomes of *A. oxyphylla* and the other nine species. We then annotated the obtained gene families using the Pfam V33.1 database (Mistry et al., 2021). Using the identified gene families and predicted divergence time, we constructed a phylogenetic tree of these species and analyzed gene family expansion and contraction using CAFE (Han et al., 2013). In CAFE, a random birth and death model is proposed, allowing for the study of gene gain or loss across a specified phylogenetic tree. We calculated a conditional p-value for each gene family and considered those with a conditional p-value of less than 0.05 to have an accelerated rate of gene gain or loss.

Whole-genome duplication and the insert time of LTR calculation

The identification of whole-genome duplication events in *A. oxyphylla* was performed using the synonymous mutation rate (Ks) method and the fourfold synonymous third-codon transversion rate (4DTv) method. Initially, the software wgd (v1.1.1), developed by Zwaenepoel and Van De Peer (2019), and a custom script (https://github.com/JinfengChen/Scripts) were utilized for this purpose. The identification of full-length long terminal repeat retrotransposons (fl-LTR-RTs) was achieved through the utilization of both LTRharvest (v1.5.10) (Ellinghaus et al., 2008) and LTR_finder (v1.07) (Xu and Wang, 2007). LTR_retriever (Ou and Jiang, 2018) was then used to produce high-quality intact fl-LTR-RTs and a non-redundant LTR library. To determine the distance between the flanking sequences on both sides of LTR, mafft (v7.205) (Katoh et al., 2009) was used for comparison, and the Kimura model in EMBOSS (v6.6.0) (Rice et al., 2000) was employed for distance calculation.

RP-HPLC analysis

After 14 days of drying at 45°C in a drying oven, the fruit was polished into powder and accurately weighed, 25mL 70% ethanol was added, then ultrasonic extraction was performed for three times, each time for 30 minutes, centrifugation was performed several times and the supernatant was taken to obtain the test product solution. Chromatographic column: Phenomenex Gemini C6-phenyl (250 mm× 4.6mm, 5μm); mobile phase: acetonitrile (A)–water (B) solution, gradient elution [0-5 min, A-B(40:60); 6-26, A-B(60:40); 27-32, A-B(40:60)]; flow rate: 1.0 ml/min^{-1}; detection wavelength: 240 nm; column temperature: 30°C. The result is shown in Supplementary Table 3-6.

Transcriptome sequencing and analysis

In this study, total RNA was extracted from A. oxyphylla using the RNAsecure Plant Kit (TIANGEN). The resulting RNA was assessed for purity using a NanoPhotometer® spectrophotometer (IMPLEN, CA, USA) and for integrity using the RNA Nano 6000 Assay Kit on the Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, USA). Sequencing libraries were generated

using the NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (NEB, USA), following the manufacturer's recommendations, with index codes added to attribute sequences to each sample. The total RNA content used was 1.5 μg. To cluster the index-coded samples, we employed the TruSeq PE Cluster Kit v3-cBot-HS (Illumina) and followed the manufacturer's instructions. The library preparations were subsequently sequenced using the Illumina HiSeq platform, generating paired-end reads.

Qualitative and quantitative analysis of metabolites

The metabolites were analyzed by injecting them into an LC-MS/MS system manufactured by Thermo Fisher in the USA, as described by Bondia-Pons et al. (2013). Subsequently, the raw files obtained by mass spectrometry were processed using CompoundDiscoverer 3.1 software, also developed by Thermo Fisher Scientific. Spectrogram processing and database searches were conducted to obtain qualitative and quantitative results for the metabolites. Quality control analysis was then performed to ensure the accuracy and reliability of the data. Multivariate statistical analysis methods, such as principal component analysis (PCA) and partial least square discriminant analysis (PLS-DA), were applied to the data to identify and analyze the metabolites. Finally, the biological significance of the metabolites was established using functional analysis of metabolic pathways.

Differential expression analysis and co-expression network analysis

The differential expression analysis of the two groups was carried out using the R package 'DESeq' version 1.10.1. To control the false discovery rate (FDR), the P-value was adjusted using the method of Benjamini and Hochberg (1995). Genes with an adjusted P-value of less than 0.05 were classified as differentially expressed. The R package 'GOseq', which utilizes the Wallenius non-central hypergeometric distribution to account for gene-length bias in DEGs, was used for GO enrichment analysis of the identified DEGs (Young et al., 2010). To test the statistical enrichment of DEGs in the KEGG pathways, the KOBAS software (Mao et al., 2005) was employed. The collected multidimensional data were subjected to regression and reduction analysis, including PCA and PLS-DA, with a focus on preserving original information to the fullest. This approach facilitated the identification of differential metabolites. Subsequently, correlation analysis between significantly altered genes from the transcriptome analysis and significantly altered metabolites from the metabolomics analysis was performed using the Pearson correlation coefficient (Pearson's r). This measure helped quantify the degree of association between differential genes and differential metabolites. Finally, to better understand the involvement of differential genes and differential metabolites in biochemical pathways and signal transduction pathways, all the obtained differential genes and differential metabolites were simultaneously mapped onto the KEGG pathway database.

qRT-PCR

Several genes and transcription factors (TFs) were chosen for RT-qPCR examination. The initial strand cDNA was produced through the application of the NovoScript® Plus all-in-one First Strand cDNA Synthesis SuperMix (gDNA Purge, Novoprotein, Shanghai, China). The gene-specific primers are enumerated in Supplementary Table 7.

Statistics and reproducibility

The data consisted of a minimum of three biological replicates. To compare different groups in pairwise fashion, statistical analysis was conducted through one-way ANOVA, which was followed by Dunnett's test with a significance threshold of $p < 0.05$.

Results

Determination of genome size and heterozygosity

The present study conducted chromosome number measurements on the root tips of a cultivated individual of *A. oxyphylla*, which had previously been sequenced for its genome. The results revealed the karyotype formula of *A. oxyphylla* is 2n=2x=44m+2sm, displaying a relative length range falling between 6.23% and 3.15% with an asymmetry coefficient of 57.6%. The karyotype type is 1A. There exist 24 pairs of 48 chromosomes, with clear 1-2 banding on the 1st, 2nd, 3rd, 6th, and 11th chromosome pairs, including satellited chromosomes. It was indicated that *A. oxyphylla* is a homologous diploid (Supplementary Figure 1). The genome size is 1.79Gb, which was determined through flow cytometry, and *Manihot esculenta* and *Solanum lycopersicum* were selected as reference species (Supplementary Figure 2 and Supplementary Table 8). Additionally, K-mer (Li et al., 2010) analysis was employed to evaluate the *A. oxyphylla* genome. The analysis indicated approximately 86.64 Gb of modified 17-mers, a primary peak distribution frequency appearing at depth = 40, and an estimated genome size of 2.14 Gb with 0.99% heterozygosity (Supplementary Figure 3 and Supplementary Table 9). Together, these findings provide evidence that the sequenced material possesses a diploid nature, confirming the karyotype formulae 2n=2x=48 in *A. oxyphylla*, which differs from report that suggested 2n=4x=48 in the genus *Alpinia* (Saenprom et al., 2018).

Genome sequencing, assembly, and annotation

The genome of *A. oxyphylla* was sequenced using PacBio and Illumina platforms. This resulted in clean subreads of 80.10Gb with 37.35X coverage depth and clean reads of 115.85Gb with 54.02X coverage depth, as shown in Supplementary Table 10. Additionally,

high-throughput chromosome conformation capture (Hi-C) libraries were constructed for *A. oxyphylla*, resulting in contigs totaling 2.08Gb in length, with a high contig N50 value of 76.96Mb and the longest contig of 152.6Mb. Scaffolds of 2.08Gb were collected with a scaffold N50 of 83.05Mb, with 24 scaffolds (2.00Gb) accounting for approximately 96.08% of all sequences anchored into 24 pseudochromosomes. As a result, we obtained a chromosome-level genome of *A. oxyphylla* consisting of 24 chromosomes with a total size of 2.08Gb, as indicated in Figure 2 and Supplementary Table 11.

The completeness of the assembled genome of *A. oxyphylla* was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO) and the Core Eukaryotic Genes Mapping Approach (CEGMA). Our BUSCO analysis revealed that 94.6% of the complete single-copy genes were assembled from 1614 Embryophyta-wide conserved single-copy genes. The fragmented and missing categories accounted for 1.9% and 3.5%, respectively. Additionally, CEGMA evaluation used 248 conserved genes from six eukaryotic model organisms to form a core gene library. Our evaluation showed that 235 genes were assembled with 94.76% accuracy (Supplementary Table 12). To further validate the accuracy of our assembly, fragments from the small fragment library were aligned to the assembled genome. Our results indicated that the alignment rate of all small reads fragments to

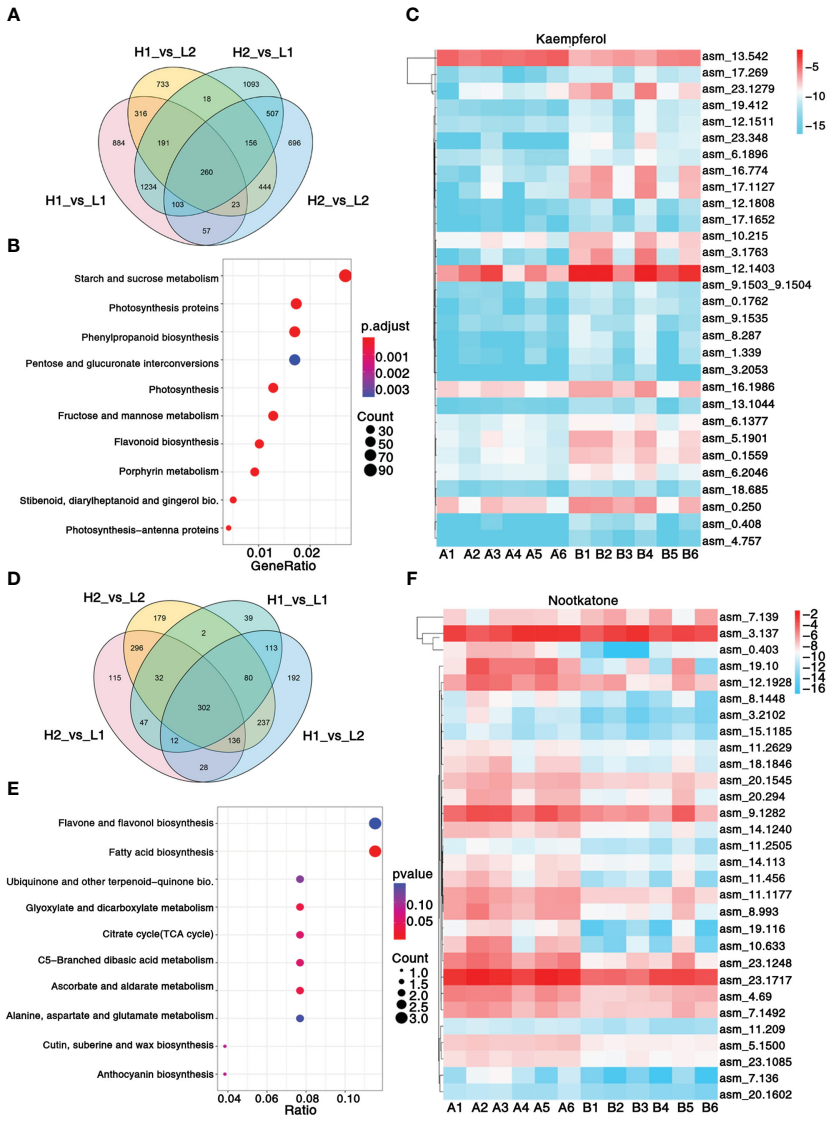


FIGURE 2
Transcriptomic and metabolic analysis and identification of differentially expressed genes. **(A)** Venn diagram shows the number of differentially expressed genes in four regions. **(B)** KEGG enrichment analysis of differentially expressed genes. **(C)** Heatmap showing the expression level of top 30 genes associated with pharmacodynamic component kaempferol. **(D)** Venn diagram shows the number of differentially expressed metabolites in four regions. **(E)** KEGG enrichment analysis of differentially metabolites. **(F)** Heatmap showing the expression level of top 30 genes associated with pharmacodynamic component nootkatone. The metabolisms KEGG annotation and correlation coefficient between genes and metabolites of this figure are from Supplementary Data 2, 3. Sample class: H1 (A1-A3): Danzhou; H2 (A4-A6): Baoting; L1 (B1-B3): Nanning; L2 (B4-B6): Zhangpu. The fpkm values are log2-based. Red and blue indicate high and low expression levels, respectively.

the genome was about 99.30%, while the coverage rate was roughly 99.95%, indicating that there was a good consistency between reads and the assembled genome (Supplementary Table 13). Furthermore, our analysis of the heterozygous SNP ratio showed that the *A. oxyphylla* genome assembly had a high single base accuracy of 0.481%. Additionally, our distribution analysis of GC content (39.69%) and average depth confirmed that the sample was not contaminated (Supplementary Figure 3 and Supplementary Table 14). Overall, these quality control metrics indicate that our *A. oxyphylla* genome assembly is complete, precise, and high quality.

In this study, a combination of homology-based searches and *de novo* annotation was employed to identify repeat sequences in *A. oxyphylla*. The total length of these sequences was found to be approximately 1.83Gb (1,836,775,398), accounting for 88.06% of the whole genome. It was observed that a large proportion of these sequences were transposable elements (TE), which constituted 87.82% of the entire genome (Figure 1B). The most abundant class of TE was found to be the long terminal repeats (LTR), which accounted for 61.70% of the genome. Protein-coding gene models were predicted through a combination of ab initio prediction, incorporating transcriptome, and homology (Table 1 and Supplementary Table 15). A total of 38,178 protein-coding genes were predicted in *A. oxyphylla*, with an averaged gene length and CDS length of approximately 5.60Kb and 1.12Kb, respectively. Of these genes, 4,982 had homologous support, 1,229 were supported by RNA-Seq, and 593 stemmed from *de novo* gene predictions (Supplementary Figure 5 and Supplementary Table 16). Then, protein sequences were predicted based on gene structure with known protein libraries, such as Swissprot, Nr, InterPro, Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO), and Pfam, with 75.4%, 93.9%, 54.6%, 73.7%, 81.1%, and 73.3% of these genes being functionally assigned, respectively. A total of 35,917 genes were annotated, while 2,261 genes remained unannotated. Of all the genes, 96.08% were assigned to 24 chromosomes, with a total GC content of 39.28%. These genes were unevenly distributed along the chromosomes. Most of them were focused on both ends of the chromosome, and the repetitive sequence appeared to be complementary and centromere focused (Figure 1B and Supplementary Table 17). The study also identified 534 micro RNAs (miRNAs), 3,928 tRNAs, 10,423 small nuclear RNAs (snRNAs), and 9,249 rRNAs in the *A. oxyphylla* genome (Supplementary Table 18).

TABLE 1 Statistics for the *A. oxyphylla* genome and gene prediction.

Assembly features	Size or number
Estimate of genome size (flow cytometry)	1.831Gb
Estimate of genome size (survey)	2.144Gb
Assembled genome size	2.08Gb
Total length of contigs	2,085,722,329bp
Total number of contigs	572
Contig N50	76,960,141bp
Total length of scaffolds	2,085,726,029bp
Scaffolds N50	83,048,840bp
GC content	39.69%
Complete BUSCO	94.6%
Annotation features	**Size or number**
Number of protein-coding genes	38,178
Long terminal repeat (LTR) density	61.7%
Total repetitive sequence	1,836,775,398bp
Rate of repetitive sequence	88.06%

Phylogenetic evolution and whole-genome duplication

In order to investigate the evolutionary status of *A. oxyphylla*, we conducted a comparative analysis of the available genomes of nine angiosperm species (Figure 3A). All species analyzed shared a total of 33,536 gene families, with 6,132 gene families being common across all species, while 440 gene families (equivalent to 1,747 genes) were unique to *A. oxyphylla* (Supplementary Figure 6). Using orthologs alignment of 832 single-copy gene families acquired in *A. oxyphylla* and nine other species, we constructed a phylogenetic tree (Figures 3A, B). The findings were consistent with the current understanding of the relationships among the ten species (Li et al., 2021a; Liao et al., 2022; Yang et al., 2022). This indicated that *A. oxyphylla* was firstly grouped with *W. villosa*, and these two genera were considered as a sister monophyletic group (Li et al., 2020a). *Z. officinale* and *C. alismatifolia* were the closest relatives, forming a parallel group that belonged to Zingiberaceae. The split time of *A. oxyphylla* and *W. villosa* was estimated at 13.7 (2.6-23) million years ago (Mya), while that of *Z. officinale* and *C. alismatifolia* was approximately 16.6 (2.6-23) Mya. The division of these two groups from Zingiberaceae occurred approximately 21.7 (2.6-23) Mya. In addition, based on the known divergence times of eudicots, monocots, Bromeliaceae, and Gramineae, we estimated that Zingiberaceae separated from Musaceae around 77 Mya (Li et al., 2021a), as shown in Supplementary Figure 7.

A total of 8,628 collinearity gene pairs were identified (Figure 1B) in our intergenomic analysis, which revealed strong linear relationships among these species of Zingiberaceae, and most of the chromosomes corresponded one to one. For example, 39,400 collinear genes between *A. oxyphylla* and *Z. officinale* were identified, indicating that 51.66% of the *A. oxyphylla* genome is colinear with the *Z. officinale* genome (Figure 3C).

This study aimed to estimate potential whole-genome duplication (WGD) events in the evolutionary history of *A. oxyphylla* by characterizing the distributions of four-fold synonymous third-codon transversion (4DTv) and synonymous substitution rates (Ks) of inter- and intra-*A. oxyphylla* and *Z. officinale*, *C. alismatifolia*, *W. villosa*, and *M. acuminate*. The sharp peak of Ks was about 0.05, and 4DTv was 0.125, in intra-*A. oxyphylla* and *C. alismatifolia*, *W. villosa*, and *Z. officinale*, suggesting their WGD events occurred after the divergence of Musaceae and Zingiberaceae (Ks,~0.75; 4DTv, ~0.25). We also determined that differentiation between *A. oxyphylla* and *W. villosa*.

10.3389/fpls.2023.1161257

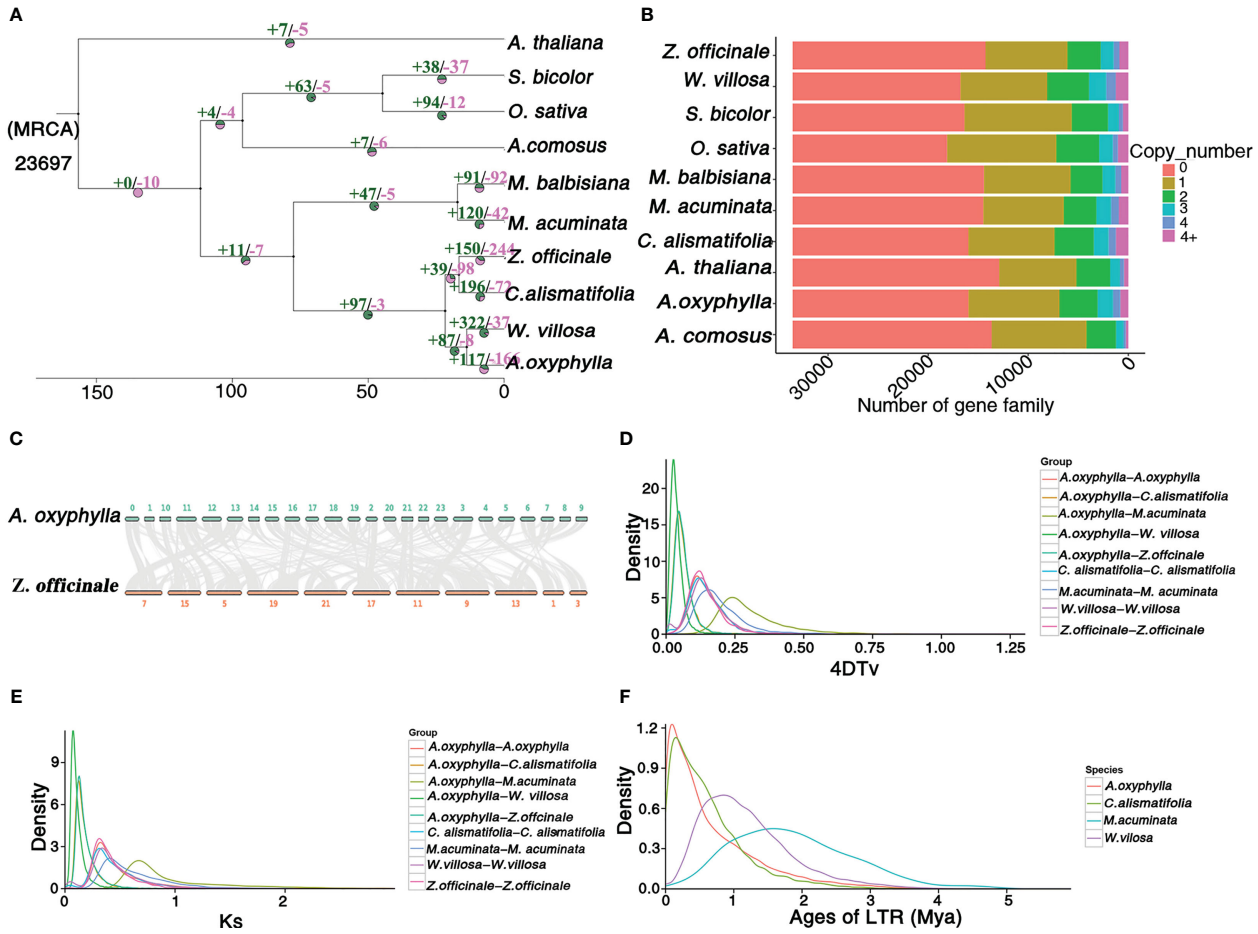


FIGURE 3
Genome and gene family evolution of *A. oxyphylla*. **(A)** Phylogenetic tree of 10 species *A. thaliana*, S. bicolor, O.sativa, A. comosus, M. balbisiana, M. acuminata, Z. officinale, C. alismatifolia, W. villosa, and *A. oxyphylla*. Gene family expansions/contractions are indicated in green/pink. Inferred divergence times (MYA, million years ago) are denoted at each node. **(B)** Gene categories used from all the species (the resource of this figure is from Supplementary Data 1). **(C)** Syntenic blocks between *A. oxyphylla* and Z. officinale. **(D, E)** Distribution of 4DTV values and Ks values of syntenic orthologous genes in the genomes of five species (M. acuminata, Z. officinale, C. alismatifolia, W. villosa, and *A. oxyphylla*. **(F)** Insertion time of LTRs in M. acuminata, C. alismatifolia, W. villosa, and *A. oxyphylla*.

occurred approximately 14(9-19) Mya (Ks,~0.01;4DTv, ~0.025), while speciation of *A. oxyphylla*, *C. alismatifolia*, and *Z. officinale* occurred approximately 22(15-29) Mya (Ks,~0.02;4DTv, ~0.05) (Figures 3D, F). Additionally, we inferred the age of LTRs in four Musales species, finding that *A. oxyphylla* was the first to finish the expansion of LTRs (~0.01 Mya), followed by *C. alismatifolia* (~0.02 Mya), while *W. villosa* and *M. acuminate* expansions occurred much later (~1.00 and 1.50 Mya, respectively) (Figure 3E). In conclusion, the above collinearity analysis of inter- and intra-*A. oxyphylla* genomes helps to confirm the WGD event and LTR amplification involvement in *A. oxyphylla* speciation.

Gene family analysis

Based on the analysis of protein sequences from above 10 species, 30,717 genes were assigned to 15,972 families in *A. oxyphylla*, and 440 gene families, including 1,747 genes, were unique to *A. oxyphylla*, as illustrated in Figure 4A, Supplementary Figure 6, and Supplementary Table 19. To identify the shared and unique gene families, four other Musales species were selected for further analysis. The results indicated that 9,100 gene families were shared among the five species, and 758 unique gene clusters were identified in the *A. oxyphylla* genome (Figure 4A). Additionally, the specific KEGG pathway of *A. oxyphylla* genome was analyzed, and it was found that certain pathways, such as protein export, RNA transport, glutathione metabolism, and vitamin B6 metabolism, were significantly enriched (P<0.05) (Supplementary Figure 8).

After the divergence of *A. oxyphylla*, 117 gene families experienced significant expansion, while another 166 gene families showed significant contraction, as shown in Figure 3A. Our KEGG enrichment analysis indicates that the expanded gene families were involved in biosynthesis pathways such as ribosome, photosynthesis, sesquiterpenoid and triterpenoid biosynthesis, and pyruvate metabolism, relative to the biosynthesis of terpenoids in plastid (Figure 4B). In contrast, the contracted gene families showed

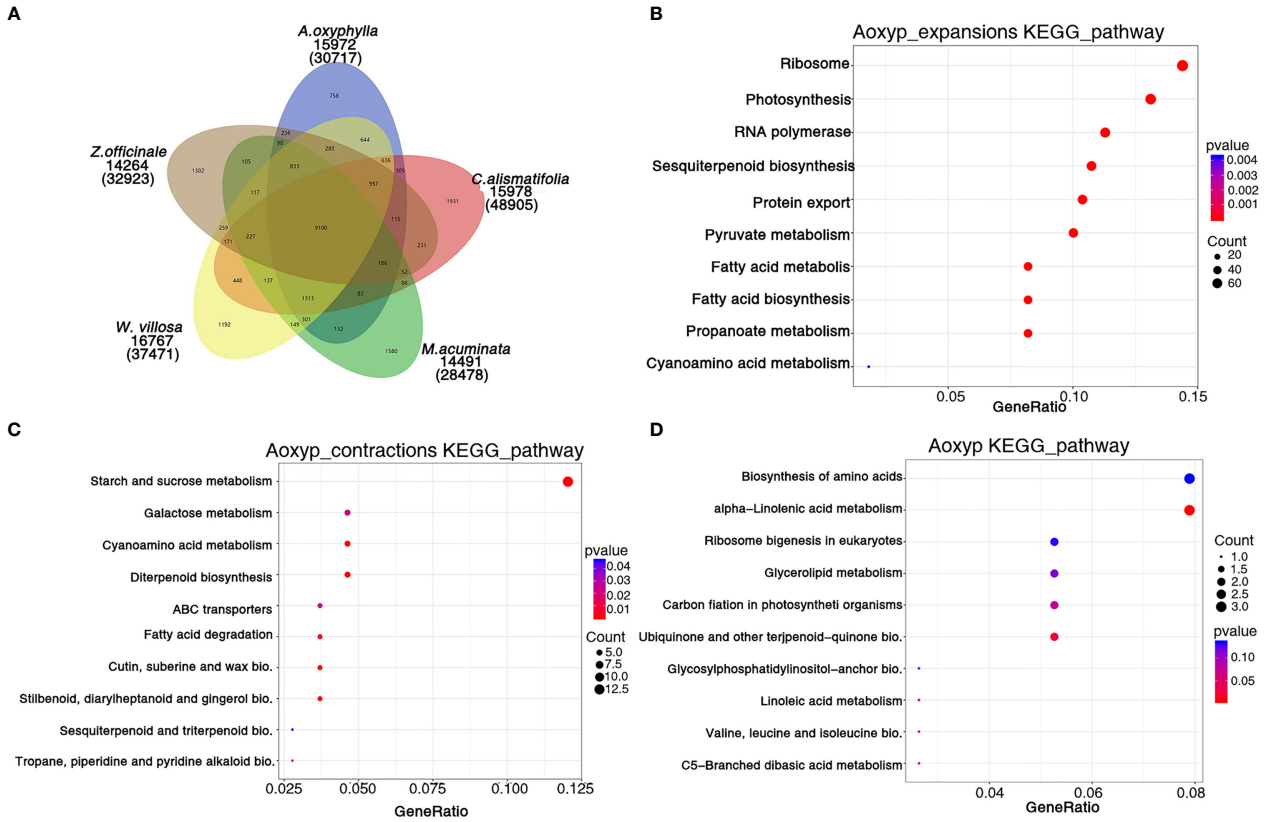


FIGURE 4
Gene family and functional enrichment analysis of *A. oxyphylla*. **(A)** The shared and unique gene families among five species (*M. acuminata, Z. officinale, C. alismatifolia, W. villosa* and *A. oxyphylla*). **(B–D)** KEGG enrichment analysis of expanded gene families, contracted gene families, and positively selected gene families in *A. oxyphylla*.

enrichment in biosynthesis pathways including starch and sucrose metabolism, cyanoamino acid metabolism, diterpenoid biosynthesis, and stilbenoid, diarylheptanoid, and gingerol biosynthesis (Figure 4C). Additionally, our GO analysis revealed that the expanded gene families were mainly enriched in the malonyl-CoA biosynthetic process, acetyl-CoA carboxylase activity, and plastid, as shown in Supplementary Figure 9B. Conversely, the contracted gene families were mainly enriched in the defense response, extracellular region, and ADP binding (Supplementary Figure 9C).

When genes undergo strong positive selection, they play a critical role in generating novel functions within a species. In this study, we analyzed the genes that were subject to positive selection in *A. oxyphylla*. We identified a total of 106 positively selected genes and further performed KEGG analysis to explore their functions. This analysis revealed that several KEGG pathways, such as alpha-linolenic acid metabolism, ubiquinone, and other terpenoid–quinone biosynthesis, were significantly enriched (Figure 4D). GO analysis demonstrated that these positively selected genes were mainly associated with chloroplast organization, chloroplast structure, peptidyl-prolyl cis-trans isomerase activity, and serine-type endopeptidase activity (Supplementary Figure 9D). These positively selected genes, which are specific and expanded, could contribute to the biosynthesis of various secondary metabolites such as volatile terpenoids and flavones.

Differentially expressed genes and characteristic metabolites analyses of *A. oxyphylla* from four different regions

Reverse-phase high-performance liquid chromatography (RP-HPLC) was utilized to analyze the levels of nootkatone, kaempferol, tectochrysin, and six other characteristic metabolites present in *A. oxyphylla* populations from 17 different regions (see Supplementary Table 1). The results of the variance analysis indicated that there were notable differences in four of these regions: Danzhou (A1-A3) and Baoting (A4-A6) received high comprehensive evaluation, while Nanning (B1-B3) and Zhangpu (B4-B6) received low comprehensive evaluation (Supplementary Table 3-6). Consequently, transcriptome and metabolome analyses of the same batch of materials from these four regions were performed with the aim of elucidating significant differences in the genes and metabolites.

A total of 103.49Gb of clean data, with 7.94 9.37 Gb per sample, were collected. On average, 86.32% of reads were mapped to the genome (Supplementary Table 20), resulting in the identification of 3,309 non-communal genes among the four regions (Figure 2A). The KEGG enrichment analysis revealed that 10 pathways, including phenylpropanoid biosynthesis, flavonoid biosynthesis, and stilbenoid, diarylheptanoid, and gingerol biosynthesis, were significantly enriched between A1-A6 and B1-B6 (Figure 2B). Additionally, GO analysis

highlighted significant enrichment in photosystem II, protein polymerization, transferase activity transferring acyl groups other than amino-acyl groups, and terpene synthase activity (Supplementary Data 6). A total of 302 common metabolites were found among the four regions (Figure 2D), which were mainly enriched in ubiquinone and other terpenoid–quinone biosynthesis, fatty acid biosynthesis, and flavone and flavonol biosynthesis pathways (Figure 2E). These pathways allow for insight into the metabolic processes underlying the significant variations in metabolites content among the different regions of *A. oxyphylla*.

In order to unravel the molecular mechanism underlying the variation in *A. oxyphylla* quality across the different regions, we performed an integrated analysis of the transcriptome and metabolome to identify the top 30 hub genes central to the biosynthesis of nootkatone and kaempferol, which are the major metabolites in *A. oxyphylla* fruit. Using the fragments per kilobase of exon model per million mapped fragments (FPKM) data, we created a heatmap visualization to map the distribution of these 30 genes across different regions (Figures 2C, F). Our results showed that the genes involved in nootkatone biosynthesis were highly expressed in the A1-A6 regions, which have higher pharmacodynamic components (Figure 2C). All the annotated genes were functional except for asm_4.69, which was predicted to be an ethylene-responsive transcription factor 3, suggesting its critical regulatory function. Isopentenyl diphosphate delta-isomerase I (IPPI) (asm_23.1717) is the key enzyme of terpenoid synthesis, and it was highly expressed in all regions, representing its essential roles in sesquiterpene biosynthesis. Among the top 30 ranked genes involved in kaempferol biosynthesis, five transcription factors (TFs) (asm_13.1044, 23.348, 18.685, 12.1808, and 4.757) were identified, including MYB98 (MYB98, KAN4, DIVARICATA), KAN4, DIVARICATA, and two bZIP TFs, implying their potential roles in flavonol biosynthesis (Figure 2F).

To further identify the key enzyme genes responsible for the differences in nootkatone and kaempferol content across the four regions of *A. oxyphylla*, we compared and screened the genes involved in terpenoid and flavonols backbone biosynthesis. Our results identified 87 and 35 genes, respectively, in relation to their relevant biosynthesis pathways. The enzymes AACT, DXS, HDS, HDR, and valencene synthase showed higher copy numbers in the nootkatone biosynthesis pathway, while the PAL, C4H, 4CL, and CHS genes exhibited higher copy numbers in the kaempferol biosynthesis pathway (Figure 5), potentially indicating rate-limiting enzymes. However, on examining their expression profiles in the four regions, few genes in the terpenoid and flavonoid backbone biosynthesis pathway were specifically highly expressed, except valencene synthase, which belongs to the downstream genes in the volatile terpenoid biosynthesis pathway. Therefore, we postulate that terpene synthases (TPSs) are likely responsible for the region-specific differences in sesquiterpenoids accumulation observed in *A. oxyphylla* fruit.

Genome-wide detection of TPS genes in *A. oxyphylla*

TPSs play a crucial role in catalyzing GPP, FPP, and GGPP, which produce the skeletons of monoterpenes, sesquiterpenes, and diterpenes, respectively. These enzymes have evolved different-sized subfamilies in various plant species, but typical plant TPSs are a valuable tool to examine plant evolution since they belong to a mid-sized gene family that is conserved more by lineage than by function (Jia et al., 2022). The previous phylogenetic tree of TPS genes from gymnosperms and angiosperms divided the TPSs into seven subfamilies (TPS-a to TPS-g) (Bohlmann J and Croteau, 1998; Dudareva et al., 2003; Martin and Bohlmann, 2004). Among them, TPS-d is specific to gymnosperm (Chen et al., 2011). In this study, we identified 56 putative *AoxTPSs* based on the assembly genome of *A. oxyphylla* and performed phylogenetic analysis to understand their evolutionary relationship with *O.sativa*, *A. thaliana*, and *O. sanctum* (Figure 6A). We observed that 165 TPSs were classified into five subfamilies: TPS-a, TPS-b, TPS-c, TPS-e/f, and TPS-g. Most of the predicted *AoxTPS* genes (46) were clustered into TPS-a (36) and TPS-b (10) subfamilies, suggesting their significant expansion in the genome of *A. oxyphylla* and their possible contribution to mass-producing sesquiterpenoids and monoterpenoids in the fruit. The TPS-c, TPS-e/f, and TPS-g subfamilies in the phylogenetic tree showed 3, 4, and 1 members of *AoxTPSs*, respectively, and the TPS-b and TPS-g subfamilies were mainly responsible for producing monoterpenoids; thus, the TPS-b and TPS-g subfamilies were shown in a combined state with the representation of separate clusters. *AoxTPS* 1, *AoxTPS* 52, and *AoxTPS*53 were not attributed to any subfamilies, indicating that these gene copies originated from dispersed or segmental duplication after species divergence (Li et al., 2022). These 59 *AoxTPSs* were distributed on 13 chromosomes, with 20 genes (*AoxTPS*32-*AoxTPS* 51) located on the chromosome 17 (Figure 6B). Terpenoid biosynthesis genes are generally organized into tandem metabolic gene clusters (Osbourn, 2010; Nützmann and Osbourn, 2014). We found three tandem gene clusters distributed on chromosome 1, 23, and 17, with 10, 2, and 6 *AoxTPS* genes located, respectively, suggesting that tandem duplication events have participated in the expansions of *AoxTPSs*. Further validation is needed to establish whether these expansions contribute to its terpene biosynthesis. We also compared the expression profiles of *AoxTPSs* in four different regions (Figure 6C) and found 14 *AoxTPSs* (consisting of eight genes belonging to the TPS-a subfamily and six genes belonging to the TPS-b subfamily) that exhibited higher transcript abundance in high pharmacodynamic components regions (A1-A6). Some key genes were further validated by qRT-PCR, and the expression level of valencene synthase gene copy *AoxTPS* 34 and *AoxTPS* 50 was consistent with the accumulation of nootkatone from 17 different regions (Supplementary Figure 10). These results above suggest that these genes are important in the different quality formation of *A. oxyphylla* among the 17 regions based on nootkatone content.

Discussion

A. oxyphylla is one of the "Four Famous South Medicines" in China, which significantly contributes to the understory planting

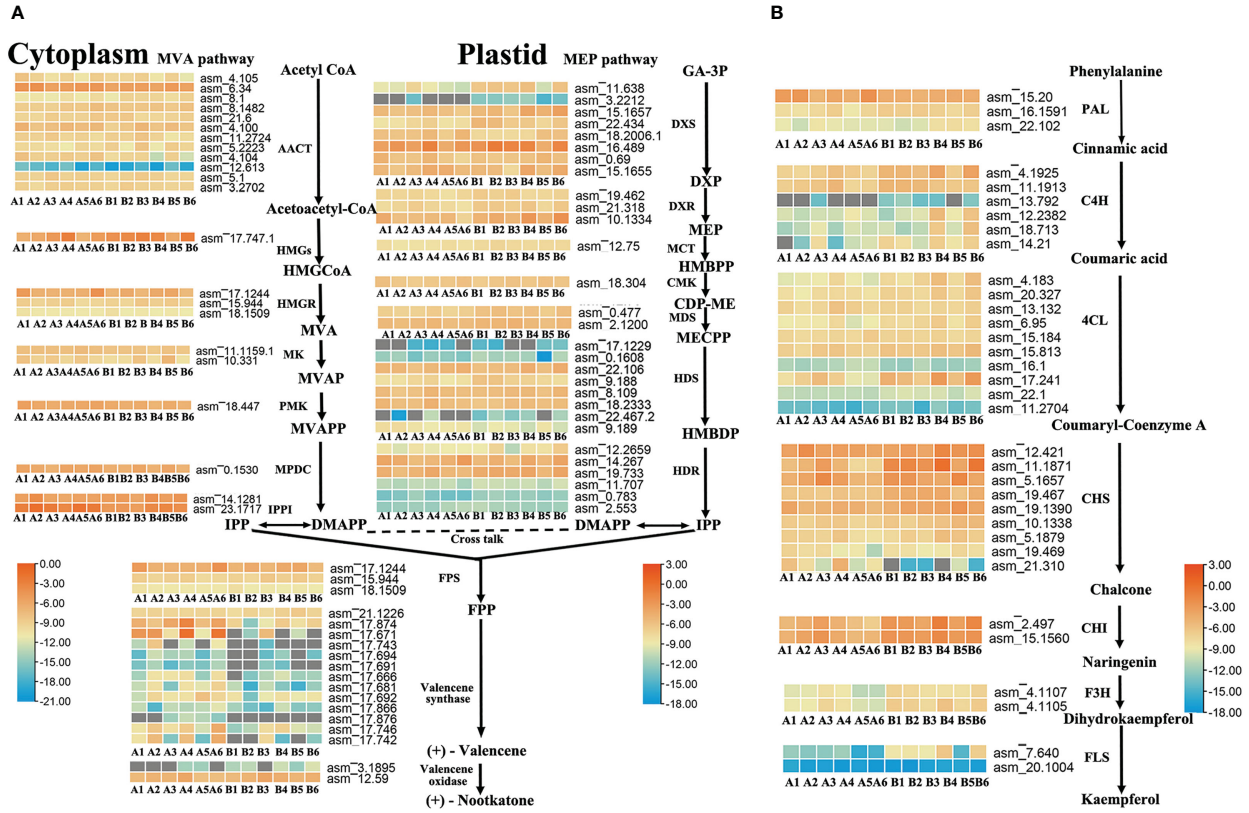


FIGURE 5
Terpenoid and flavonoid metabolic pathways involved in the biosynthesis of nootkatone and kaempferol in *A. oxyphylla*. Heatmap showing the expression level of candidate genes involved in nootkatone **(A)** and kaempferol **(B)** synthesis in different regions. Based on the million mapped fragments (FPKM) value, genes with the identity>70% were selected on the nootkatone and kaempferol biosynthesis pathway; for functional annotation of related genes, see Supplementary Data 4, 5. The FPKM values are log2-based. Red and blue indicate high and low expression levels, respectively. Enzyme abbreviations: MVA pathway: AACT (acetyl-CoA acetyltransferase); HMGS (3-hydroxyl-3-methylglutaryl-CoA synthase); HMGR (3-hydroxy-3-methylglutaryl-CoA reductase); MK (mevalonate kinase); PMK (phosphomevalonate kinase); MPDC (pyrophosphomevalonate diphosphate decarboxylase); IPPI (IDP isomerase); MEP pathway: DXS (1-deoxy-D-xylulose-5-phosphate synthase); DXR (1-deoxy-D-xylulose-5-phosphate reductoisomerase); MCT (2-C-methyl-D-erythritol-4-phosphate cytidylyltransferase); CMK (4-(cytidine 5-diphospho)-2-C-methylerythritol kinase); MDS (2-C-methul-D-erytyritol 2,4-cyclodiphosphate synthase); HDS (4-hydroxy-3-methylbut-2-enyl-diphosphate synthase); HDR (4-hydroxy-3-methylbut-2-enyl-diphosphate reductase); PAL (phenylalanine ammonia-lyase); C4H (cinnamate-4-hydroxylase); 4CL (4-coumarate CoA ligase); CHS (chalcone synthase); CHI (chalcone isomerase); F3H (flavanone 3-hydroxylase); FLS (flavonol synthase). Compound abbreviations: HMGCoA (3-hydroxyl-3-methyglutaryl-CoA); MVA (mevalonate); MVAP (mevalonate-5-phosphate); MVAPP (mevalonate-5-diphosphate); IPP (isopentenyl diphosphate); DMAPP (dimethylallyl diphosphate); GA-3P (D-Glyceraldehyde 3-phosphate); DXP (1-Deoxy-D-xylulose 5-phosphate); MEP (2-C-Methyl-D-erythritol 4-phosphate); HMBPP (4-(Cytidine 5′-diphospho)-2-C-methyl-D-erythritol); CDP-ME (2-Phospho-4-(Cytidine 5′-diphospho)-2-C-methyl-D-erythritol); MECPP (2-C-Methyl-D-erythritol 2,4-cyclodiphosphate); HMBDP (4-Hydroxy-3-methylbut-2-enyl-diphosphate); FPP (farnesyl diphosphate).

economy. The genome, metabolome, and transcriptome data collected for *A. oxyphylla* constitute essential genetic, genomic, and transcriptome resources that can be utilized in future research to comprehend its evolution, biosynthesis of pharmacodynamics components, and quality difference formation. These resources hold significance for studying other species of the Zingiberaceae family and have economic, ecological, and research value. The basic number and ploidy level of Zingiberaceous species are various. In this study, we report the somatic chromosome number of *A. oxyphylla* as 2n=48, which agrees with the previous cytological study by Saenprom et al. (2018). However, our cytomorphological and genome survey results indicate that, contrary to previous beliefs (Eksomtramage and Boontum, 1995; Saenprom et al., 2018), *A. oxyphylla* is a diploid rather than tetraploid as it belongs to the Alpinia genus. Our analysis involved a combination of PacBio and Hi-C technology, which resulted in the assembly of a 2.08 Gb chromosome-scale genome with a contig N50 of 76.96 Mb and scaffold N50 of 83.04 Mb. Moreover, 96.08% of contigs were anchored to the 24 chromosomes. The quality of this assembly is superior to that of recently published species from the Zingiberaceae family, including *W. villosa* (contig N50 of 9.13 Mb) (Yang et al., 2022), *Z. officinale* (contig N50 of 12.68 Mb) (Cheng et al., 2021b), *A. tsao-ko* (contig N50 of 4.8 Mb) (Li et al., 2022), and *C. alismatifolia* (N50 of 57.51Mb) (Liao et al., 2022). Additionally, we annotated 38,178 genes, which is more than *Z. officinale* (36,503) but less than *C. alismatifolia* (57,534) (Liao et al., 2022) and *W. villosa* (42,588) (Yang et al., 2022). The divergence between *A. oxyphylla* and *W. villosa* was approximately 13.7 Mya, while the speciation of *C. alismatifolia* and *Z. officinale* occurred around 16.9 Mya, which is earlier than the estimation of Liao et al. (2022), who proposed that *C.*

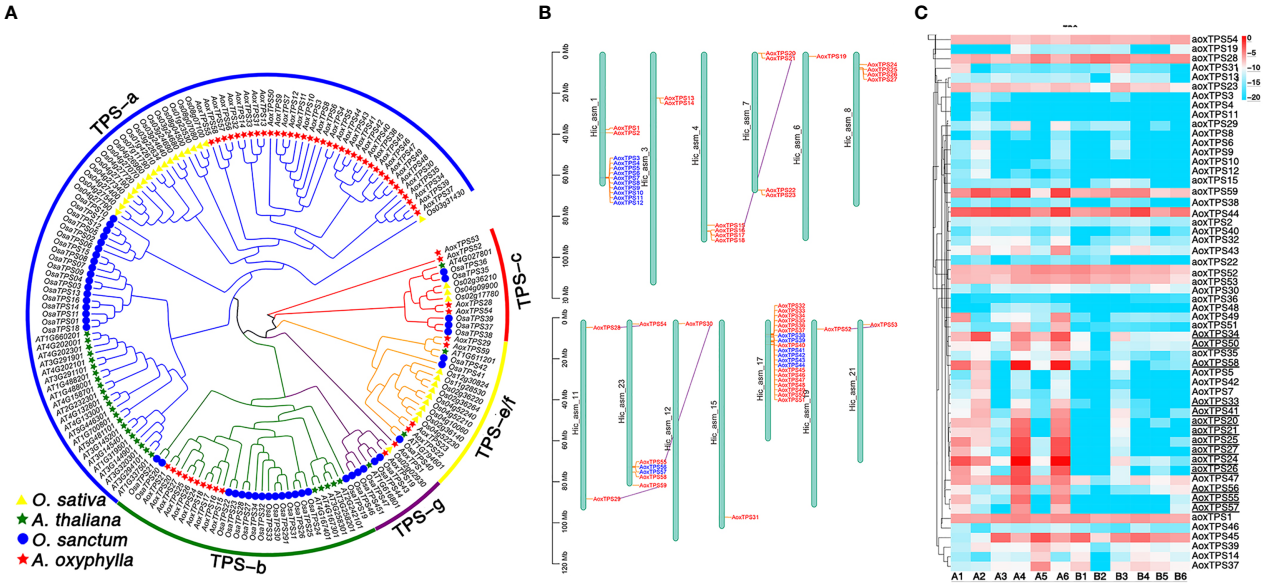


FIGURE 6
Analysis of TPS gene family in *A. oxyphylla*. **(A)** Phylogenetic tree of TPS genes from *A. oxyphylla* (59 genes), *O. sativa* (32 genes), *A. thaliana* (31 genes), and *O. sanctum* (47 genes). The outer circle and branch colors represent different TPS gene subfamilies. **(B)** Schematic map presentation of the genomic localization of 59 *AoxTPSs*. Blue represents tandem duplicates, and purple-linked genes represent genome-wide replication. **(C)** Heatmap (row scale) showing the differentially expressed of *AoxTPSs* according to the transcriptome data in four regions. Uppercase AOX represents *AoxTPSs* from a or b subfamily, and lowercase aox represents *AoxTPSs* from other subfamilies. Underlined *AoxTPSs* represents highly expressed in A1-A6 areas. Sample class: H1 (A1-A3): Danzhou; H2(A4-A6): Baoting; L1 (B1-B3): Nanning; L2(B4-B6): Zhangpu.

alismatifolia and *Z. officinale* diverged approximately 11.9 Mya. The distributions of Ks and substitution rate of 4DTv suggest that a recent WGD event was shared by *A. oxyphylla*, *C. alismatifolia*, *W. villosa*, and *Z. officinale*. This observation corroborates the recent WGD reported in other species (Li et al., 2021a; Liao et al., 2022; Yang et al., 2022) and provides further evidence of the shared WGD in the Zingiberaceae family (Cheng et al., 2021b).

The main pharmacodynamic components of FAO are terpenoids and flavonoids, with sesquiterpene nootkatone and flavonal kaempferol having the highest content, respectively. Nootkatone is predominantly found in the seeds, which is the traditional medicinal part, while kaempferol is mainly deposited in the capsules (Chen et al., 2014). Compared to other Zingiberaceae species such as *W. villosa*, *A. oxyphylla* has experienced significant expansion in 117 gene families. KEGG terms related to pyruvate metabolism and sesquiterpenoid and triterpenoid biosynthesis were significantly enriched in these gene families, and GO analysis showed that expanded gene families were mainly enriched in the malonyl-CoA biosynthetic process, acetyl-CoA carboxylase activity, and plastid. This suggests that *A. oxyphylla* has accumulated genes involved in terpenoid and flavonoids synthesis in recent evolutionary history.

Geo-herbalism is an important index reflecting the quality of traditional Chinese medicine as it is mainly influenced by heredity factors and the environment. One crucial element in geo-herbalism is the content of medicinal ingredients. Accordingly, this study analyzed the gene expression differences in the nootkatone and kaempferol biosynthesis pathway among samples from four different *A. oxyphylla*-growing regions. The results indicated that ethylene-responsive transcription factor IPPI exhibited a critical regulation function in sesquiterpene biosynthesis. Ethylene-responsive transcription factors also play an important role in various abiotic stresses, and they can induce terpenoid synthesis (*OsTPS33, OsTPS14, OsTPS3*) in *O. sativa* in a drought stress environment (Jung et al., 2021) and accelerate the metabolic flux of tanshinone (a type of diterpene) accumulation in *S. miltiorrhiza* (Bai et al., 2018). The interconversion of isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) is mediated by IPPI, which is the only enzyme shared by the mevalonic acid (MVA) and methylerythritol phosphate (MEP) pathways. Many plants contain two IPPI isoforms with different expression profiles encoding proteins and subcellular locations. It has been reported that OsIPPI1 is predominantly responsible for the synthesis of MVA pathway-derived terpenoids, while OsIPPI2 is responsible for the synthesis of MEP pathway-derived terpenoids, such as chlorophylls and carotenoids (Jin et al., 2020). Like many other secondary metabolites, flavonoids play important roles in the interaction of plants with their environment. Additionally, three MYB (MYB98, KAN4, DIVARICATA) and two bZIP TFs were found to be closely related to flavonols biosynthesis in *A. oxyphylla*. For example, KAN4, which belongs to the MYB family, has been previously shown to regulate the biosynthesis of flavonols in Arabidopsis seeds (Gao et al., 2010). Similarly, another MYB gene, *Sm*MYB98, can activate the transcription of the *SmGGPPS1, SmPAL1, and SmRAS1* genes and play a positive regulatory role in the synthesis of tanshinone in *S. miltiorrhiza* (Hao et al., 2020). bZIP was found to focus on the regulation of genes in the upstream synthesis of phenylalanine but inhibit the formation of flavones (flavonol synthase) (Mei et al., 2021). Hence, these genes or TFs may

contribute to the biosynthesis of sesquiterpenoids and flavonols in *A. oxyphylla*.

The present study investigated the genes that contribute to the biosynthesis of kaempferol and nootkatone in *A. oxyphylla*. Specifically, the study focused on PAL, C4H, 4CL, CHS, AACT, DXS, HDS, HDR, and valencene synthase, which were found to have expanded gene families in *A. oxyphylla*. PAL catalyzes the first step of flavonoid biosynthesis pathway and was shown in a previously conducted study to be significantly upregulated as the *A. oxyphylla* fruit matures (Pan et al., 2022). Among these genes, C4H, 4CL, and CHS are involved in regulating other primary steps of flavonoid biosynthesis and show differential expression in different tissues of *A. oxyphylla* (Yuan et al., 2021). AACT, DXS, HDS, and HDR encode key enzymes of terpenoid backbone biosynthesis, and AACT is the initiation enzyme of the MVA pathway, which predominantly provides the precursors for the cytosolic biosynthesis of sesquiterpenoids and for terpenoid biosynthesis in mitochondria. DXS, HDS, and HDR serve as rate-limiting or regulatory enzymes in the MEP pathway and are preferably used for the biosynthesis of monoterpenoids, diterpenoids, carotenoids, and other compounds (Tholl, 2015). But the expression of all the aforementioned genes did not show any evident difference between the A and B regions, except valencene synthase, which belongs to the TPS family and showed considerable expansion in *A. oxyphylla*. This coincides with reports on other plant species, such as *O. sanctum* (Kumar et al., 2018), *W. villosa* (Yang et al., 2022), *Z. officinale* (Cheng et al., 2021b), and Citrus grandis 'Tomentosa' (Xian et al., 2022). Previous studies have suggested that the expansion of the TPS-a and TPS-b subfamilies contributes to the diversity and content enrichment of sesquiterpenoids and monoterpenoids, respectively (Chaw et al., 2019; Yang et al., 2022). In the current study, eight AoxTPS genes belonging to the TPS-a or TPS-b subfamilies (AoxTPS34, AoxTPS50, AoxTPS58, AoxTPS33, AoxTPS41, AoxTPS55, AoxTPS56, and AoxTPS57) were highly expressed in most regions with higher pharmacodynamic components. This finding suggests that these genes may play a role in the regionally different quality formation of *A. oxyphylla*. Furthermore, these genes were validated and selected as candidate genes for utilization in molecular breeding.

Conclusion

This study presents a high-quality chromosome-level reference genome of *A. oxyphylla*. We conducted comprehensive genomic, transcriptomic, and metabolic analyses on population materials from four different plant regions. Our findings reveal that materials from the Hainan region contain higher levels of pharmacodynamic components and confirm that their geo-herbalism properties are attributed to the higher content of nootkatone. Furthermore, we identified the valencene synthase gene, which is likely responsible for the efficient nootkatone synthesis ability across different regions. Therefore, these results contribute significantly to the assessment of *A. oxyphylla* quality in production practices and also to functional genomic research and genome-assisted breeding.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Author contributions

KP, SZ and JZ designed the project and contributed to the original concept of the manuscript. SZ performed *de novo* genome assembly and annotation and analyzed all the data. KP collected the material from 17 regions and wrote the manuscript. SD completed the karyotype analysis of *A. oxyphylla* and revised the manuscript. BG analyzed the transcriptome and metabolism data. JL performed the DNA/RNA extraction and RT-qPCR analysis. ML and SL participated in the RP-HPLC analysis. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by the National Natural Science Foundation of China (No. 81660629) and the Hainan Province Science and Technology Special Fund (ZDYF2022XDNY170).

Acknowledgments

We are grateful for the help of the Fujian Agriculture and Forestry University and Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, for the sample collection.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1161257/full#supplementary-material

References

Bai, Z., Li, W., Jia, Y., Yue, Z., Jiao, J., Huang, W., et al. (2018). The ethylene response factor SmERF6 co-regulates the transcription of SmCPS1 and SmKSL1 and is involved in tanshinone biosynthesis in salvia miltiorrhiza hairy roots. *Planta* 248, 243–255. doi: 10.1007/s00425-018-2884-z

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate a practical and powerful a pproach to multiple testing. *J. R. Statist. Soc B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and genomewise. *Genome Res.* 14, 988–995. doi: 10.1101/gr.1865504

Bohlmann J, M.-G. G., and Croteau, R. (1998). Plant terpenoid synthases molecular biology and phylogenetic analysis. *ProcNatl Acad. Sci.* 95, 4126–4133. doi: 10.1073/pnas.95.8.4126

Bondia-Pons, I., Barri, T., Hanhineva, K., Juntunen, K., Dragsted, L. O., Mykkanen, H., et al. (2013). UPLC-QTOF/MS metabolic profiling unveils urinary changes in humans after a whole grain rye versus refined wheat bread intervention. *Mol. Nutr. Food Res.* 57, 412–422. doi: 10.1002/mnfr.201200571

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., and Shendure, J. (2013). Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* 31, 1119–1125. doi: 10.1038/nbt.2727

Chakraborty, A., Mahajan, S., Jaiswal, S. K., and Sharma, V. K. (2021). Genome sequencing of turmeric provides evolutionary insights into its medicinal properties. *Commun. Biol.* 41193. doi: 10.1038/s42003-021-02720-y

Chaw, S. M., Liu, Y. C., Wu, Y. W., Wang, H. Y., Lin, C. I., Wu, C. S., et al. (2019). Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nat. Plants* 5, 63–73. doi: 10.1038/s41477-018-0337-0

Chen, F., Li, H.-L., Tan, Y.-F., Guan, W.-W., Zhang, J.-Q., Li, Y.-H., et al. (2014). Different accumulation profiles of multiple components between pericarp and seed of *Alpinia oxyphylla* capsular fruit as determined by UFLC-MS/MS. *Molecules* 19, 4510–4523. doi: 10.3390/molecules19044510

Chen, L., Ma, S., Yan, H., Wang, L., and Li, J. (2017). Geo-herbalism research of polygalae radix based on element profiles and chemometrics. *Spectrosc. Lett.* 50, 352–357. doi: 10.1080/00387010.2017.1332648

Chen, Z., Ni, W., Yang, C., Zhang, T., Lu, S., Zhao, R., et al. (2018b). Therapeutic effect of *Amomum villosum* on inflammatory bowel disease in rats. *Front. Pharmacol.* 9. doi: 10.3389/fphar.2018.00639

Chen, F., Tholl, D., Bohlmann, J., and Pichersky, E. (2011). The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* 66, 212–229. doi: 10.1111/j.1365-313X.2011.04520.x

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018a). Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021a). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18, 170–175. doi: 10.1038/s41592-020-01056-5

Cheng, S. P., Jia, K. H., Liu, H., Zhang, R. G., Li, Z. C., Zhou, S. S., et al. (2021b). Haplotype-resolved genome assembly and allele-specific gene expression in cultivated ginger. *Hortic. Res.* 8, 188. doi: 10.1038/s41438-021-00599-8

Christenhusz, M. J. M., and Byng, J. W. (2016). The number of known plants species in the world and its annual increase. *Phytotaxa* 261 (3), 201–217. doi: 10.11646/phytotaxa.261.3.1

Dudareva, N., Martin, D., Kish, C. M., Kolosova, N., Gorenstein, N., Faldt, J., et al. (2003). (E)-beta-ocimene and myrcene synthase genes of floral scent biosynthesis in snapdragon: function and expression of three terpene synthase genes of a new terpene synthase subfamily. *Plant Cell* 15, 1227–1241. doi: 10.1105/tpc.011015

Eksomtramage, L., and Boontum, K. (1995). Chromosome counts of zingiberaceae. *Songklanakarin J. Sci. Technol.* 17, 291–297.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinf.* 9, 18. doi: 10.1186/1471-2105-9-18

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238. doi: 10.1186/s13059-019-1832-y

Gao, P., Li, X., Cui, D., Wu, L., Parkin, I., and Gruber, M. Y. (2010). A new dominant arabidopsis transparent testa mutant, sk21-d, and modulation of seed flavonoid biosynthesis by KAN4. *Plant Biotechnol. J.* 8, 979–993. doi: 10.1111/j.1467-7652.2010.00525.x

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9, R7. doi: 10.1186/gb-2008-9-1-r7

Han, M. V., Thomas, G. W., Lugo-Martinez, J., and Hahn, M. W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol. Biol. Evol.* 30, 1987–1997. doi: 10.1093/molbev/mst100

Hao, X., Pu, Z., Cao, G., You, D., Zhou, Y., Deng, C., et al. (2020). Tanshinone and salvianolic acid biosynthesis are regulated by SmMYB98 in salvia miltiorrhiza hairy roots. *J. Adv. Res.* 23, 1–12. doi: 10.1016/j.jare.2020.01.012

He, B., Xu, F., Xiao, F., Yan, T., Wu, B., Bi, K., et al. (2018). Neuroprotective effects of nootkatone from alpiniae oxyphyllae fructus against amyloid-beta-induced cognitive impairment. *Metab. Brain Dis.* 33, 251–259. doi: 10.1007/s11011-017-0154-6

Jia, Q., Brown, R., Kollner, T. G., Fu, J., Chen, X., Wong, G. K., et al. (2022). Origin and early evolution of the plant terpene synthase family. *Proc. Natl. Acad. Sci. U.S.A.* 119, e2100361119. doi: 10.1073/pnas.2100361119

Jin, X., Baysal, C., Gao, L., Medina, V., Drapal, M., Ni, X., et al. (2020). The subcellular localization of two isopentenyl diphosphate isomerases in rice suggests a role for the endoplasmic reticulum in isoprenoid biosynthesis. *Plant Cell Rep.* 39, 119–133. doi: 10.1007/s00299-019-02479-x

Jung, S. E., Bang, S. W., Kim, S. H., Seo, J. S., Yoon, H. B., Kim, Y. S., et al. (2021). Overexpression of OsERF83, a vascular tissue-specific transcription factor gene, confers drought tolerance in rice. *Int. J. Mol. Sci.* 22 (14), 7656. doi: 10.3390/ijms22147656

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110, 462–467. doi: 10.1159/000084979

Katoh, K., Asimenos, G., and Toh, H. (2009). Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.* 537, 39–64. doi: 10.1007/978-1-59745-251-9_3

Kumar, Y., Khan, F., Rastogi, S., and Shasany, A. K. (2018). Genome-wide detection of terpene synthase genes in holy basil (*Ocimum sanctum* l.). *PloS One* 13, e0207097. doi: 10.1371/journal.pone.0207097

Li, P., Bai, G., He, J., Liu, B., Long, J., Morcol, T., et al. (2022). Chromosome-level genome assembly of *Amomum tsao-ko* provides insights into the biosynthesis of flavor compounds. *Hortic. Res.* 9, uhac211. doi: 10.1093/hr/uhac211

Li, J., Du, Q., Li, N., Du, S., and Sun, Z. (2021b). Alpiniae oxyphyllae fructus and alzheimer's disease: an update and current perspective on this traditional chinese medicine. *BioMed. Pharmacother.* 135, 111167. doi: 10.1016/j.biopha.2020.111167

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows wheeler transform. *bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, Y., Kong, D., Fu, Y., Sussman, M. R., and Wu, H. (2020b). The effect of developmental and environmental factors on secondary metabolites in medicinal plants. *Plant Physiol. Biochem.* 148, 80–89. doi: 10.1016/j.plaphy.2020.01.006

Li, Y. H., Tan, Y. F., Wei, N., and Zhang, J. Q. (2016). Diuretic and anti-diuretic bioactivity differences of the seed and shell extracts of *Alpinia oxyphylla* fruit. *Afr J. Tradit Complement Altern. Med.* 13, 25–32. doi: 10.21010/ajtcam.v13i5.4

Li, H. L., Wu, L., Dong, Z., Jiang, Y., Jiang, S., Xing, H., et al. (2021a). Haplotype-resolved genome of diploid ginger (*Zingiber officinale*) and its unique gingerol biosynthetic pathway. *Hortic. Res.* 8, 189. doi: 10.1038/s41438-021-00627-7

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., et al. (2010). *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20, 265–272. doi: 10.1101/gr.097261.109

Li, D. M., Zhu, G. F., Xu, Y. C., Ye, Y. J., and Liu, J. M. (2020a). Complete chloroplast genomes of three medicinal alpinia species: genome organization, comparative analyses and phylogenetic relationships in family zingiberaceae. *Plants (Basel)* 9 (2), 286. doi: 10.3390/plants9020286

Liao, X., Ye, Y., Zhang, X., Peng, D., Hou, M., Fu, G., et al. (2022). The genomic and bulked segregant analysis of *Curcuma alismatifolia* revealed its diverse bract pigmentation. *aBIOTECH* 3, 178–196. doi: 10.1007/s42994-022-00081-6

Ma, Z., Li, S., and Zhang, M. (2010). Light intensity affects growth, photosynthetic capability, and total flavonoid accumulation of anoectochilus plants. *HORTSCIENCE* 45, 863–867. doi: 10.21273/HORTSCI.45.6.863

Ma, S., Zhu, G., Yu, F., Zhu, G., Wang, D., Wang, W., et al. (2018). Effects of manganese on accumulation of glycyrrhizic acid based on material ingredients distribution of glycyrrhiza uralensis. *Ind. Crops Products* 112, 151–159. doi: 10.1016/j.indcrop.2017.09.035

Mao, X., Cai, T., Olyarchuk, J. G., and Wei, L. (2005). Automated genome annotation and pathway identification using the KEGG orthology (KO) as a controlled vocabulary. *Bioinformatics* 21, 3787–3793. doi: 10.1093/bioinformatics/bti430

Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Martin, D. M., and Bohlmann, J. (2004). Identification of *Vitis vinifera* (-)-alpha-terpineol synthase by in silico screening of full-length cDNA ESTs and functional characterization of recombinant terpene synthase. *Phytochemistry* 65, 1223–1229. doi: 10.1016/j.phytochem.2004.03.018

Mei, X., Wan, S., Lin, C., Zhou, C., Hu, L., Deng, C., et al. (2021). Integration of metabolome and transcriptome reveals the relationship of benzenoid-phenylpropanoid pigment and aroma in purple tea flowers. *Front. Plant Sci.* 12, 762330. doi: 10.3389/fpls.2021.762330

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: the protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi: 10.1093/nar/gkaa913

Mulder, N., and Apweiler, R. (2007). InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol. Biol.* 396, 59–70. doi: 10.1007/978-1-59745-515-2_5

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Nützmann, H.-W., and Osbourn, A. (2014). Gene clustering in plant specialized metabolism. *Curr. Opin. Biotechnol.* 26, 91–99. doi: 10.1016/j.copbio.2013.10.009

Osbourn, A. (2010). Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends Genet.* 26, 449–457. doi: 10.1016/j.tig.2010.07.001

Ou, S., and Jiang, N. (2018). LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 176, 1410–1422. doi: 10.1104/pp.17.01310

Pan, K., Yu, X., Wang, S., Hou, J., Luo, Y., and Gao, B. (2022). Dynamic changes of transcriptome and metabolites during ripening of *Alpinia oxyphylla* fruit (AOF). *J. Plant Biol* 65 (6), 445–457. doi: 10.1007/s12374-022-09354-5

Ra Mans, D., Djotaroeno, M., Friperson, P., and Pawirodihardjo, J. (2019). Phytochemical and pharmacological support for the traditional uses of zingiberacea species in suriname - a review of the literature. *Pharmacog J.* 11, 1511–1525. doi: 10.5530/pj.2019.11.232

Ranavat, S., Becher, H., Newman, M. F., Gowda, V., and Twyford, A. D. (2021). A draft genome of the ginger species *Alpinia nigra* and new insights into the genetic basis of flexistyly. *Genes (Basel)* 12 (9), 1297. doi: 10.3390/genes12091297

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS-the european molecular biology open software suite. *Trends Genet.* 16 (6), 276–7. doi: 10.1016/s0168-9525(00)02024-2

Saenprom, K., Saensouk, S., Saensouk, P., and Senakun, C. (2018). Karyomorphological analysis of four species of zingiberaceae from Thailand. *Nucleus* 61, 111–120. doi: 10.1007/s13237-018-0235-x

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., et al. (2015). HiC-pro: an optimized and flexible pipeline for Hi-c data processing. *Genome Biol.* 16, 259. doi: 10.1186/s13059-015-0831-x

Tang, S., Lomsadze, A., and Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* 43, e78. doi: 10.1093/nar/gkv227

Thapa, P., Lee, Y. J., Nguyen, T. T., Piao, D., Lee, H., Han, S., et al. (2021). Eudesmane and eremophilane sesquiterpenes from the fruits of alpinia oxyphylla with protective effects against oxidative stress in adipose-derived mesenchymal stem cells. *Molecules* 26 (6), 1762. doi: 10.3390/molecules26061762

Tholl, D. (2015). Biosynthesis and biological functions of terpenoids in plants. *Adv. Biochem. Eng. Biotechnol.* 148, 63–106. doi: 10.1007/10_2014_295

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Wang, H., Liu, X., Wen, M., Pan, K., Zou, M., Lu, C., et al. (2012). Analysis of the genetic diversity of natural populations ofAlpinia oxyphyllaMiquel using inter-simple sequence repeat markers. *Crop Sci.* 52, 1767–1775. doi: 10.2135/cropsci2011.06.0323

Xian, L., Sahu, S. K., Huang, L., Fan, Y., Lin, J., Su, J., et al. (2022). The draft genome and multi-omics analyses reveal new insights into geo-herbalism properties of *Citrus grandis* 'Tomentosa'. *Plant Sci.* 325, 111489. doi: 10.1016/j.plantsci.2022.111489

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Yang, P., Zhao, H. Y., Wei, J. S., Zhao, Y. Y., Lin, X. J., Su, J., et al. (2022). Chromosome-level genome assembly and functional characterization of terpene synthases provide insights into the volatile terpenoid biosynthesis of *Wurfbainia villosa*. *Plant J* (3), 630–645. doi: 10.1111/tpj.15968

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq accounting for selection bias. *Genome Biol.* 11, R14. doi: 10.1186/gb-2010-11-2-r14

Yu, S. H., Kim, H. J., Jeon, S. Y., Kim, M. R., Lee, B. S., Lee, J. J., et al. (2020). Anti-inflammatory and anti-nociceptive activities of *Alpinia oxyphylla* miquel extracts in animal models. *J. Ethnopharmacol* 260, 112985. doi: 10.1016/j.jep.2020.112985

Yuan, L., Pan, K., Li, Y., Yi, B., and Gao, B. (2021). Comparative transcriptome analysis of *Alpinia oxyphylla* miq. reveals tissue-specific expression of flavonoid biosynthesis genes. *BMC Genom Data* 22, 19. doi: 10.1186/s12863-021-00973-4

Zhang, J., Wang, S., Li, Y., Xu, P., Chen, F., Tan, Y., et al. (2013). Anti-diarrheal constituents of *Alpinia oxyphylla*. *Fitoterapia* 89, 149–156. doi: 10.1016/j.fitote.2013.04.001

Zou, Y., Zou, P., Liu, H., and Liao, J. (2013). Development and characterization of microsatellite markers for alpinia oxyphylla (Zingiberaceae). *Appl. Plant Sci.* 1 (4), apps.1200457. doi: 10.3732/apps.1200457

Zwaenepoel, A., and Van De Peer, Y. (2019). Wgd-simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* 35, 2153–2155. doi: 10.1093/bioinformatics/bty915

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 12 June 2023
DOI 10.3389/fpls.2023.1183481

OPEN ACCESS

EDITED BY
Kai-Hua Jia,
Shandong Academy of Agricultural Sciences, China

REVIEWED BY
Xian-Ge Hu,
Zhejiang Agriculture and Forestry University, China
Yingyue Li,
Beijing Forestry University, China
Juan Guo,
Chinese Academy of Medical Sciences and Peking Union Medical College, China

*CORRESPONDENCE
Baolin Guo
blguo@implad.ac.cn

RECEIVED 10 March 2023
ACCEPTED 11 May 2023
PUBLISHED 12 June 2023

CITATION
Xu C, Liu X, Shen G, Fan X, Zhang Y, Sun C, Suo F and Guo B (2023) Time-series transcriptome provides insights into the gene regulation network involved in the icariin-flavonoid metabolism during the leaf development of *Epimedium pubescens*. *Front. Plant Sci.* 14:1183481. doi: 10.3389/fpls.2023.1183481

Time-series transcriptome provides insights into the gene regulation network involved in the icariin-flavonoid metabolism during the leaf development of *Epimedium pubescens*

Chaoqun Xu, Xiang Liu, Guoan Shen, Xuelan Fan, Yue Zhang, Chao Sun, Fengmei Suo and Baolin Guo*

Key Laboratory of Bioactive Substances and Resources Utilization of Chinese Herbal Medicines, Ministry of Education & National Engineering Laboratory for Breeding of Endangered Medicinal Materials, Institute of Medicinal Plant Development, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China

Herba Epimedii (Epimedium) leaves are rich in prenylated flavonol glycosides (PFGs) with high medicinal value. However, the dynamics and regulatory network of PFG biosynthesis remain largely unclear. Here, we combined metabolite profiling (targeted to PFGs) and a high-temporal-resolution transcriptome to elucidate PFGs' regulatory network in *Epimedium pubescens* and identified key candidate structural genes and transcription factors (TFs) involved in PFG accumulation. Chemical profile analysis revealed that PFG content was quite different between buds and leaves and displayed a continuous decline with leaf development. The structural genes are the determinant reasons, and they are strictly regulated by TFs under temporal cues. We further constructed seven time-ordered gene co-expression networks (TO-GCNs) of PFG biosynthesis genes (including *EpPAL2*, *EpC4H*, *EpCHS2*, *EpCHI2*, *EpF3H*, *EpFLS3*, and *EpPT8*), and three flavonol biosynthesis routines were then predicted. The TFs involved in TO-GCNs were further confirmed by WGCNA analysis. Fourteen hub genes, comprising 5 MYBs, 1 bHLH, 1 WD40, 2 bZIPs, 1 BES1, 1 C2H2, 1 Trihelix, 1 HD-ZIP, and 1 GATA were identified as candidate key TFs. The results were further validated by TF binding site (TFBS) analysis and qRT-PCR. Overall, these findings provide valuable information for understanding the molecular regulatory mechanism of PFGs biosynthesis, enriching the gene resources, which will guide further research on PFG accumulation in *Epimedium*.

KEYWORDS

***Epimedium pubescens*, prenylated flavonol glycosides, time-series transcriptome, gene regulatory network, transcription factor**

Introduction

Epimedium herb (yin-yang-huo), a well-known traditional Chinese medicine (TCM), is recognized as a prominent prenylflavonol glycoside producer with high medicinal value. To date, only *Epimedium* and *Vancouveria* (sister genus of *Epimedium*, distributed in North America) have a high content of these prenylated flavonol glycosides (PFGs). Pharmacological evidence suggested that PFGs are the major active ingredients, and four of these (Epimedin A, Epimedin B, Epimedin C, and icariin) are used as important bioactive markers for quality control (Xie et al., 2010; Ma et al., 2011). PFGs possess superior capability in being neuro- and cardio-protective (Wang et al., 2007) and have also been used for enhancing reproductive function and anti-aging (Kang et al., 2012). In 2022, icaritin, an aglycone of all PFGs, was approved as a new drug to inhibit hepatocellular carcinoma (HCC) initiation and malignant growth (Zhu et al., 2011; Zhao et al., 2015). *Epimedium* species was predicted as an important and promising medicinal plant with broad market demand, but wild resources of medicinal *Epimedium* species have declined dramatically in recent years due to over-harvesting and habitat destruction (Zhang et al., 2009).

Epimedium is a leafy herbal medicinal plant. The information regarding PFG accumulation dynamics and regulation is scarce. The total PFG content, especially at harvest, is usually what is referred to in traditional use. A systematic understanding of PFG content dynamics could explore the genetic mechanisms and guide the harvesting practice. Huang et al. (2015) reported a study on the dynamic changes of PFGs in leaf developmental process of *E. sagittatum* and suggested that the total content (sum of the content of Epimedin A, B, C and icariin) peaked at folded young leaf with erected petiole stage and then sequentially decreased. Epimedin C constituted the main component and showed a similar trend.

To date, the majority of structural genes of PFGs biosynthesis pathway have been identified in *Epimedium* (Zeng et al., 2013a; Zeng et al., 2013b; Huang et al., 2015; Pan et al., 2017; Feng et al., 2018; Feng et al., 2019; Lyu, 2020; Yang X et al., 2020; Liu Y et al., 2021; Wang et al., 2021; Shen et al., 2022). Of these, four flavonoid skeleton genes, *EwPAL*, *Ew4CL1*, *EwCHS1*, and *EwCHI1* (Liu Y et al., 2021), one *EpsFLS* gene (Pan, unpublished), modification enzyme genes including two PTs, *EsPT2*, and *EpPT8* (Wang et al., 2021; Shen et al., 2022), prenylated flavonoid glycosides with a kind of glycosylation at the 7-OH position of the A-ring, *EpsGT8*, *EsGT1*, and *Ep7GT* (Feng et al., 2019; Yang X et al., 2020; Yao et al., 2022a), glycosylation at the 3-OH position of the C-ring, *EpsGT8*, *EsGT1*, *Ep7GT*, *Ek3RT*, and *Eps3RT* (Feng et al., 2018; Lyu, 2020), glycosylation at the 3-*O*-rhamnoside position, *EpF3R2″XylT* (Yao et al., 2022b), and one OMT gene, *EkOMT1* (Zhou J, 2021), have been functionally verified. The available public gene resources can be of great help in exploring the regulation of PFGs biosynthesis.

There have been few studies on TFs that focused on the regulation of the structural genes of flavonoids in *Epimedium*. Some TFs acted in a manner of MYB-bHLH-WD40 (MBW) complex. *EsAN2* (Huang et al., 2016b) and *EsMYBA1* (Huang et al., 2013) were reported to be involved in anthocyanin biosynthesis pathways and significantly enhanced the anthocyanin accumulation. *EsAN2* can significantly upregulate the expression of *CHS*, *CHI*, and *ANS*, while *EsMYBA1* regulated *CHS*, *CHI*, *F3H*, *DFR*, and *ANS*. In addition, *EsMYB7* and *EsMYB10* (Huang et al., 2012) were reported to regulate the PA biosynthesis, *EsTT8* or *EsGL3* (bHLH) and *EsTTG1* (WD40) may be the co-factors (Huang et al., 2015). Another type worked only by MYB, such as *EsMYBF1* (highly homologous with SG7), and positively regulated flavonol accumulation in a leaf-specific manner by strongly activating the expression of *EsF3H* and *EsFLS* (Huang et al., 2016a). *EsMYB12* and *EsMYB1*, belonging to SG4, have been implicated as transcriptional repressors and negatively regulated anthocyanin biosynthesis in all tissues and the biosynthesis of flavonoids in root, respectively (Huang et al., 2012). However, whether more TFs were involved in transcriptional regulation network remain unknown.

Chang et al. (2019) predicted a regulatory cascade of Kranz anatomy development, which is a structure crucial for the high efficiency of photosynthesis in C4 plants by establishing a time-ordered gene co-expression network (TO-GCNs) method that could use 3D (gene expression, condition, and time) time-series transcriptome data. The time order of TF genes in each gene co-expression network (GCN) was assigned by the breadth-first search algorithm initiated from a seed node which is monotonically increased or decreased. TO-GCNs can effectively elucidate relationships among TF *vs* TF and TF *vs* key genes during continuous development stages. Based on TO-GCNs, the potential regulators and cascade regulatory networks related to flower coloring of *Rhododendron simsii* and *Syringa oblata* were predicted (Yang F. et al., 2020; Ma et al., 2022), the UVB- and UVC-induced early physiological stress responses and the molecular mechanism were characterized in *Pinus tabuliformis* (Xu et al., 2021; Xu et al., 2022), and recently, poplar '84 K' to salt treatment at time series was analyzed, and the physiological dynamics and the potential regulatory mechanism were solved (Zhao et al., 2023). Therefore, the application of time-series transcriptome can provide a new insight into the gene regulation network involved in PFGs.

Here, we report a comprehensive high-temporal-resolution investigation of transcriptome and metabolome (targeted to PFGs) of leaves at six development stages in *E. pubescens*. This study highlighted the regulatory mechanism underlying PFGs biosynthesis. We constructed seven TO-GCNs of TFs regulating structural genes in PFG pathways, and a regulation mechanism model was finally proposed. This study provides a road map for understanding the molecular regulatory mechanism of PFGs biosynthesis, which will facilitate further research on PFGs accumulation in *Epimedium*.

Materials and methods

Plant materials

Plant material *E. pubescens* was obtained from cultivation bases, Leshan, Sichuan province (43°50′9.66″N, 81°10′21.73″E), during spring to autumn (from 26 March to 26 August) of 2021. Analysis of the PFG content of the mature leaves of seven plants was conducted (three biological replicates for each plant, each

repetition has six leaves) prior to the experiment, and three individuals with the closest content were used. The leaf width was preliminarily used as the criterion for determining different developmental stages. Leaf width of 0.5 ± 0.2 ~ 5 ± 0.2 cm with increments of 0.5 cm were collected. Samples were collected at 10:00~11:30 am of a sunny day and were respectively categorized into two parts used for PFG extraction and transcriptome. Each was treated with liquid nitrogen immediately after grafting, stored with dry ice, and quickly transported to Beijing for -80°C conservation under ultralow temperature and further used for RNA extraction and chemical component identification. In total, 42 samples were collected, including a terminal bud as well as 13 leaf sampling points, each with three replicates. Thirty-nine samples were finally used for transcriptome and metabolome determination (leaf width of 4.5 cm was not employed). Due to requirements of sequencing library construction and amount of extraction, 6 and 7 samples are missed, respectively. Finally, 33 and 32 samples were used, respectively, for transcriptome and metabolome analysis. Based on the PCA results of expression levels and PFG contents, the developmental stages were divided, and the sample numbers for each stage were reassigned, as detailed in Supplemental Table 1 and Supplemental Table 2.

PFGs identification and quantification

PFGs were extracted with 99.8% methanol and detected by ultra-high-performance liquid chromatography (UHPLC). Briefly, approximately 0.1 g of each sample was extracted using 1 ml of extraction solution by vortexing at 4°C and subsequent sonification in ultrasonic bath (RK100, Bandelin, Berlin, Germany) for 20 min, then the samples were centrifuged at 12,000 rpm for 10 min, and the supernatants were filtered through a 0.22 μm membrane. Chromatographic separations of compounds in methanol extracts were performed using a Waters ACQUITY I-Class UHPLC system coupled with photo-diode array and quadrupole time-of-flight mass spectrometry (UHPLC-PDA-Q-TOF/MSE) (Waters, Manchester, the United Kingdom). UHPLC-Q-TOF/MSE combined with the UNIFI data analysis platform were adopted to identify the PFGs. UHPLC-PDA was used to determine the relative content of PFGs. Chromatographic settings were as follows: the separation medium was performed on a Waters ACQUITY^{TM} HSS T3 C18 column (100 mm × 2.1 mm) with 1.8 μm particle size (Waters, Ireland) at 40°C. The binary gradient elution system consisted of 0.1% formic acid-water (A) and acetonitrile (B) with a flow rate of 0.6 mL/min, and the absorbance was monitored at 270 nm. Separation was achieved using the following gradient: 0~1.5 min (21% B), 1.5~3 min (24% B), 3~4 min (25% B), 4~6.5 min (29% B), 6.5~7 min (32% B), 7~8 min (44% B), 8~9 min (45% B), 9~11 min (46% B) and 11~20 min (95% B). The injection volume was set to 2 μL. The mass spectrometer (MS) conditions were as follows: electronic impact ion source temperature, 110°C; auxiliary gas (N_2) flow rate and temperature, 850 L/h and 450°C, respectively; negative and positive ionization mode were operated, and the capillary voltage was 2.5 and 3 kV, respectively; high and low scanning energy was 30-50 and 4 eV, respectively; the taper hole voltage, 50 V; the scanning range of molecular weight, 100-1,600 Da; and leucineenk ephalin solution was used to correct the accurate mass number.

Masslynx (version: 4.1) was used to analyze the chromatograms and mass spectra. Target compounds were identified by referring to Zhou M. et al. (2021). Peak area was utilized for quantification of all the target compounds. The PFGs standards used were homemade, including Hexandraside F, Epimedin A, Epimedin B, Epimedin C, icariin, 3''''-carbonyl-2''-β-*L*-quinovosyl-icariin, Ikarisoside B, 2''-*O*-rhamnopyranosyl Ikarisoside A, Ikarisoside A, Sagittatoside A, Sagittatoside B, icariside I, 2''-*O*-rhamnopyranosyl icariside II, icariside II and icaritin. The purity is more than 98%. Three independent experiments were performed, and the mean value was used for further analysis. Principal component analysis (PCA) was provided by R package factoextra. Log transformed and normalized PFGs were used as the inputs. Detailed scripts can be seen in Supplemental File 1.

RNA exaction, library construction, and sequencing

Total RNA was extracted using TRIZOL reagent (Invitrogen, Life Technologies, USA) according to the manufacture's protocol. NanoDrop ND 1000 (Nanodrop technologies) was initially used to detect the protein contamination, the ratio of OD260/OD280 was strictly controlled at 1.9~2.1, and then the RNA Integrity Number (RIN) was assessed by Agilent Technologies 2100 bioanalyzer (Agilent, Santa Clara,CA). Only when RIN > 8 and 28S/18S ≥ 0.7 was sequencing performed. 39 sequence libraries were constructed and sequenced on Illumina HiSeq 2500 platform in BioMED (https://www.biomedi.com.cn/).

Transcriptome analysis

Trimmomatic (version: 0.36) was utilized to make quality control, raw reads were trimmed *via* removing adapters, low quality sequences or bases, and contaminations or overrepresented sequences. The clean data were mapped to the *E. pubescens* genome (Shen et al., 2022) by using HISAT2 (Kim et al., 2015), and hisat2-build and hisat2 were employed to build the index and make alignments, respectively. R package Rsubread (Liao et al., 2019) was adopted to perform gene expression quantification. Gene expression levels were calculated and normalized to transcripts per million (TPM) reads. Differentially expressed genes (DEGs) between each stage were identified with DESeq2 (Love et al., 2014). Genes with Benjamini-Yekutieli false discovery rate (FDR) < 0.05 and $|\log_2(\text{fold change})| > 1$ were considered to be DEGs. DEGs were subjected to enrichment analysis through gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) by using the R package clusterProfiler (version: 3.6.0) (Yu G et al., 2012). PCA was provided by R package PCAtools (Blighe and Lun, 2019). Log transformed and normalized gene expression data was used as the inputs. Detailed scripts can be seen in Supplemental File 1.

TF prediction and PFG biosynthetic candidate gene identification

PlantTFDB database (http://planttfdb.gao-lab.org/index.php) and iTAK software (Zheng et al., 2016) were utilized for TF identification. PlantRegMap database (http://plantregmap.gao-lab.org/) was used to identify the WD40 family by homology to *Arabidopsis thaliana*. PFGs biosynthetic candidate genes were retrieved by using blast or the Hidden Markov Model (HMM) method embedded in HMMER (version: 3.0) (http://hmmer.org/). ClustalW2 (Larkin et al., 2007) and IQ-TREE (Minh et al., 2020) were used for sequence alignment and phylogenetic tree construction, and trees were visualized and modified using iTOL (https://itol.embl.de/) (Letunic and Bork, 2019).

TO-GCNs construction

Three major steps were included for TO-GCNs (Chang et al., 2019): 1) Co-expression cutoffs determination, 2) gene co-expression network (GCN) construction, 3) Time-order of TF gene expression determination. TO-GCNs inputs were the expression profile for each expressed gene (with average TPM > 0.5), which consists of four time points (S1~S4). Pearson correlation coefficient (PCC) values for TFs and gene pairs were calculated, and the cutoff of positive co-expression PCC ≥ 0.85 ($p < 0.01$) was determined. TFs amounting to 1,020 and genes amounting to 20,966 above PCC ≥ 0.85 in "C1 + C2 +" GCN were constructed. MFSelector (Wong et al., 2020) was applied to identify the seed genes with ascending or descending monotonic patterns. A TF gene *FAR1* (*Ebr04G048560*) with the strongest ascending monotonic pattern was selected as the initial node to generate all time-ordered levels of nodes in the TO-GCNs by breath-first search algorithm. TO-GCNs were visualized by Cytoscape (version: 3.6.1) (Shannon et al., 2003). Detailed scripts can be seen in Supplemental File 1.

Candidate PFGs genes regulatory network inference and TFBS analysis

TO-GCNs was firstly used to predict the candidate direct regulators, which should be co-expressed at the same level as or at one level earlier than the structural gene. Similarly, the second-, third-, and fourth-order candidate TFs were inferred, respectively. Secondly, TFBS analysis of each regulatory network of PFGs biosynthesis genes were predicted by extracting the 5' upstream 2 Kb sequences and queried against PlantRegMap database. Thirdly, the presence of the TFBS in the promoter region of each network node were further checked.

WGCNA analysis

All expressed genes (with average TPM > 0.5) were applied. Four major steps were included: 1) Co-expression modules were constructed by using the automatic network construction function blockwiseModules with parameters "soft thresholding power = 9", "mergeCutHeight = 0.25", and "minModuleSize = 50"; 2) An adjacency matrix and a subsequently topological overlap matrix (TOM) were constructed and converted, respectively; 3) Eigengene for each module was calculated and was used to correlate to PFG content. The networks were visualized by Cytoscape (version: 3.6.1) (Shannon et al., 2003). Detailed scripts can be seen in Supplemental File 1.

qRT-PCR analysis

The pre-extracted RNA was reverse transcribed into cDNA using a HiScript II Reverse Transcriptase-based two-step qPCR kit (Vazyme Biotech Co. Ltd., Nanjing, China). Nine gene pairs were selected for validation. beta-Actin-1 was selected as the internal reference gene. Primer 5.0 software was used for primer design. The amplification system was constructed using a LineGene 9600 Plus quantitative real-time PCR detection system (Bioer, Hangzhou, China) and placed in CFX Connect (Bio-Rad Laboratories Inc. Hercules, CA, USA). Three technical replicates were used for each gene, and three biological replicates were used for samples of each developmental stage. The relative expression of genes was calculated using the $2^{-\Delta\Delta Ct}$ method. Origin (version: 2019) was used for correlation analysis to verify the credibility of transcriptome.

Results

Characterization and identification of PFGs in leaves of *E. pubescens*

The PFGs were identified as reported previously (Zhou M. et al., 2021). Here, a case for identifying Epimedin A [Supplemental Figure 1, the peak visible at retention time (R_t) of 4.11 min] were illustrated. In high collision energy (CE) of ESI- mode, the peak eluting at R_t = 4.11 produced a fragment ion at *m/z* 367 [M+HCOO-2Glc-Rha]-, in combination with the evidence of a fragment ion at *m/z* 313 [M+H-2Glc-Rha-C_4H_8]+ in high CE of ESI+, the peak eluting at R_t = 4.11 was deduced to be a flavonoid of Type I (Supplemental Figure 2). Then, based on the fragment ion at *m/z* 883 [M + HCOO]- in low CE of ESI- or at *m/z* 839 [M+H]+ in low CE of ESI+, the molecular mass was confirmed. Then the glycosyl chain at the C-3 site was inferred from the presence of *m/z* 839 [M+H]+ and *m/z* 677 [M+H-Glc]+ in low CE of ESI+, and a loss of 162 Da means the glycosyl ligand connected here is a glucose. The glycosyl chain at the C-7 site was evidenced by *m/z* 883 [M + HCOO]- and *m/z* 675 [M +HCOO-Glc]- in low CE of ESI-, and *m/z* 677 [M+H-glc]+ and *m/z* 531 [M+H-Glc-Rha]+ in low CE of ESI+, which suggested one glucose and rhamnose moiety, according to the database matching by UNIFI and literature reported (Zhao et al., 2008).Taken together, the peak eluting at R_t = 4.11 min was tentatively identified as Epimedin A (Supplemental Figure 3C). Detailed mass spectrogram of chromatographic for all identified PFGs can be seen in Supplemental Figure 3. Finally, a database was built based on the identified PFGs by UHPLC-Q-TOF/MS and PDA chemometric data,

wherein only the commonly observed peaks in PDA chemometric data were used as the marker compounds. As a result, 14 types of PFGs were identified (Table 1 and Supplemental Figure 1). Two backbone types of PFGs, belonging to anhydroicaritin (Type I, C-4′ linked methoxy) and demethylanhydroicaritin (Type II, C-4′ linked hydroxyl), were identified in all samplings (Supplemental Figure 2), with ten and four PFGs included, respectively (Table 1). Detailed major aglycone types, sugarmoieties, and substituent groups in leaves of *E. pubescens* are summarized in Table 1.

Division of leaf development stages

The development stages were defined based on the integration of PCA analyses against the metabolome (targeted to PFGs) (Supplemental Table 1) and transcriptome Supplemental Table 2) data, respectively. PCA of metabolome data (Figure 1A) showed that the first two PCs cumulatively accounted for ~70% of the total variance. PC1 revealed a clear separation among samples of 1~4 (leaf width of 0.5~1 cm), 5~9 (leaf width of 1.5~2 cm), and 10~30 (leaf width of 2.5~5 cm). These groupings were arranged in a clear time-series manner (from left to right), but samples of 28~30 (old leaf with highly leathery) displayed a different characteristic with samples of 10~27 (leaf width of 2.5~4 cm with middle leathery) in PC2 (Supplemental Table 1). PCA of transcriptome data (expressed genes, defined as average TPM > 0.5) exhibited a similar trend (Figure 1B). PC1 showed dynamic changes over the time-series (from left to right). Samples of 4~10 (leaf width of 0.5~1 cm) and 9~14 (leaf width of 1.5~2 cm) revealed significantly different characteristics and were distinguished from 15~33 (leaf width of 2.5~5 cm), which showed a different characteristic in PC2. Notably, samples of 1~3 (bud stage) was with significantly different characteristics Supplemental Table 2). To sum up, the leaf development can be segregated into six stages: 1) Stage 0 (S0), bud stage; 2) Stage 1 (S1), leaf width is 0.5~1 cm, with low-degree of leathery; 3) Stage 2 (S2), leaf width is 1.5~2 cm, with low-degree of leathery; 4) Stage 3 (S3), leaf width is 2~4 cm, with low-degree of leathery; 5) Stage 4 (S4), leaf width is 5 cm, with middle-degree of leathery; 6) Stage 5 (S5), leaf width is 5 cm, with high-degree of leathery (Figure 2A).

Dynamic changes of PFGs with leaf development

The dynamic changes of PFGs were further revealed. Firstly, bud stage (S0) showed the exclusivity of chemical composition and content accumulation. Ikarisoside A and Rhamnose-ikarisoside A were observed as marker components, but they were not detected at leaf stages (S1~S4), Epimedin A and B were almost non-existent in bud stage (S0), with a very low proportion in leaf stages and were almost unchanged (4.97~7.03% for Epimedin A, 6.56~9.69% for Epimedin B) with leaf development (S1~S4). In addition, Epimedoside A and Diphylloside B showed higher contents in bud stage (S0), which was much lower at leaf stages (~39.60% in S0 and ~2.0% in S1~S4 for Epimedoside A; ~6.37% in S0 and ~0.70% in S1~S4 for Diphylloside B). Secondly, leaf stages (S1~S4) showed similar performance in chemical compositions but with significant changes in content accumulation. Total PFGs content were almost unchanged from S0 to S1 (~1.30% increased) and peaked at S1 (the highest accumulation) and then decreased rapidly, followed by 11.23%, 37.81% and 49.43% decreases from S1 to S2, S1 to S3, and S1 to S4, respectively. Similarly, Epimedin C showed the largest proportion of S1 (~53.41% of the total PFGs) and with changes similar to total PFG content, with an ~11.87% increase from S0 to S1, followed by continuous decrease, with an ~8.11%, ~16.14%, and ~24.15% decrease from S1 to S2, S1 to S3, and S1 to S4, respectively. However, the content of icariin was gradually increased, with amplification of ~5.76%, ~13.11%, ~18.71%, and ~23.36% in S0 to S1, S1 to S2, S1 to S3, and S1 to S4, respectively (Figure 2B, Table 1 and Supplemental Table 3).

Transcriptome profiles at different development stages of *E. pubescens*

We generated 5.62~9.04 Gb clean bases per library and a total of 923 Mb clean pair-end reads after filtering and removing the adapter sequences. Q20, Q30, and GC content were higher than 97, 93, and 44%, respectively Supplemental Table 4). The clean reads were mapped to the *E. pubescens* reference genome with an average alignment rate of 87.90% Supplemental Table 5), and 21,345 genes were found to be expressed in at least one sample (Supplemental Table 2).

By comparing the overrepresented GO categories among the DEGs, the biological processes of each stage were outlined. Compared with S0, it was notable that S1 showed an upregulation of basic energy metabolism and antioxidant capacity (Supplemental Figure 4 and Supplemental Table 6). Compared with S1, the up-regulated genes at S2 mostly have function in processes relevant to "cell wall organization or biogenesis (GO:0071555, GO:0042546)", indicating a shift to plant protection, and this change lasted until S3. Multiple enzyme-encoding genes, for example, chitinase, laccase, peroxidase and pectin esterase, were involved. Notably, the PFG content revealed a rapid decline between S2 and S3, and further analysis showed that "secondary metabolic process (GO:0019748)" was overrepresented in the S2 *vs* S3 up-regulated gene set. This GO term included 6 genes, and 3 genes (*Ebr03G037830*, *Ebr03G037820*, and *Ebr02G014470*) were Glutathione S-transferase (GST), which may affect the accumulation of PFG content (Supplemental Table 6). Compared with S3, GO terms related to "abscisic acid-activated signaling pathway" emerged at up-regulated gene set of S4 *vs* S3, however, biological processes relevant to "cell wall organization or biogenesis", "lignin catabolic process", and "hormone-mediated signaling pathway", especially "ethylene-mediated signaling pathway" were overrepresented in the down-regulated gene set of S4 *vs* S3 (Supplemental Figure 4 and Supplemental Table 6). This observation indicates that ABA and ethylene signal transduction tend to play a major role in regulating leaf development or functional transition. To test this hypothesis, we further explored the gene expression pattern from S1 to S4, which reflected the largest variation of PFG content. The results of GO (Supplemental Figure 5) and KEGG enrichment (Supplemental

TABLE 1 UHPLC-Q-TOF/MS metabolic fingerprinting of methanol extracts of *E. pubescens* buds (S0) and leaves of five developmental stages (S1~S5).

Peak No.	Rt (min)	Compound	Molecular formula	Calculated mass (m/z)	Fragment ions (m/z)	Aglycone type	R1	R2	S0	S1	S2	S3	S4	S5
1	2.07	Diphyllodside B	$C_{38}H_{48}O_{19}$	807.2701	645.2173, 353.1035 (-)	II	Rha-Rha	Glc	+	+	+	+	+	+
				809.2852	517.1718, 355.1573, 299.0548 (+)									
2	2.19	Epimedoside A	$C_{32}H_{38}O_{15}$	661.2251	499.1683, 353.1010 (-)	II	Rha	Glc	+	+	+	+	+	+
				663.2525	517.1907, 355.1298 (+)									
3	4.11	Epimedin A	$C_{39}H_{50}O_{20}$	837.294	675.2376, 367.1208 (-)	I	Rha(2-1)Glc	Glc	–	+	+	+	+	+
				839.3212	677.2615, 531.2004, 369.1444, 313.0804 (+)									
4	4.32	Epimedin B	$C_{38}H_{48}O_{19}$	853.2689	645.2260, 367.1208 (-)	I	Rha-Xyl	Glc	–	+	+	+	+	+
				809.3088	677.2648, 531.2004, 369.1444, 313.0804 (+)									
5	4.52	Epimedin C	$C_{39}H_{50}O_{19}$	867.3208	659.2418, 513.1812, 367.1208 (-)	I	Rha(2-1) Rha	Glc	+	+	+	+	+	+
				823.3237	677.2615, 531.2004, 369.1240, 313.0804 (+)									
6	4.71	Icariin	$C_{33}H_{40}O_{15}$	721.2438	513.1832, 367.1208 (-)	I	Rha	Glc	+	+	+	+	+	+
				677.2681	531.2062, 369.1469, 313.0827 (+)									
7	5.5	3'''-carbonyl-2''-β-*L*-quinovosyl-icariin	$C_{39}H_{48}O_{19}$	819.2817	657.2256, 513.1832, 367.1208(-)	I	Rha(2-1)Qui	Glc	–	+	+	+	+	+
				821.3079	531.2004, 369.1444, 313.0804 (+)									
8	6.15	Anhydroicaritin-3-*O*-(acetyl)rhamnopyranosyl-xylopyranosyl-7-*O*-glucopyranoside	$C_{40}H_{50}O_{20}$	819.2817	657.2256, 513.1832, 367.1208(-)	I	Rha(OAc) Xyl	Glc	–	+	+	+	+	+
				821.3079	531.2004, 369.1444, 313.0804 (+)									
9	7.36	2''-*O*-rhamnosyl-ikarisoside A	$C_{32}H_{38}O_{14}$	645.2173	352.0936(-)	II	Rha-Rha	H	+	–	–	–	–	–
				647.2309	501.1755,355.1150 (+)									
10	7.42	Anhydroicaritin-3-*O*-(acetyl) rhamnopyranosyl-(acetyl) xylopyranosyl-7-*O*-glucopyranoside or its isomers	$C_{42}H_{52}O_{21}$	937.3149	729.2535, 367.1233 (-)	I	Rha(OAc)-Xyl(OAc)	Glc	–	+	+	+	+	+
				893.3442	719.2871, 531.2092, 369.1493 (+)									

(Continued)

10.3389/fpls.2023.1183481

TABLE 1 Continued

Peak No.	Rt (min)	Compound	Molecular formula	Calculated mass (m/z)	Fragment ions (m/z)	Aglycone type	R1	R2	S0	S1	S2	S3	S4	S5
11	7.46	Anhydroicaritin-3-*O*-(acetyl) rhamnopyranosyl-(acetyl) xylopyranosyl-7-*O*-glucopyranoside or its isomers	$C_{42}H_{52}O_{21}$	937.3149	729.2535, 367.1233 (-)	I	Rha(OAc)-Xyl(OAc)	Glc	-	+	+	+	+	+
				893.3442	719.2871, 531.2092, 369.1493 (+)									
12	7.53	Ikarisoside A	$C_{26}H_{28}O_{10}$	499.1626	353.1018 (-)	II	Rha	H	+	-	-	-	-	-
13	8.26	2''-*O*-rhamnopyranosyl icariside II	$C_{33}H_{40}O_{14}$	659.2418	367.1208, 352.0966 (-)	I	Rha(2-1) Rha	H	+	+	+	+	+	+
				661.2669	515.2066, 369.1444, 313.0804 (+)									
14	8.77	Icariside II	$C_{27}H_{30}O_{10}$	513.1832	366.1153, 351.0904, 323.0949 (-)	I	Rha	H	+	+	+	+	+	+
				515.2066	369.1444, 313.0804 (+)									

Figure 6) of up-regulated gene set of S1 *vs* S4 were in line with the above-mentioned studies (Supplemental Table 6).

Mining of PFGs biosynthetic genes and TFs

PFGs biosynthetic genes and TFs needed to be mined before constructing TO-GCNs. A total of 259 PFGs biosynthetic genes including 7 PALs (*EpPAL1~EpPAL7*) (Xu et al. unpublished), 1 C4H (*EpC4H*), 14 4CLs (*Ep4CL1~Ep4CL14*), 12 CHSs (*EpCHS1~EpCHS12*) (Shen et al. unpublished), 2 CHIs (*EpCHI1~EpCHI2*) (Fan et al. unpublished), 1 F3H (*EpF3H*), 3 FLSs (*EpFLS1~EpFLS3*), 19 PTs (*EpPT1~EpPT19*) (Shen et al., 2022), 183 UGTs (Yao et al., 2022a) and 17 OMTs (*EpOMT1~EpOMT17*) (Shen et al. unpublished) were identified. Through blast results with the activity verification reported genes (Supplemental Table 7), the matching between expression level with PFGs content during leaf development and the mutation analysis of key sites (including substrate-binding site, active site or phosphorylation site) Supplemental Table 8), 9 genes [including *EpPAL2* (*Ebr04G040710*), *EpC4H* (*Ebr01G074580*), *Ep4CL2* (*Ebr04G003020*), *EpCHS2* (*Ebr05G049130*), *EpCHI2* (*Ebr06G004160*), *EpCHIL* (*Ebr01G073610*), *EpF3H* (*Ebr04G062950*), *EpFLS3* (*Ebr04G051790*) and *EpPT8* (*Ebr02G069700*)] were selected as the candidate genes that participated in the PFGs biosynthesis of *E. pubescens* (Figure 3, Supplemental Figure 7, Supplemental Table 8). A total of 2,249 TFs were detected in *E. pubescens* genome, which were classified into 59 families according to the PlantTFDB database (Supplemental Table 9). A total of 1,208 TF genes were expressed (average TPM > 0.5) in S1~S4. WD40, MYB, bHLH, ERF, and C2H2 families accounted for the largest portion, comprising 195, 97, 94, 80, and 52 members, respectively (Supplemental Table 10).

TO-GCNs regulatory network construction

Between any two leaf developmental stages (S1~S4), 20,943 genes (1,208 TFs and 19,735 structural genes) were expressed (average TPM > 0.5). A TF gene *FAR1* (*Ebr04G048560*), expressing in a low level at S1 and monotonically increasing until S4, was selected as the initial node to build a TO-GCNs network. Eight time-series expression levels (L1~L8, nodes > 10) centering on TFs were finally constructed using the suggested positive/negative cutoff values (0.85; -0.61). Finally, 1,124 genes including 1,022 TFs and 102 PFGs biosynthesis genes made up the TO-GCNs specific to PFGs biosynthetic pathway (Figure 2C). These eight levels were corresponded to the average expression levels at the four developmental stages (S1~S4), as shown by the red squares (high expression levels) along the diagonal in the heatmap, which formed the basis for the inference of upstream and downstream genes/metabolites regulatory relationships (Figure 2D).

With regard to the established TO-GCNs, the major PFGs biosynthetic genes were mainly expressed at the earlier stages (S1~S2) (Supplemental Table 11). There were 18 and 31 PFGs biosynthetic genes expressed in S1 and S2, respectively. The co-expression genes of S1 and S2 reflected a higher expression in early time, and a lower expression in lateral time, and this was in line with PFGs content changes. *EpPT8* (*Ebr02G069700*), *EpPAL2*

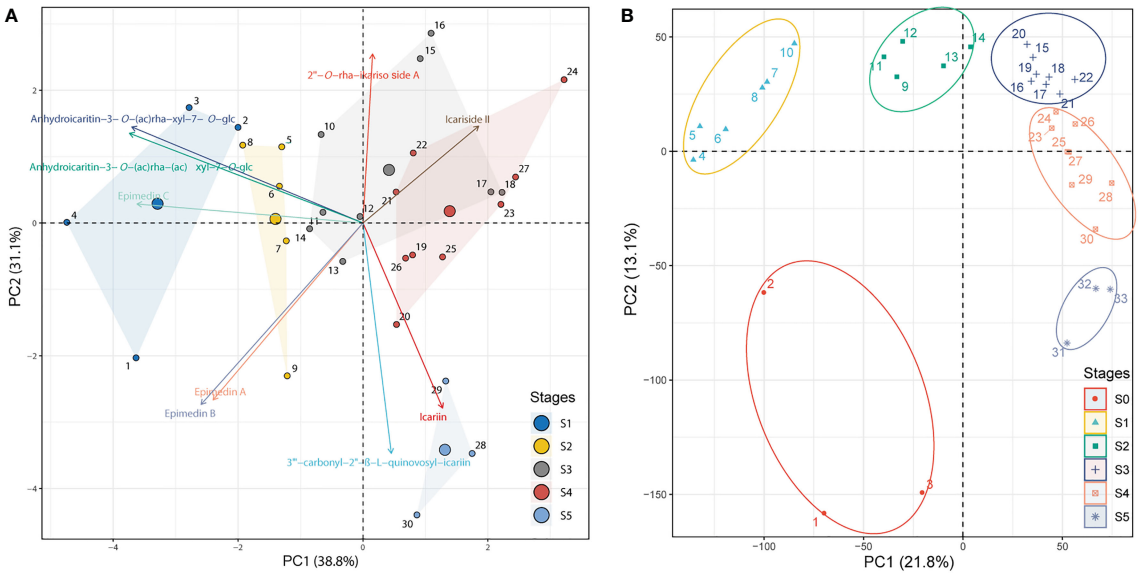


FIGURE 1
PCA of identified PFGs and gene expression levels. **(A)** PCA of the content of identified PFGs in all collected samples from S1~S5. The PCA biplot shows both the PC scores of zero-centered and unit-scaled compound quantity data (the dots represented the sampling individuals) and the loadings of variables (the vectors represented the identified PFGs). Seven individuals are not displayed because the insufficient samplings could not meet the extraction requirements. See Supplemental Table 1 for the numerical symbols. **(B)** PCA of gene expression levels in S0~S5. All expressed genes (average TPM > 0.5) for all collected samples except for six ambiguous individuals were displayed. See Supplemental Table 2 for the numerical symbols.

(*Ebr04G040710*), and *EpFLS3* (*Ebr04G051790*) were presented in L7. *Ep4CL12* (*Ebr04G003020*), *EpC4H* (*Ebr01G074580*), *EpCHS2* (*Ebr05G049130*), *EpCHI2* (*Ebr06G004160*), *EpCHIL* (*Ebr01G073610*), and *EpF3H* (*Ebr04G062950*) were found in L6. Very few causal genes of PFGs biosynthesis could be detected in L1~L4.

Regulatory network prediction of PFG biosynthetic genes

The regulatory network of seven candidate genes (*EpPAL2*, *EpC4H*, *EpCHS2*, *EpCHI2*, *EpF3H*, *EpFLS3*, and *EpPT8*) were predicted, and the TFBS results of each regulatory network are shown in Supplemental Table 12. Take *EpFLS3* for example (Figure 4). We suggest that *EpFLS3* may be regulated in a hierarchical order by three routines: 1) *WRKY* (*Ebr03G071730*) acted as the fourth regulator. It directly regulated the third regulator *MYB* (*Ebr02G010220*), then regulated *MYB* (*Ebr05G056880*), then regulated *MYB* (*Ebr05G057070*), and finally regulated *EpFLS3*. 2) *WRKY* (*Ebr02G071190*) acted as the fourth regulator, either *C3H* (*Ebr06G026230*) or *C2H2* (*Ebr01G055910*) acted as the third regulators, both may regulate *Trihelix* (*Ebr01G020500*), which acted as second regulator and regulated *MYB* (*Ebr05G057070*), and *MYB* (*Ebr05G057070*) served as the direct regulator of *EpFLS3*; on the other hand, *C2H2* (*Ebr01G055910*) could also regulate *MYB* (*Ebr05G057070*) and then regulated *EpFLS3*. 3) *MYB* (*Ebr04G060880*) acted as the second regulator, regulated *MYB* (*Ebr0G003750*), then regulated *EpFLS3*. The routines of 1 and 3 were the most likely regulatory pathways, as *MYB* (*Ebr02G010220*) and *MYB* (*Ebr0G003750*) have been proven to participate in flavonol biosynthesis, which were homologous genes with *A. thaliana* MYB genes *AtMYB111*, *AtMYB11*, and *AtMYB12*. *MYB* (*Ebr02G010220*), *MYB* (*Ebr0G003750*), and *MYB* (*Ebr05G057070*) may act as the core regulated genes for *EpFLS3* regulation.

Both *EpPAL2* and *EpCHS2* showed more complex regulatory networks than other PFGs biosynthetic genes. In brief, the direct regulators of *EpPAL2* with correlation level over 0.8 were six genes except WD40. These genes [*MYB* (*Ebr05G057070*), *C2H2* (*Ebr0G014410*), *HD-ZIP* (*Ebr0G012350*), *MYB* (*Ebr0G003750*), *GATA* (*Ebr02G046010*), and *bHLH* (*Ebr05G004010*)] were further regulated by the second regulators with different correlation levels, and finally, the core regulatory networks of *EpPAL2* were predicted. 1) *WRKY* (*Ebr03G071730*) -> *MYB* (*Ebr02G010220*) -> *MYB* (*Ebr05G056880*) -> *MYB* (*Ebr05G057070*) -> *EpPAL2*. 2) *WRKY* (*Ebr03G071730*) -> *MYB* (*Ebr02G010220*) -> *MYB* (*Ebr05G057060*) -> *MYB* (*Ebr05G057070*) -> *EpPAL2* (Figure 5). Similarly, the most probably regulatory routines of *EpCHS2* were as follows: 1) *WRKY* (*Ebr03G071730*) -> *MYB* (*Ebr02G010220*) -> *MYB* (*Ebr05G056880*) -> *EpCHS2*. 2) *WRKY* (*Ebr03G071730*) -> *MYB* (*Ebr02G010220*) -> *bZIP* (*Ebr05G038380*) -> *EpCHS2*. In addition, *WRKY* (*Ebr03G071730*) -> *MYB* (*Ebr02G010220*) -> *MYB* (*Ebr02G055930*) -> *EpCHS2*, *MYB-related* (*Ebr03G041510*) -> *C3H* (*Ebr05G000930*) -> *MYB* (*Ebr02G055930*) -> *EpCHS2*, *MYB-related* (*Ebr03G041510*) -> *bHLH* (*Ebr06G000830*) -> *MYB* (*Ebr02G055930*) -> *EpCHS2* may also be possible candidate routines, as *MYB* (*Ebr02G055930*) is involved in flavonol biosynthesis (Figure 6). By further analysis of the regulatory network of *EpC4H* (Supplemental Figure 8), *EpCHI2* (Supplemental Figure 9), and *EpF3H* (Supplemental Figure 10), it was found that those genes harbored similar regulatory relationships to *EpPAL2*, *EpCHS2*, and *EpFLS3*. Our research suggests that a set or several sets of TFs,

10.3389/fpls.2023.1183481

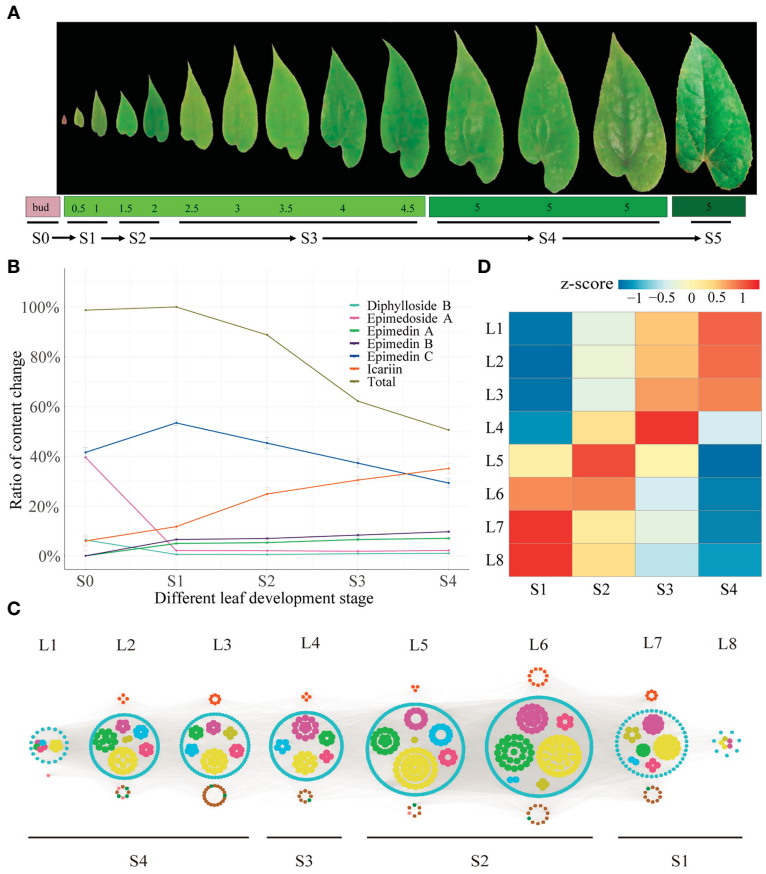


FIGURE 2
Time-ordered gene co-expression network related to leaf development in *E. pubescens*. **(A)** Schematic representation of the design for samplings of *E pubescens*. Leaf width was preliminarily taken as the division of leaf development stages, 42 samples from 14 sampling points, each with three replicates, were collected for metabolic profiling and RNA-seq. Six stages were defined: 1) Stage 0 (S0, bud stage), 2) Stage 1 (S1, leaf width of 0.5~1 cm), 3) Stage 2 (S2, leaf width of 1.5~2 cm), 4) Stage 3 (S3, leaf width of 2~4 cm), 5) Stage 4 (S4, leaf width of 5 cm with middle degree of leathery), and 6) Stage 5 (S5, leaf width of 5 cm with high degree of leathery); **(B)** Proportion of content for the main PFGs in S1~S5 of *E pubescens*. Gray line: total PFGs, blue line: Epimedin C, orange line: icariin, purple line: Epimedin B, green line: Epimedin A, cyan line: Diphylloside B, pink line: Epimedoside A; **(C)** Predicted regulatory network and the connection among TFs and the structural genes involved in PFGs biosynthesis pathway. Inside the cyan circles, purple nodes represented *MYB* genes, green nodes represented *bHLH* genes, yellow nodes represented WD40 genes, brown nodes represented *bZIP* genes, light-blue nodes represented *WRKY* genes, light-green nodes represented *ARF* genes. Outside cyan circles, red nodes, locating on the top, represented *PAL*, *C4H*, *4CL*, *CHS*, *CHI*, *F3H*, and *FLS* genes, and green, brown, and pink nodes on the bottom represented *PT*, *UGT*, and *OMT* genes, respectively; **(D)** The heatmaps of average normalized TPM (z-score) at S1~S4 stages at each level were identified in the time-ordered gene co-expression network. Four stages of leaf with different types of flavonoids accumulation were identified, S1 (L8 and L7), S2 (L6 and L5), S3 (L4), and S4 (L3, L2 and L1) based on the expression profile. The bar represents the expression level of each gene (z-score). Low to high expression is indicated by a change in color from blue to red.

acting in a collaborative manner, regulated the biosynthesis pathway of PFGs.

We further predicted the regulatory network of *EpPT8*, which was the important gene for the formation of active ingredients of *Epimedium*. *MYB* (*Ebr05G057070*), *MYB* (*Ebr0G003750*), *C2H2* (*Ebr0G014410*), *ARF* (*Ebr05G013780*), and *MYB* (*Ebr04G001770*) were predicted to be the direct regulators. *Trihelix* (*Ebr01G020500*), *C2H2* (*Ebr05G010200*), *bZIP* (*Ebr05G038380*), and *TCP* (*Ebr03G011330*) acted as the secondary regulators. The predicted regulatory routines may be as follows: 1) *WRKY* (*Ebr03G071730*) -> *MYB* (*Ebr02G010220*) -> *C2H2* (*Ebr05G010200*) -> *MYB* (*Ebr05G057070*) -> *EpPT8*; 2) *WRKY* (*Ebr03G071730*) -> *MYB* (*Ebr02G010220*) -> *bZIP* (*Ebr05G038380*) -> *MYB* (*Ebr05G057070*) -> *EpPT8*; and 3) *WRKY* (*Ebr03G071730*) -> *MYB* (*Ebr02G010220*) -> *MYB* (*Ebr04G060880*) -> *MYB* (*Ebr0G003750*) -> *EpPT8* (Figure 7).

Exploring TFs involved in PFG accumulation based on WGCNA analysis

WGCNA analysis was employed to construct the co-expression network to further test whether the predicted TFs were involved in PFGs accumulation. A total of 21,345 expressed genes (average TPM > 0.5) were clustered into 18 modules comprising 146~2,240 genes, and each module harbored TFs varying from 3 to 199 (Figure 8A, Supplemental Table 13). Based on the correlation analysis between the module eigengene and the abundance of four PFGs (Epimedin A, Epimedin B, Epimedin C, and icariin) and total PFG content, the blue and brown module was significantly positively and negatively correlated with Epimedin C and total PFGs content, respectively (Figure 8A). We selected blue module (containing 2,240 genes) for further analysis given that most genes relevant to PFGs biosynthesis came from this module.

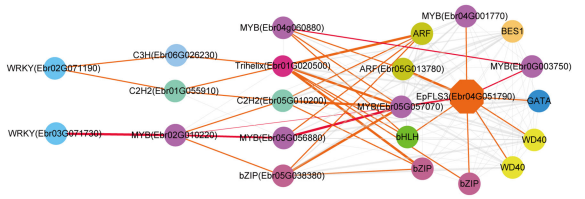


FIGURE 4
Resolved hierarchical regulation for *EpFLS3*. Thickness of the edges in any gene pair represent the strength of co-expression relationships. Thick red and orange edges mean a strong correlation (> 0.8), thin red and orange edges, a middle correlation (0.7~0.8), and gray edges, a weak correlation (0.66~0.7). Red edges represented the most probable pathway related to PFGs biosynthesis.

In the blue module, we firstly extracted a subnetwork comprising of 1,898 genes, and identified 126 TFs and 37 PFGs biosynthesis genes, and then we filtered by TOM value or weight (threshold: 0.24), at which value, the maximum PFGs biosynthesis genes can be retained. The subnetwork (containing 650 genes) was further filtered by threshold of |MM| > 0.8 and |GS| > 0.2 (MM: module membership or eigengene-based connectivity, GS: gene significance). Finally, we completed network construction related to Epimedin C or total PFG accumulation and hub gene identification (Figure 8B).

The core co-expression network contained 52 TFs, which belonged to 18 families, typified by MYB (12 genes), bHLH (6 genes), WD40 (8 genes), and bZIP (5 genes). Hub genes were those with higher intramodular connectivity, which have been visualized as larger circles (Figure 8B), including 5 MYBs, 1 bHLH, 1 WD40, 2 bZIPs, 1 BES1, 1 C2H2, 1 Trihelix, 1 HD-ZIP, and 1 GATA. Further examinations revealed almost all of the TFs existed in the 7 TO-GCNs of PFGs biosynthesis genes (Figures 4–7, Supplemental Figures 8–10), which confirmed that these TFs affect PFG content accumulation by participating in the regulation of target PFGs biosynthesis genes. In addition, almost all of the nine causal PFGs biosynthesis genes were included in this core network (Figure 8B, Supplemental Table 8). This confirmed the reliability of the established TO-CNs network.

Verification of gene regulatory relationships by qRT-PCR

Nine gene pairs including TFs and their predicted target genes were verified by qRT-PCR (see the primers in Supplemental Table 14). These gene pairs included *MYB* (*Ebr05G057070*) and *EpPT8*; *MYB* (*Ebr0G003750*) and *EpPAL2*; *bZIP* (*Ebr05G038380*) and *EpCHS2*; *MYB*

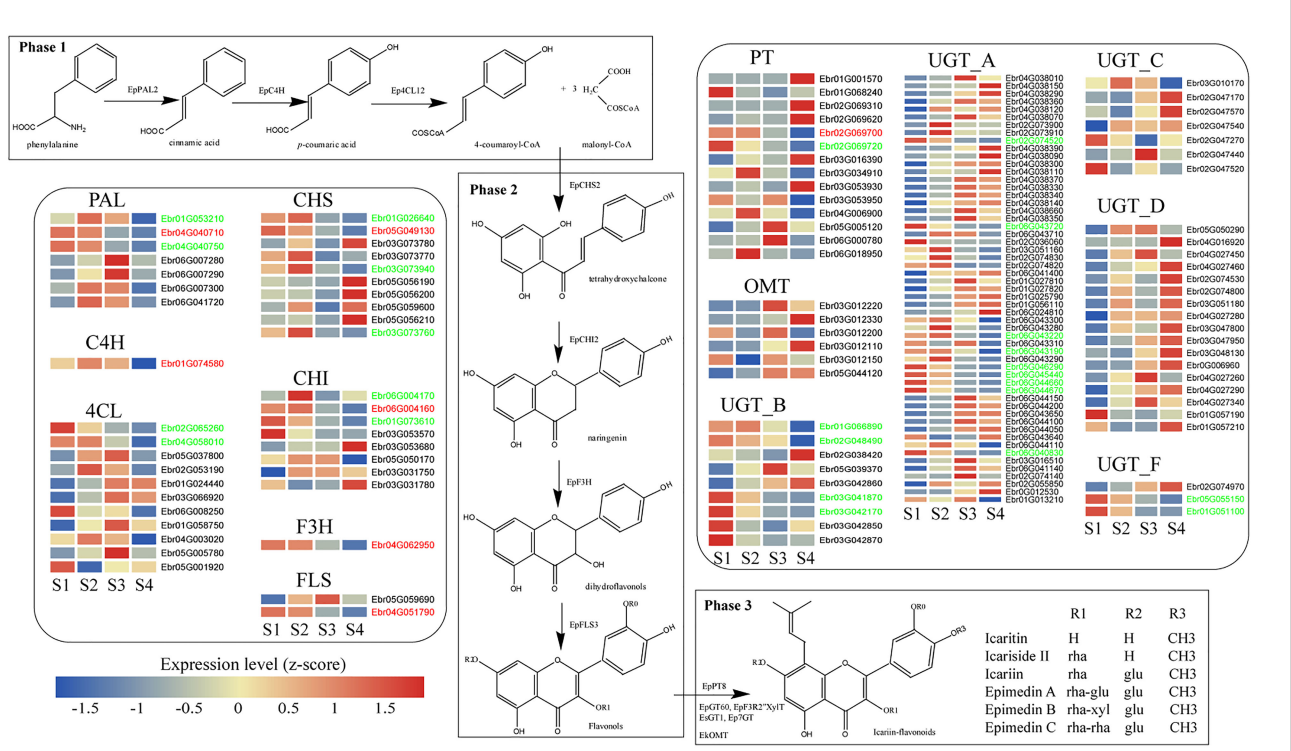


FIGURE 3
The biosynthesis pathway of PFGs in *E. pubescens*. The PFG biosynthesis pathway was divided into three phases. Phase 1: the common phenylpropanoid metabolism, including PAL: phenylalanine ammonia-lyase, C4H: cinnamate-4-hydroxylase, and 4CL: 4-coumarate CoA ligase; Phase 2: core pathway, including CHS: chalcone synthase, CHI: chalcone isomerase, F3H: flavanone 3-hydroxylase, and FLS: flavonol synthase; Phase 3: further enzymatic modification, including PT: prenyltransferase, UGT: UDP-glycosyltransferase, and OMT: *O*-methyltransferase; this study showed genes related to flavonoid biosynthesis (group A, B, C, D and F of UGT gene family, referring to Yao et al., 2022a). Gene expression profiles (in normalized TPMs) in different stages (here S1~S4, from left to right in each heatmap panel) are presented in the heatmap alongside the gene names. Gene names marked in red color represent genes with functional verification, genes marked in green color represent genes consistent with content change during leaf development. The bar represents the expression level of each gene (z-score). Low to high expression is indicated by a change in color from blue to red.

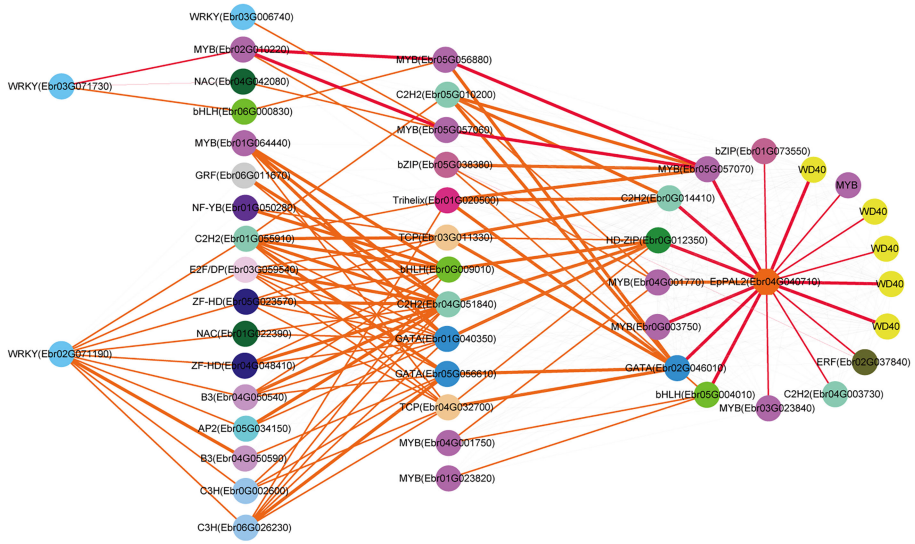


FIGURE 5
Resolved hierarchical regulation for *EpPAL2*. The notion is the same as Figure 4.

(*Ebr0G003750*) and *EpFLS3*; *MYB* (*Ebr02G010220*) and *EpF3H*; *bZIP* (*Ebr05G038380*) and *EpC4H*; *MYB* (*Ebr02G010220*) and *MYB* (*Ebr05G056880*); *WRKY* (*Ebr03G071730*) and *MYB* (*Ebr02G010220*); and *MYB* (*Ebr01G039680*) and *MYB* (*Ebr01G039880*). The results showed that the gene pairs investigated exhibited strong correlations (Pearson correlation coefficient > 0.7), which further verified the reliability of the regulatory network (Figure 9).

Discussion

Epimedium herb has been widely used as important medicine due to its rare content of PFGs, which are known for their outstanding role in inhibiting hepatocellular carcinoma initiation and malignant growth. However, the biosynthesis and regulation mechanism of PFGs have not been systematically summarized and discussed with regard to *Epimedium*. Through high-temporal-resolution transcriptome and metabolome (targeted to PFGs) analysis during early leaf development, the molecular mechanism of PFGs accumulation and its regulation can be unraveled.

Metabolic profiling differences of PFGs between buds and leaves in *Epimedium*

Epimedium species are herbaceous perennials grown from woody rhizomes, in which the meristem of buds triggered the emergence of leaves and flowers. A few studies have conducted the metabolic profiling on leaves, stems, and rhizomes (Xu et al., 2007; Yu J et al., 2012; Zhou M et al., 2021). However, there are no reports on buds. In this study, a significantly different metabolic profiling of PFGs was detected between buds and leaves (Table 1). The main PFG in buds was Epimedoside A (demethylanhydroicaritin backbone, C-4′ linked hydroxyl), which was apparently higher than that in leaves; this is similar to the stem and rhizome reported by Zhou M et al. (2021). Ikarisoside A and 2′-*O*-rhamnosyl-Ikarisoside A (backbone of Type II) were only detected in buds. The main PFGs in leaves were dominant by

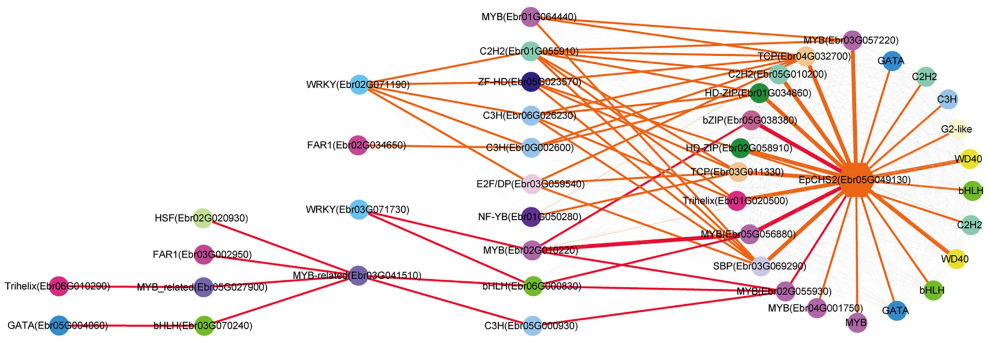


FIGURE 6
Resolved hierarchical regulation for *EpCHS2*. The notion is the same as Figure 4.

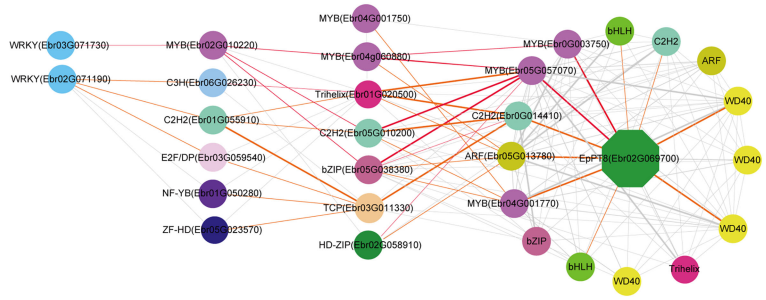


FIGURE 7
Resolved hierarchical regulation for *EpPT8*. The notion is the same as Figure 4.

Epimedin C and icariin (anhydroicaritin backbone, C-4′ linked methoxy) (Table 1, Supplemental Table 1). Different *O*-methylation modification constituted the chemical diversity between buds and leaves. The higher enzyme activity of 4′-OMT in leaves may be the possible molecular basis, which can provide an explanation of the spatial distributions of metabolites and their chemical structures. Similar studies, e.g. the differential distribution of compounds in aerial and underground *Scutellaria baicalensis* (Zhao et al., 2016) and biosynthesis-based spatial metabolome of *Salvia miltiorrhiza*, have been in-depth discussed in depth (Tong et al., 2022).

Possible reasons for the decrease of PFGs contents

Flavonoids have a strong antioxidant capacity and play an important role in promoting growth and development, which function at cellular-level processes, including cell division, membrane integrity, and ROS scavenging (Yadav et al., 2018). The significant metabolic flow transition for different stages of leaf development (S1~S4) (Supplemental Figure 5, Supplemental Table 6) suggested that the declined tendency of PFGs may be closely related to the demand for antioxidant ability and plant protection in young leaves. Studies of flavonoids in leaves of *Amygdalus pedunculata* (He et al., 2021), *Cistus ladanifer* (Valares Masa et al., 2016), and *Ginkgo biloba* (Wang et al., 2022) provide further evidence. In addition to S1, many biological processes towards "cell wall biosynthesis" were enriched in S2~S4, therefore, we deduced that with leaves aged, the increase in cell wall flavonoids may be paralleled by a decrease in soluble flavonoids in *Epimedium*. Since flavonoids are synthesized *via* a multienzyme complex localized in the endoplasmic reticulum (Zhu et al., 2016), the flavonoid transport deliver system in S1 may experience an efficient transport to each membrane-limited compartments, including nucleus

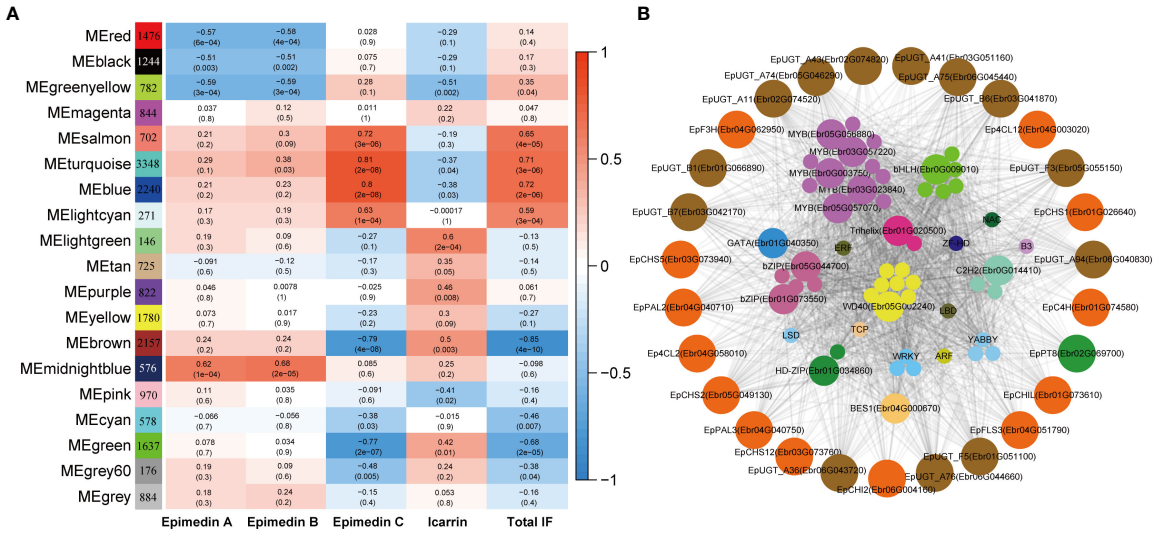


FIGURE 8
WGCNA of all expressed genes. **(A)** Module-PFGs relationship. Each row represented a module, which consists of genes with similar expression pattern. Each column represented the specific PFGs compound content, including Epimedin A, B, C, and icariin, and the sum of all PFGs. The value in each cell at the row-column intersection represents the correlation coefficient between the module and the specific compound content and is displayed according to the color scale on the right. The value in parentheses in each cell represents the *P* value. **(B)** Regulatory network of PFG biosynthesis in *E pubescens*. The outside circle with different colors indicates different families of structural genes associated with Epimedin C or total PFGs biosynthesis in the blue module. The inside circles with different colors indicate different families of TFs characterized in the same module whose transcripts are highly correlated with the expression of structural genes.

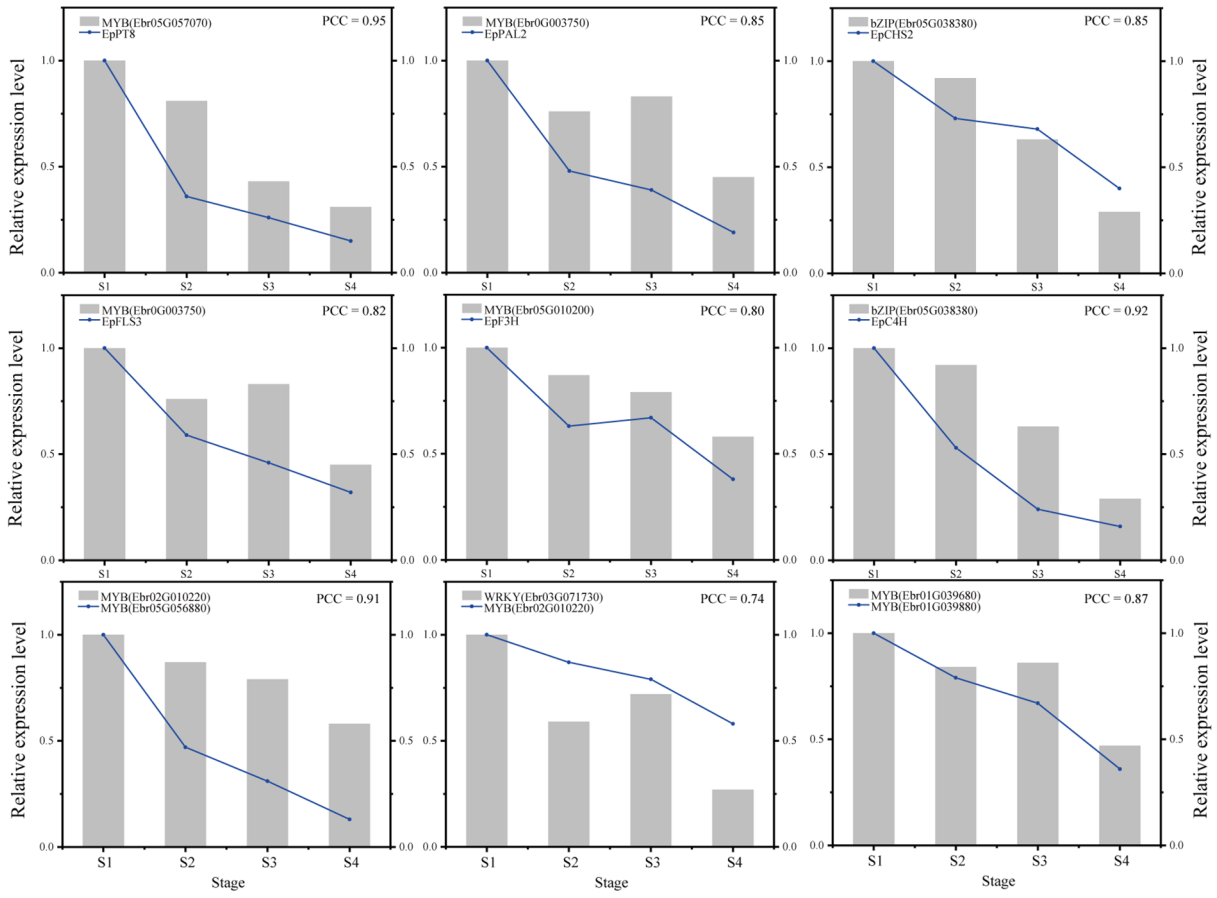


FIGURE 9
Verification of genes expression and regulatory relationships by qRT-PCR.

and chloroplast. S2~S4 may occur vacuolar efflux of these soluble flavonoids and deposited in the cell wall (Zhao and Dixon, 2010). Further experiment with confocal laser scanning microscopy can provide more direct evidence.

PFGs biosynthesis genes constitute the direct reason for PFG contents variation. Transcriptome analysis showed that these genes, including *EpPAL2*, *EpCHS2*, *EpCHI2*, *EpF3H*, *EpFLS3*, and *EpPT8*, had expressions which were down-regulated gradually with leaf maturity (Figure 3), which are in agreement with the metabolomics data, showing the highest PFG levels in S1 *Epimedium* leaves and its decrease during further leaf development. Similar results were observed in leaves of *E. sagitattum* (Huang et al., 2015), *A. pedunculata* (He et al., 2021), *C. ladanifer* (Valares Masa et al., 2016), and *G. biloba* (Wang et al., 2022).

In view of the collaborative expression of the PFGs biosynthetic genes, it could be presumed that these genes have been strictly regulated by TFs under temporal cues. This was in line with previous studies (Stracke et al., 2010; Naik et al., 2022). To explain the regulatory background of *Epimedium* PFGs metabolism during leaf development, we obtained 14 candidate TFs through WGCNA analysis (Figure 8). It has been well established that SG7 MYB TFs contain SG7 motif ([K/R][R/x][R/K]xGRT[S/x][R/G]xx [M/x]K) and SG7-2 motif ([W/x][L/x]LS) at their C-termini, such as *AtMYB11*, *AtMYB12*, and *AtMYB111*, which function was direct activators to affect flavonol biosynthesis (Mehrtens et al., 2005; Stracke et al., 2007). The homologous genes have been investigated in *E. sagittatum* (Huang et al., 2015), grape (Czemmel et al., 2009), *Fagopyrum tataricum* (Yao et al., 2020), *Nicotiana tabacum* (Song et al., 2020), and pear (Zhai et al., 2019). Within the 14 TFs, one MYB *Ebr0G003750*, which was identified as one of the hub genes in co-expression network of Epimedin C or total PFGs, was the homolog of *EsMYBF1* identified in *E. sagittatum* (with 91.86% identity), which may be one of the direct activators involving in the regulating of PFG content. The consistency of the expression pattern of MYB *Ebr0G003750* with metabolomic data further verified this point. Other candidate MYB hub genes like MYB *Ebr05G057070* and *Ebr05G056880* (belonging to SG1, involved in environmental stress), and MYB *Ebr03G057220* (belonging to SG15, involving in epidermal cells) may involve indirect ways. MYB *Ebr03G023840* belongs to SG6 (involving in anthocyanin biosynthesis), and this gene may regulate both anthocyanin and flavonol biosynthesis pathways simultaneously. This type of gene has been reported in *Gerbera hybrida*, which is involved in the regulation of both flavonoids, as they share the same subcellular localization and common biosynthetic substrates, which may compete for substrates (Zhong et al., 2020).

TO-GCN is conductive to discover new TF genes

We present here a TO-GCN approach to provide regulatory dynamics during leaf development. Compared to other time-series analysis method like Mfuzz (Kumar and Futschik, 2007), maSigPro (Conesa et al., 2006), and ImpulseDE2 (Fischer et al., 2018), TO-GCNs could predict upstream regulators of any genes in the GCNs (Chang et al., 2019). Given the important role of TFs as major drivers of genetic variation (Wallace et al., 2014), to understand which TFs control which sets of PFG biosynthesis genes, it is important for the rational metabolic engineering of plants with altered metabolites.

In general, it is much more difficult to predict an upstream regulator than a downstream target one. In this study, we revealed the cascade regulations of seven PFGs biosynthesis genes. Similar examples related to cascade regulation have been reported (Zhou et al., 2015; Li et al., 2020; Zhao et al., 2021). More than 1,022 TFs are assigned to the leaf development TO-GCN, providing a global picture of all gene regulatory relationships, and ~50 TFs were specially extracted through all 7 TO-GCNs related to PFG biosynthesis pathways (Figure 4–7, Supplemental Figures 8-10). This enriched our genetic resources with PFG regulation. In contrast, only one TF (*EsMYBF1*) (Huang et al., 2016a), two TFs (*EsAN2* and *EsMYBA1*) (Huang et al., 2012; Huang et al., 2013), and two TFs (*EsMYB7* and *EsMYB10*) (Huang et al., 2012) were identified as being involved in regulating flavonol, anthocyanin, and PAs biosynthesis in previous studies of *Epimedium*.

The TFs involved in flavonoid regulation have been reviewed (Li, 2014; Xu et al., 2015; Liu W et al., 2021). Besides the deeply studied *MYB11*, *MYB12*, and *MYB111*, which are functionally redundant and control the flavonol biosynthesis *via* activating the early biosynthetic genes such as *CHS*, *CHI*, *F3H*, *F3'H*, and *FLS*, many new TFs have been revealed, including activators *TCP3* (*bHLH*) (Li and Zachgo, 2013), *CsbZIP1* (Zhao et al., 2021), *VvibZIPc22* (Malacarne et al., 2016), and *AtWRKY23* (Grunewald et al., 2012), and repressors *CsPIF3* (Zhao et al., 2021), *FaMYB1* (Aharoni et al., 2001), and *BES1* (Liang et al., 2020). Based on TO-GCNs, we identified some unconfirmed TFs, which may be potential regulatory genes for flavonol synthesis. Li et al. (2021) reported that C2H2 and Trihelix indirectly promoted the synthesis of flavonoids by regulating abscisic acid (ABA) levels. The homolog genes of C2H2-type zinc finger (*Ebr0G014410*) and Trihelix DNA-binding factors (*Ebr01G020500*) were included in the TO-GCN of *EpFLS3*, and both genes positively correlated with the expression of *EpFLS3*, which may promote flavonol synthesis by reducing the inhibition of ABA on *EpFLS3* expression. Similarly, *BES1* (Liang et al., 2020) and *CsbZIP1* (Zhao et al., 2021) were recently reported, and these genes were included in the established TO-GCNs. It is reported that *BES1* served as a positive regulator in brassinosteroid signaling, inhibiting the transcription of *MYB11*, *MYB12*, and *MYB111*, thereby decreasing flavonol biosynthesis (Liang et al., 2020). In the tea plant, UV-B irradiation-mediated *bZIP1* upregulation leads to the promotion of flavonol biosynthesis by binding to the promoters of *MYB12*, *FLS*, and *UGT* and activating their expression (Zhao et al., 2021).

Complex regulation mechanisms of PFGs biosynthesis were revealed in *Epimedium*

In this study, the interactivity of TFs with structural genes were highly complex, especially for genes like *EpPAL2* (49 TFs were predicted) and *EpCHS2* (43 TFs were predicted) (Figures 5, 6), which served as a control point of metabolic flow. Their *cis*-elements, including Silencer, H-box, ACE elements, AT-rich element, Box I and Box II, were regulated by many TFs and environmental factors (Liu et al., 2011; Pant and Huang, 2022). Genes like *EpPAL2*, *EpFLS3*, *EpC4H*, *EpCHS2*, *EpCHI2*, *EpF3H*, and *EpPT8* interacted with 11 or more regulatory partners, and expression of most of these genes was activated at S1~S2 or even before that (Figures 4–7, Supplemental Figures 7–9 and Supplemental Table 11). Our data suggested that gene interactions were at their highest complexity at the initiation of leaf, as has been described in petal color regulation in *R. simsii* (Yang F et al., 2020) and *S. oblata* (Ma et al., 2022).

We found almost all the PFG biosynthesis genes are co-regulated by a set of commonly shared TFs; this was evidenced in *CsMYBF1* (Liu et al., 2016), *GtMYBP3*, and *GtMYBP4* (Nakatsuka et al., 2012). Grotewold (2008) reported that the control of secondary metabolism is often carried out by TFs that are specialized in controlling particular branches of a pathway, often by activating or repressing the expression of a few genes encoding metabolic enzymes. The shared TFs in 7 TO-GCNs related to PFG biosynthesis pathways might the key regulators, including *MYB* (*Ebr02G010200*), *bZIP* (*Ebr05G038380*), *C2H2* (*Ebr05G010200*), *MYB* (*Ebr05G057070*), *MYB* (*Ebr0G003750*), *MYB* (*Ebr02G055930*), *MYB* (*Ebr01G039880*), and *WRKY* (*Ebr03G071730*) (Figures 4–7, Supplemental Figure 7–9). SG7 group MYBs, *Ebr0G003750*, *Ebr02G055930*, and *Ebr02G010200* can serve as the marker genes, which allowed us to find possible candidate pathways controlling flavonol pathway in *Epimedium*.

Based on the aforementioned description, a model is proposed to elucidate the PFG accumulation pattern and how the aforementioned fine-tuners control flavonoid biosynthesis (Figure 10). Although partial important TF and TF pairs or TF and biosynthesis genes relationships have been tested by qRT-PCR (Figure 9), the regulatory relationships between more TFs need to be further confirmed.

Conclusion

By performing a combined analysis of metabolite profiling (targeted to prenylated flavonol glycosides) and a high-temporal-resolution transcriptome analysis in *E. pubescens*, the overall decline pattern of PFGs accumulation was clarified. Based on TO-GCNs, a TF gene homologous to *EsMYBF1* was found, multiple new TFs were predicted, and cascade regulatory networks related to prenylated flavonol glycosides were established. Partial TFs have been validated by WGCNA analysis and qRT-PCR. This is the first time that high-temporal-resolution transcriptome was performed to explore the cascade regulation of structural genes related to active ingredients accumulation in medicinal plants, which provide guidance for further studies on the role of TFs involved in PFGs biosynthesis and breeding programs. Although the TO-GCN results

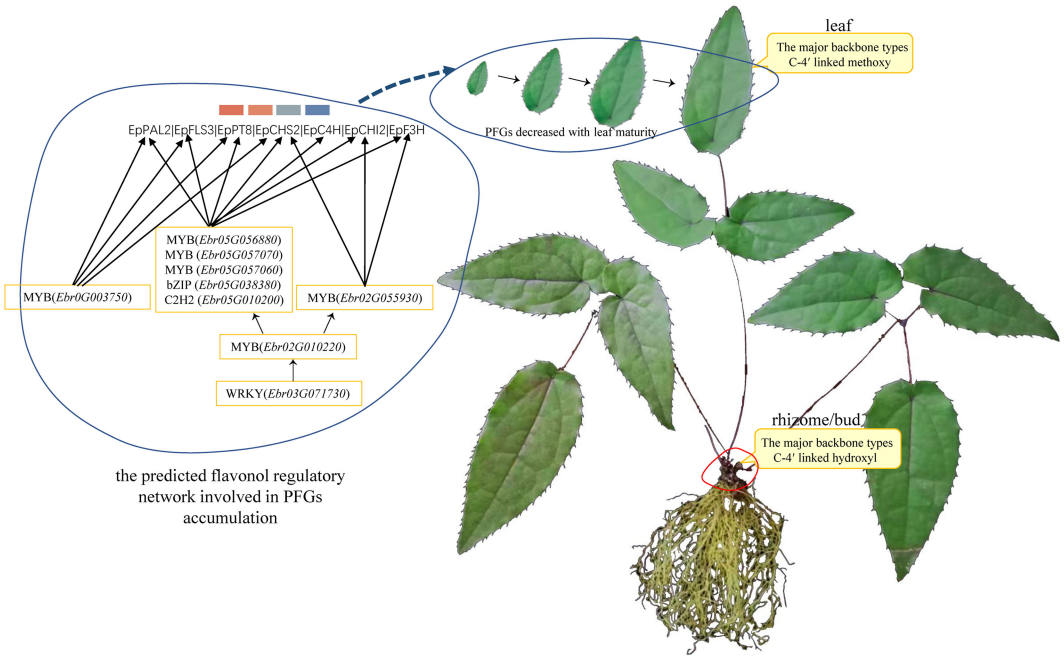


FIGURE 10
Possible model for this study.

suggest the potential regulatory relationship between TFs and TFs, and between TFs and structural genes, further studies are needed in order to confirm the connections. In future work, molecular biology experiments such as subcellular localization and genetic transformation of hairy roots and yeast one-hybrid could be utilized to further verify the candidate TF activity in regulating the prenylated flavonol glycoside content and the binding ability to target gene promoters.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Author contributions

BG and CX conceived and designed the study; CX, YZ, and XL prepared the materials, conducted the experiments, and analyzed all data; CX wrote the manuscript; XF, FS, GS, CS, BG, and CX were involved in data interpretation and finalizing the manuscript draft. All authors read and approved the final draft.

Funding

This work was financially supported by the CAMS Innovation Fund for Medical Sciences (CIFMS) (2021-I2M-1-031).

Acknowledgments

We are grateful to traditional Chinese medicine professional technical cooperative of Changhong for help in sample collection.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1183481/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Different UPLC-PDA fingerprint characteristics in different leaf development stages (S0~S5) of *E. pubescens*. (Rt=2.07 min) Diphyllodside B; (Rt=2.19 min)

Epimedoside A; (Rt=4.11 min) Epimedin A; (Rt=4.32 min) Epimedin B; (Rt=4.52 min) Epimedin C; (Rt=4.71 min) icariin; (Rt=5.50 min) 3‴-carbonyl-2″-β-L-quinovosyl-icariin; (Rt=6.15 min) Anhydroicaritin-3-*O*-(acetyl) rhamnopyranosyl-xylopyranosyl-7-*O*-glucopyranoside; (Rt=7.36 min) 2″-*O*-rhamnosyl-ikarisoside A; (Rt=7.42 min and Rt=7.46 min) Anhydroicaritin-3-*O*-(acetyl) rhamnopyranosyl-(acetyl)xylopyranosyl-7-*O*-glucopyranoside or its isomers; (Rt=7.53 min) Ikarisoside A; and (Rt=8.26 min) 2″-*O*-rhamnopyranosyl icariside II; (Rt=8.77 min) icariside II.

SUPPLEMENTARY FIGURE 2
Mass spectrogram of chromatographic notable peak. Each PFG was identified under negative- and positive ion mode. **(A)** Diphyllodside B; **(B)** Epimedoside A; **(C)** Epimedin A; **(D)** Epimedin B; **(E)** Epimedin C; **(F)** icariin; **(G)** 3‴-carbonyl-2″-β-L-quinovosyl-icariin; **(H)** Anhydroicaritin-3-*O*-(acetyl) rhamnopyranosyl-xylopyranosyl-7-*O*-glucopyranoside; **(I)** 2″-*O*-rhamnosyl-ikarisoside A; **(J)** Anhydroicaritin-3-*O*-(acetyl) rhamnopyranosyl-(acetyl)xylopyranosyl-7-*O*-glucopyranoside or its isomers; **(K)** Ikarisoside A; **(L)** 2″-*O*-rhamnopyranosyl icariside II; **(M)** icariside II.

SUPPLEMENTARY FIGURE 3
Mother nucleus structure of PFGs in all stages of *E. pubescens* leaves.

SUPPLEMENTARY FIGURE 4
Comparative analysis of biological processes by GO enrichment in multiple leaf development stages.

SUPPLEMENTARY FIGURE 5
GO enrichment of S1 *vs* S4 up-regulated genes.

SUPPLEMENTARY FIGURE 6
KEGG enrichment of S1 *vs* S4 up-regulated genes.

SUPPLEMENTARY FIGURE 7
Classification of 4CL genes of *E. pubescens* determined by the classification system of *A. thaliana*.

SUPPLEMENTARY FIGURE 8
Resolved hierarchical regulation for *EpC4H*. The notion is the same as .

SUPPLEMENTARY FIGURE 9
Resolved hierarchical regulation for *EpCHI2*. The notion is the same as .

SUPPLEMENTARY FIGURE 10
Resolved hierarchical regulation for *EpF3H*. The notion is the same as.

References

Aharoni, A., De Vos, C. R., Wein, M., Sun, Z., Greco, R., Kroon, A., et al. (2001). The strawberry *FaMYB1* transcription factor suppresses anthocyanin and flavonol accumulation in transgenic tobacco. *Plant J.* 28 (3), 319–332. doi: 10.1046/j.1365-313X.2001.01154.x

Blighe, K., and Lun, A. (2019). *PCAtools: everything principal components analysis. r package version 2.0.* https://github.com/kevinblighe/PCAtools

Chang, Y. M., Lin, H. H., Liu, W. Y., Yu, C. P., Chen, H. J., Wartini, P. P., et al. (2019). Comparative transcriptomics method to infer gene coexpression networks and its applications to maize and rice leaf transcriptomes. *Proc. Natl. Acad. Sci.* 116, 3091–3099. doi: 10.1073/pnas.1817621116

Conesa, A., Nueda, M. J., Ferrer, A., and Talón, M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics* 22 (9), 1096–1102. doi: 10.1093/bioinformatics/btl056

Czemmel, S., Stracke, R., Weisshaar, B., Cordon, N., Harris, N. N., Walker, A. R., et al. (2009). The grapevine R2R3-MYB transcription factor *VvMYBF1* regulates flavonol synthesis in developing grape berries. *Plant Physiol.* 151 (3), 1513–1530. doi: 10.1104/pp.109.142059

Feng, K., Chen, R., Xie, K., Chen, D., Guo, B., Liu, X., et al. (2018). A regiospecific rhamnosyltransferase from *Epimedium pseudowushanense* catalyzes the 3-*O*-rhamnosylation of prenylflavonols. *Org. Biomol. Chem.* 16, 452–458. doi: 10.1039/C7OB02763J

Feng, K., Chen, R., Xie, K., Chen, D., Liu, J., Du, W., et al. (2019). Ep7GT, a glycosyltransferase with sugar donor flexibility from *Epimedium pseudowushanense*, catalyzes the 7-*O*-glycosylation of baohuoside. *Org. Biomol. Chem.* 17, 8106–8114. doi: 10.1039/C9OB01352K

Fischer, D. S., Theis, F. J., and Yosef, N. (2018). Impulse model-based differential expression analysis of time course sequencing data. *Nucleic Acids Res.* 46 (20), e119–e119. doi: 10.1093/nar/gky675

Grotewold, E. (2008). Transcription factors for predictive plant metabolic engineering: are we there yet? *Curr. Opin. Biotechnol.* 19, 138–144. doi: 10.1016/j.copbio.2008.02.002

Grunewald, W., De Smet, I., Lewis, D. R., Löfke, C., Jansen, L., Goeminne, G., et al. (2012). Transcription factor *WRKY23* assists auxin distribution patterns during arabidopsis root development through local control on flavonol biosynthesis. *Proc. Natl. Acad. Sci.* 109 (5), 1554–1559. doi: 10.1073/pnas.1121134109

He, Y., Pan, L., Yang, T., Wang, W., Li, C., Chen, B., et al. (2021). Metabolomic and confocal laser scanning microscopy (clsm) analyses reveal the important function of flavonoids in amygdalus pedunculata pall leaves with temporal changes. *Front. Plant Sci.* 12, 648277. doi: 10.3389/fpls.2021.648277

Huang, W., Khaldun, A., Chen, J., Zhang, C., Lv, H., Yuan, L., et al. (2016a). A R2R3-MYB transcription factor regulates the flavonol biosynthetic pathway in a traditional Chinese medicinal plant, *Epimedium sagittatum*. *Front. Plant Sci.* 7, 1089. doi: 10.3389/fpls.2016.01089

Huang, W., Khaldun, A., Lv, H., Du, L., Zhang, C., and Wang, Y. (2016b). Isolation and functional characterization of a R2R3-MYB regulator of the anthocyanin biosynthetic pathway from *Epimedium sagittatum*. *Plant Cell Rep.* 35, 883–894. doi: 10.1007/s00299-015-1929-z

Huang, W., Sun, W., Lv, H., Luo, M., Zeng, S., Pattanaik, S., et al. (2013). A R2R3-MYB transcription factor from *Epimedium sagittatum* regulates the flavonoid biosynthetic pathway. *PloS One* 8, e70778. doi: 10.1371/journal.pone.0070778

Huang, W., Sun, W., Lv, H., Xiao, G., Zeng, S., and Wang, Y. (2012). Isolation and molecular characterization of thirteen R2R3-MYB transcription factors from *Epimedium sagittatum*. *Int. J. Mol. Sci.* 14, 594–610. doi: 10.3390/ijms14010594

Huang, W., Zeng, S., Xiao, G., Wei, G., Liao, S., Chen, J., et al. (2015). Elucidating the biosynthetic and regulatory mechanisms of flavonoid-derived bioactive components in *Epimedium sagittatum*. *Front. Plant Sci.* 6, 689. doi: 10.3389/fpls.2015.00689

Kang, H. K., Choi, Y. H., Kwon, H., Lee, S. B., Kim, D. H., Sung, C. K., et al. (2012). Estrogenic/antiestrogenic activities of a *Epimedium koreanum* extract and its major components: *in vitro* and *in vivo* studies. *Food Chem. Toxicol.* 50, 2751–2759. doi: 10.1016/j.fct.2012.05.017

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Kumar, L., and Futschik, M. E. (2007). Mfuzz: a software package for soft clustering of microarray data. *Bioinformation* 2 (1), 5. doi: 10.6026/97320630002005

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., Mcwilliam, H., et al. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Letunic, I., and Bork, P. (2019). Interactive tree of life (iTOL) v4: recent updates and new developments. *Nuc. Acids Res.* 47, 256–259. doi: 10.1093/nar/gkz239

Li, S. (2014). Transcriptional control of flavonoid biosynthesis: fine-tuning of the MYB-bHLH-WD40 (MBW) complex. *Plant Signaling Behav.* 9 (1), e27522. doi: 10.4161/psb.27522

Li, C., Wu, J., Hu, K. D., Wei, S. W., Sun, H. Y., Hu, L. Y., et al. (2020). PyWRKY26 and PybHLH3 cotargeted the PyMYB114 promoter to regulate anthocyanin biosynthesis and transport in red-skinned pears. *Hortic. Res.* 7, 37–48. doi: 10.1038/s41438-020-0254-z

Li, S., and Zachgo, S. (2013). TCP3 interacts with R2R3-MYB proteins, promotes flavonoid biosynthesis and negatively regulates the auxin response in *Arabidopsis thaliana*. *Plant J.* 76 (6), 901–913. doi: 10.1111/tpj.12348

Li, C., Zhang, L., Niu, D., Nan, S., Miao, X., Hu, X., et al. (2021). Investigation of flavonoid expression and metabolite content patterns during seed formation of *Artemisia sphaerocephala* krasch. *Seed Sci. Res.* 31 (2), 136–148. doi: 10.1017/S096025852100012X

Liang, T., Shi, C., Peng, Y., Tan, H., Xin, P., Yang, Y., et al. (2020). Brassinosteroid-activated BRI1-EMS-SUPPRESSOR 1 inhibits flavonoid biosynthesis and coordinates growth and UV-b stress responses in plants. *Plant Cell* 32, 3224–3239. doi: 10.1105/tpc.20.00048

Liao, Y., Smyth, G. K., and Shi, W. (2019). The r package rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nuc. Acids Res.* 47 (8), e47. doi: 10.1093/nar/gkz114

Liu, W., Feng, Y., Yu, S., Fan, Z., Li, X., Li, J., et al. (2021). The flavonoid biosynthesis network in plants. *Int. J. Mol. Sci.* 22 (23), 12824. doi: 10.3390/ijms222312824

Liu, C., Long, J., Zhu, K., Liu, L., Yang, W., Zhang, H., et al. (2016). Characterization of a citrus R2R3-MYB transcription factor that regulates the flavonol and hydroxycinnamic acid biosynthesis. *Sci. Rep.* 6 (1), 1–16. doi: 10.1038/srep25352

Liu, Y., Lou, Q., Xu, W., Xin, Y., Bassett, C., and Wang, Y. (2011). Characterization of a chalcone synthase (CHS) flower-specific promoter from *Lilium orential* 'Sorbonne'. *Plant Cell Rep.* 30, 2187–2194. doi: 10.1007/s00299-011-1124-9

Liu, Y., Wu, L., Deng, Z., and Yu, Y. (2021). Two putative parallel pathways for naringenin biosynthesis in *Epimedium wushanense*. *RSC Adv.* 11, 13919–13927. doi: 10.1039/D1RA00866H

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15 (12), 550. doi: 10.1186/s13059-014-0550-8

Lyu, Y. (2020). Identification and characterization of three flavonoid 3-*O*-glycosyltransferases from *Epimedium koreanum* nakai. *Bioc. Engi. J.* 163, 107759. doi: 10.1016/j.bej.2020.107759

Ma, H., He, X., Yang, Y., Li, M., Hao, D., and Jia, Z. (2011). The genus epimedium: an ethnopharmacological and phytochemical review. *J. Ethno.* 134, 519–541. doi: 10.1016/j.jep.2011.01.001

Ma, B., Wu, J., Shi, T. L., Yang, Y. Y., Wang, W. B., Zheng, Y., et al. (2022). Lilac (*Syringa oblata*) genome provides insights into its evolution and molecular mechanism of petal color change. *Commun. Biol.* 5 (1), 686. doi: 10.1038/s42003-022-03646-9

Malacarne, G., Coller, E., Czemmel, S., Vrhovsek, U., Engelen, K., Goremykin, V., et al. (2016). The grapevine *VvibZIPC22* transcription factor is involved in the regulation of flavonoid biosynthesis. *J. Exp. Bot.* 67 (11), 3509–3522. doi: 10.1093/jxb/erw181

Mehrtens, F., Kranz, H., Bednarek, P., and Weisshaar, B. (2005). The arabidopsis transcription factor *MYB12* is a flavonol-specific regulator of phenylpropanoid biosynthesis. *Plant Physiol.* 138 (2), 1083–1096. doi: 10.1104/pp.104.058032

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Naik, J., Misra, P., Trivedi, P. K., and Pandey, A. (2022). Molecular components associated with the regulation of flavonoid biosynthesis. *Plant Sci.* 317, 111196. doi: 10.1016/j.plantsci.2022.111196

Nakatsuka, T., Saito, M., Yamada, E., Fujita, K., Kakizaki, Y., and Nishihara, M. (2012). Isolation and characterization of *GtMYBP3* and *GtMYBP4*, orthologues of R2R3-MYB transcription factors that regulate early flavonoid biosynthesis, in gentian flowers. *J. Exp. Bot.* 63 (18), 6505–6517. doi: 10.1093/jxb/ers306

Pan, J., Chen, H., Guo, B., and Liu, C. (2017). Understanding the molecular mechanisms underlying the effects of light intensity on flavonoid production by RNA-seq analysis in *Epimedium pseudowushanense* BL guo. *PloS One* 12, e0182348. doi: 10.1371/journal.pone.0182348

Pant, S., and Huang, Y. (2022). Genome-wide studies of PAL genes in sorghum and their responses to aphid infestation. *Sci. Rep.* 12 (1), 22537. doi: 10.1038/s41598-022-25214-1

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shen, G., Luo, Y., Yao, Y., Meng, G., Zhang, Y., Wang, Y., et al. (2022). The discovery of a key prenyltransferase gene assisted by a chromosome-level *Epimedium pubescens* genome. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1034943

Song, Z., Luo, Y., Wang, W., Fan, N., Wang, D., Yang, C., et al. (2020). *NtMYB12* positively regulates flavonol biosynthesis and enhances tolerance to low pi stress in *Nicotiana tabacum*. *Front. Plant Sci.* 101683. doi: 10.3389/fpls.2019.01683

Stracke, R., Ishihara, H., Huep, G., Barsch, A., Mehrtens, F., Niehaus, K., et al. (2007). Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the *Arabidopsis thaliana* seedling. *Plant J.* 50 (4), 660–677. doi: 10.1111/j.1365-313X.2007.03078.x

Stracke, R., Jahns, O., Keck, M., Tohge, T., Niehaus, K., Fernie, A. R., et al. (2010). Analysis of PRODUCTION OF FLAVONOL GLYCOSIDES-dependent flavonol glycoside accumulation in *Arabidopsis thaliana* plants reveals MYB11-, MYB12- and MYB111-independent flavonol glycoside accumulation. *New Phytol.* 188 (4), 985–1000. doi: 10.1111/j.1469-8137.2010.03421.x

Tong, Q., Zhang, C., Tu, Y., Chen, J., Li, Q., Zeng, Z., et al. (2022). Biosynthesis-based spatial metabolome of *Salvia miltiorrhiza* bunge by combining metabolomics approaches with mass spectrometry-imaging. *Talanta* 238, 123045. doi: 10.1016/j.talanta.2021.123045

Valares Masa, C., Sosa Díaz, T., Alías Gallego, J. C., and Chaves Lobón, N. (2016). Quantitative variation of flavonoids and diterpenes in leaves and stems of *Cistus ladanifer* l. at different ages. *Molecules* 21, 275. doi: 10.3390/molecules21030275

Wallace, J., Larsson, S., and Buckler, E. (2014). Entering the second century of maize quantitative genetics. *Heredity* 112, 30–38. doi: 10.1038/hdy.2013.6

Wang, Q., Jiang, Y., Mao, X., Yu, W., Lu, J., and Wang, L. (2022). Integration of morphological, physiological, cytological, metabolome and transcriptome analyses reveal age inhibited accumulation of flavonoid biosynthesis in *Ginkgo biloba* leaves. *Ind. Crops. Prod.* 187, 115405. doi: 10.1016/j.indcrop.2022.115405

Wang, P., Li, C., Li, X., Huang, W., Wang, Y., Wang, J., et al. (2021). Complete biosynthesis of the potential medicine icaritin by engineered *Saccharomyces cerevisiae* and *Escherichia coli*. *Sci. Bull.* 66, 1906–1916. doi: 10.1016/j.scib.2021.03.002

Wang, Z., Zhang, X., Wang, H., Qi, L., and Lou, Y. (2007). Neuroprotective effects of icaritin against beta amyloid-induced neurotoxicity in primary cultured rat neuronal cells *via* estrogen-dependent pathway. *Neuroscience* 145, 911–922. doi: 10.1016/j.neuroscience.2006.12.059

Wong, C. Y., Chang, Y. M., Tsai, Y. S., Ng, W. V., Cheong, S. K., Chang, T. Y., et al. (2020). Decoding the differentiation of mesenchymal stem cells into mesangial cells at the transcriptomic level. *BMC Genomics* 21, 1–14. doi: 10.1186/s12864-020-06868-5

Xie, P. S., Yan, Y. Z., Guo, B. L., Lam, C., Chui, S., and Yu, Q. (2010). Chemical pattern-aided classification to simplify the intricacy of morphological taxonomy of epimedium species using chromatographic fingerprinting. *J. Pharm. Biomed. Anal.* 52, 452–460. doi: 10.1016/j.jpba.2010.01.025

Xu, W., Dubos, C., and Lepiniec, L. (2015). Transcriptional control of flavonoid biosynthesis by MYB-bHLH-WDR complexes. *Trends Plant Sci.* 20 (3), 176–185. doi: 10.1016/j.tplants.2014.12.001

Xu, W., He, S., Huang, M., and Wang, Y. (2007). Determination of icariin contents indifferent plant parts of epimedium plants in guizhou by HPLC. *Chin. J. Exp. Tradit. Med. Form.* 13 (5), 1–3.

Xu, J., Luo, H., Zhou, S. S., Jiao, S. Q., Jia, K. H., Nie, S., et al. (2022). UV-B and UV-c radiation trigger both common and distinctive signal perceptions and transmissions in *Pinus tabuliformis* Carr. *Tree Physiol.* 42 (8), 1587–1600. doi: 10.1093/treephys/tpac021

Xu, J., Nie, S., Xu, C. Q., Liu, H., Jia, K. H., Zhou, S. S., et al. (2021). UV-B-induced molecular mechanisms of stress physiology responses in the major northern Chinese conifer *Pinus tabuliformis* Carr. *Tree Physiol.* 41 (7), 1247–1263. doi: 10.1093/treephys/tpaa180

Yadav, S. K., Kumar, V., and Singh, S. P. (2018). *Recent trends and techniques in plant metabolic engineering* (Singapore: Springer Nature Singapore Pte). doi: 10.1007/978-981-13-2251-8

Yang, X., Chen, J., Huang, W., Zhang, Y., Yan, X., Zhou, Z., et al. (2020). Synthesis of icariin in tobacco leaf by overexpression of a glucosyltransferase gene from *Epimedium sagittatum*. *Ind. Crops. Prod.* 156, 112841. doi: 10.1016/j.indcrop.2020.112841

Yang, F. S., Nie, S., Liu, H., Shi, T. L., Tian, X. C., Zhou, S. S., et al. (2020). Chromosome-level genome assembly of a parent species of widely cultivated azaleas. *Nat. Commun.* 11 (1), 5269. doi: 10.1038/s41467-020-18771-4

Yao, Y., Gu, J., Luo, Y., Wang, Y., Pang, Y., Shen, G., et al. (2022a). Genome-wide analysis of UGT gene family identified key gene for the biosynthesis of bioactive flavonol glycosides in *Epimedium pubescens* maxim. *Synth. Syst. Biotechnol.* 7, 1095–1107. doi: 10.1016/j.synbio.2022.07.003

Yao, Y., Gu, J., Luo, Y., Zhang, Y., Wang, Y., Pang, Y., et al. (2022b). A novel 3-*O*-rhamnoside: 2''-*O*-xylosyltransferase responsible for terminal modification of prenylflavonol glycosides in *Epimedium pubescens* maxim. *Int. J. Mol. Sci.* 23, 16050. doi: 10.3390/ijms232416050

Yao, P., Huang, Y., Dong, Q., Wan, M., Wang, A., Chen, Y., et al. (2020). *FtMYB6*, a light-induced SG7 R2R3-MYB transcription factor, promotes flavonol biosynthesis in tartary buckwheat (*Fagopyrum tataricum*). *J. Agric. Food Chem.* 68 (47), 13685–13696. doi: 10.1021/acs.jafc.0c03037

Yu, J., Jiang, Q., Sun, R., and Lv, H. (2012). Determination of the effective components in different parts and harvest time of *Epimedium koreanum*. *Chin. J. Exp. Tradit. Med. Form.* 18, 92–95.

Yu, G., Wang, L. G., Han, Y., and He, Q. (2012). clusterProfiler: an r package for comparing biological themes among gene clusters. *J. Integr. Biol.* 16, 284–287. doi: 10.1089/omi.2011.0118

Zeng, S., Liu, Y., Hu, W., Liu, Y., Shen, X., and Wang, Y. (2013a). Integrated transcriptional and phytochemical analyses of the flavonoid biosynthesis pathway in epimedium. *Plant Cell Tiss. Organ Cult.* 115, 355–365. doi: 10.1007/s11240-013-0367-2

Zeng, S., Liu, Y., Zou, C., Huang, W., and Wang, Y. (2013b). Cloning and characterization of phenylalanine ammonia-lyase in medicinal epimedium species. *Plant Cell Tiss. Organ. Cult.* 113, 257–267. doi: 10.1007/s11240-012-0265-z

Zhai, R., Zhao, Y., Wu, M., Yang, J., Li, X., Liu, H., et al. (2019). The MYB transcription factor *PbMYB12b* positively regulates flavonol biosynthesis in pear fruit. *BMC Plant Biol.* 19 (1), 1–11. doi: 10.1186/s12870-019-1687-0

Zhang, H., Yang, X., Guo, Y., and Wang, Y. (2009). Sustainable use of epimedium resources: current status and prospects. *Chin. Bull. Bot.* 44, 363. doi: 10.3969/j.issn.1674-3466.2009.03.014

Zhao, J., and Dixon, R. (2010). The 'ins' and 'outs' of flavonoid transport. *Trends Plant Sci.* 15, 72–80. doi: 10.1016/j.tplants.2009.11.006

Zhao, H., Guo, Y., Li, S., Han, R., Ying, J., Zhu, H., et al. (2015). A novel anti-cancer agent icaritin suppresses hepatocellular carcinoma initiation and malignant growth through the IL-6/Jak2/Stat3 pathway. *Oncotarget* 6, 31927. doi: 10.18632/oncotarget.5578

Zhao, Y., Jia, K., Tian, Y., Han, K., El-Kassaby, Y. A., Yang, H., et al. (2023). Time-course transcriptomics analysis reveals key responses of populus to salt stress. *Ind. Crops Prod.* 194, 116278. doi: 10.1016/j.indcrop.2023.116278

Zhao, H. Y., Sun, J. H., Fan, M. X., Fan, L., Zhou, L., Li, Z., et al. (2008). Analysis of phenolic compounds in epimedium plants using liquid chromatography coupled with electrospray ionization mass spectrometry. *J. @ Chrom. A* 1190, 157–181. doi: 10.1016/j.chroma.2008.02.109

Zhao, X., Zeng, X., Lin, N., Yu, S., Fernie, A. R., and Zhao, J. (2021). CsbZIP1-CsMYB12 mediates the production of bitter-tasting flavonols in tea plants (*Camellia sinensis*) through a coordinated activator–repressor network. *Hortic. Res.* 8, 110. doi: 10.1038/s41438-021-00545-8

Zhao, Q., Zhang, Y., Wang, G., Hill, L., Weng, J. K., Chen, X. Y., et al. (2016). A specialized flavone biosynthetic pathway has evolved in the medicinal plant, *Scutellaria baicalensis*. *Sci. Adv.* 2 (4), e1501780. doi: 10.1126/sciadv.1501780

Zheng, Y., Jiao, C., Sun, H., Rosli, H. G., Pombo, M. A., Zhang, P., et al. (2016). iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* 9, 1667–1670. doi: 10.1016/j.molp.2016.09.014

Zhou, J., Gao, S., Chen, J., Zeng, W., and Yu, S. (2021). *A flavonoid 4'-o-methyltransferase from epimedium koreanum and its application*. China Patent No ZL202111098372.9. (China: State Intellectual Property Office, China).

Zhong, C., Tang, Y., Pang, B., Li, X., Yang, X., Deng, Y., et al. (2020). The R2R3-MYB transcription factor *GhMYB1a* regulates flavonol and anthocyanin accumulation in *Gerbera hybrida*. *Hortic. Res.* 7, 78. doi: 110.1038/s41438-020-0296-2

Zhou, H., Wang, K., Wang, H., Gu, C., Dare, A., Espley, R., et al. (2015). Molecular genetics of blood-fleshed peach reveals activation of anthocyanin biosynthesis by NAC transcription factors. *Plant J.* 82, 105–121. doi: 10.1111/tpj.12792

Zhou, M., Zheng, W., Sun, X., Yuan, M., Zhang, J., Chen, X., et al. (2021). Comparative analysis of chemical components in different parts of epimedium herb. *J. Pharm. Biomed. Anal.* 198, 113984. doi: 10.1016/j.jpba.2021.113984

Zhu, J. H., Li, H. L., Guo, D., Wang, Y., Dai, H. F., Mei, W. L., et al. (2016). Transcriptome-wide identification and expression analysis of glutathione s-transferase genes involved in flavonoids accumulation in *Dracaena cambodiana*. *Plant Physiol. Biochem.* 104, 304–311. doi: 10.1016/j.plaphy.2016.05.012

Zhu, J., Li, Z., Zhang, G., Meng, K., Kuang, W., Li, J., et al. (2011). Icaritin shows potent anti-leukemia activity on chronic myeloid leukemia *in vitro* and *in vivo* by regulating MAPK/ERK/JNK and JAK2/STAT3/AKT signalings. *PloS One* 6, e23720. doi: 10.1371/journal.pone.0023720

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 13 June 2023
DOI 10.3389/fpls.2023.1183361

OPEN ACCESS

EDITED BY
Kai-Hua Jia,
Shandong Academy of Agricultural Sciences, China

REVIEWED BY
Hui Liu,
Chinese Academy of Sciences (CAS), China
Joaquim Martins Jr.,
Centro Nacional de Pesquisa em Energia e Materiais, Brazil

*CORRESPONDENCE
Lei Wang
2890902708@qq.com
Hao Lu
incana96@163.com

†These authors have contributed equally to this work

RECEIVED 10 March 2023
ACCEPTED 26 May 2023
PUBLISHED 13 June 2023

CITATION
Tang M, Huang J, Ma X, Du J, Bi Y, Guo P, Lu H and Wang L (2023) A near-complete genome assembly of *Thalia dealbata* Fraser (Marantaceae). *Front. Plant Sci.* 14:1183361. doi: 10.3389/fpls.2023.1183361

A near-complete genome assembly of *Thalia dealbata* Fraser (Marantaceae)

Min Tang[1†], Jialin Huang[2†], Xiangli Ma[3], Juan Du[1], Yufen Bi[3], Peiwen Guo[4], Hao Lu[5*] and Lei Wang[5*]

[1]College of Landscape and Horticulture,Yunnan Agricultural University, Kunming, China, [2]School of Chemical Biology and Environment, Yuxi Normal University, Yuxi, China, [3]College of Animal Science and Technology, Yunnan Agricultural University, Kunming, China, [4]School of Mathematical Sciences, Xiamen University, Xiamen, China, [5]Scientific Research Department, Kunming Novo Medical Laboratory Co., Ltd., Kunming, China

This study presents a chromosome-level, near-complete genome assembly of *Thalia dealbata* (Marantaceae), a typical emergent wetland plant with high ornamental and environmental value. Based on 36.99 Gb PacBio HiFi reads and 39.44 Gb Hi-C reads, we obtained a 255.05 Mb assembly, of which 251.92 Mb (98.77%) were anchored into eight pseudo-chromosomes. Five pseudo-chromosomes were completely assembled, and the other three had one to two gaps. The final assembly had a high contig N50 value (29.80 Mb) and benchmarking universal single-copy orthologs (BUSCO) recovery score (97.52%). The *T. dealbata* genome had 100.35 Mb repeat sequences, 24,780 protein-coding genes, and 13,679 non-coding RNAs. Phylogenetic analysis revealed that *T. dealbata* was closest to *Zingiber officinale*, whose divergence time was approximately 55.41 million years ago. In addition, 48 and 52 significantly expanded and contracted gene families were identified within the *T. dealbata* genome. Moreover, 309 gene families were specific to *T. dealbata*, and 1,017 genes were positively selected. The *T. dealbata* genome reported in this study provides a valuable genomic resource for further research on wetland plant adaptation and the genome evolution dynamics. This genome is also beneficial for the comparative genomics of Zingiberales species and flowering plants.

KEYWORDS

***Thalia dealbata*, near-complete genome assembly, PacBio HiFi, genome annotation, wetland plant**

1 Introduction

Wetlands, also known as the "kidneys of the earth", are of great ecological importance because they have played important roles in biodiversity conservation, carbon management, flood reduction, and water purification (Zedler and Kercher, 2005). Although wetlands cover less than 9% of the land area, they are vital habitats to many

aquatic plants and animals (Gray et al., 2013). As key components of wetland ecosystems, wetland plants function as primary producers, habitats for other taxonomic groups, and nutrient removers (Cronk and Fennessy, 2016). Almost all wetland plants are angiosperms, with a few ferns and gymnosperms. These plants are categorized into emergent, submergent, floating-leaved, and floating plants based on their growth types and morphologies (Cronk and Fennessy, 2016). Although wetland plants have developed adaptation strategies to survive periodic soil saturation and the accompanying changes in soil chemistry (Pezeshki, 2001), the underlying genetic mechanisms in survival strategies are rarely studied. With the rapid development of sequencing technologies, the characterization of more wetland plant genomes can provide deeper insights into the adaptive evolution and morphological characteristics of wetland plants.

Thalia dealbata Fraser (Marantaceae), commonly known as powdery alligator flag, is a typical emergent wetland plant native to swamps and ponds in the Southern United States of America and Mexico. It has high ornamental value, given its long-stalked canna-like foliage and violet-blue flowers. This plant is usually covered with a white and water-repellent powdery coating, which enhances its performance. *T. dealbata* is also widely used in man-made wetlands to improve water quality by breaking down or removing excess pollutants from eutrophic water (Wang et al., 2020). A recent study has presented the complete chloroplast genome of *T. dealbata* (Deng et al., 2021). However, the *T. dealbata* nuclear genome has not been sequenced or reported.

Therefore, this study aimed to reconstruct a reference genome sequence of *T. dealbata* for further genomic and genetic studies. We performed a chromosome-level assembly of the *T. dealbata*, the first sequenced genome in the Marantaceae family, by integrating PacBio high-fidelity (HiFi) sequencing and chromosome conformation capture (Hi-C) technology. Subsequently, we performed a comparative genomics analysis of *T. dealbata* and other publicly available Zingiberales species, including *Musa acuminata* Colla (D'hont et al., 2012), *M. balbisiana* Colla (Wang et al., 2019), *Zingiber officinale* Roscoe (Li et al., 2021), and *Ensete glaucum* (Roxb.) Cheesman (Wang Z. et al., 2022). The reference-level genome assembly in this study will accelerate evolutionary and morphological studies of wetland plants and further phylogenomic studies of Marantaceae and Zingiberales.

2 Materials and methods

2.1 Sample collection and sequencing

Young and healthy *T. dealbata* leaves were collected from a mature *T. dealbata* plant growing on the lakeside of Dongan Lake in Chengdu, Sichuan Province, Southwest China. The leaves were immediately frozen in liquid nitrogen and stored at -80°C, awaiting further analysis.

High-quality total genomic DNA was extracted from the *T. dealbata* leaves using the CTAB method (Doyle and Doyle, 1987). For genome survey analysis, a paired-end library with an insert size of approximately 400 bp was constructed and sequenced on a NovaSeq 6000 platform. For the *de novo* assembly of the genome, SMRTbell libraries were prepared using the PacBio 15-kb protocol (Pacific Biosciences, CA, USA) and sequenced using the circular consensus sequencing (CCS) mode on the PacBio Sequel II sequencer. Finally, a Hi-C library was constructed and sequenced for the Hi-C scaffolding analysis on a NovaSeq 6000 platform.

In addition, the *T. dealbata* total RNA was extracted from the fresh stem, leaf, and flower tissues. Next, the RNA sequencing (RNA-seq) libraries were constructed and sequenced on an Illumina NovaSeq 6000 platform. The obtained raw RNA-seq reads were filtered using Trimmomatic (version 0.36) with default parameters (Bolger et al., 2014) for downstream genome annotation and quality assessment.

2.2 Genome survey

A *k*-mer ($k = 19$) analysis of Illumina short reads was performed using the Jellyfish (version 2.2.9, parameters, -k 19, -C) (Marçais and Kingsford, 2011). The low-frequency *k*-mers (frequency < 4) were removed, and the genome size was calculated by dividing the total *k*-mer number by the homozygous peak depth in the *k*-mer distribution curve. In addition, a polyploidy peak around the homozygous peak was examined to determine the ploidy level of the *T. dealbata* genome.

2.3 Genome assembly and quality assessment

HiFi long reads were pre-processed by CCS (version 4.2.0, parameters, min-passes = 3, min-length = 10, and min-rq = 0.99; http://ccs.how). Next, the filtered HiFi reads were assembled into contigs using hifiasm (version 0.14, default parameters) (Cheng et al., 2021) and mapped using Minimap (version 2.24, default parameters) (Li, 2018). Subsequently, the low-quality contigs with read depth < 10 or GC content > 50% were removed based on the GC-depth distribution. For Hi-C scaffolding analysis, Hi-C reads were mapped using Juicebox (version 1.8.8) with default parameters (Durand et al., 2016). Uniquely mapped Hi-C reads were then used to anchor contigs into chromosomes using 3D-DNA software (Dudchenko et al., 2017). Finally, scaffolding errors were checked and corrected according to the Hi-C contact heat maps generated with Juicebox.

The final assembly quality was evaluated by re-mapping the Illumina reads against the assembly using BWA (version 0.7.17) with default parameters (Li and Durbin, 2009). Subsequently, the benchmarking universal single-copy orthologs (BUSCO) completeness score was calculated by mapping 1,614 conserved genes from Embryophyta odb10 against the assembly using BUSCO (version 3.0.2) with default parameters (Simão et al., 2015). We searched the plant telomeres that are listed in the telomerase database (Podlevsky et al., 2007) against the final assembly using an in-house perl script. In addition, we identified centromeres within the *T. dealbata* genome using quarTeT (http://www.atcgn.com:8080/quarTeT/home.html) with the similar procedures which were described in Yue et al. (2023).

2.4 Identification of repeats

The repetitive elements in the *T. dealbata* genome were annotated using RepeatMasker (version v4.07) (Tarailo-Graovac and Chen, 2009) and RepeatModeler (version v1.0.11) (Price et al., 2005). First, a repeat library was *de novo* predicted based on the final assembly using RepeatModeler with default parameters. Next, a known repeat library of green plants was extracted using the "queryRepeatDatabase.pl" script from RepeatModeler. Finally, the two repeat libraries were combined into one comprehensive library, followed by a genome-wide homology-based identification of repeats using RepeatMasker. In addtion, we annotated intact long terminal repeat (LTR) retrotransposons (LTR-RTs) by integrating the predictions from LTR_Finder (version 1.06) (Xu and Wang, 2007) and LTRharvest (version 1.5.10) (Ellinghaus et al., 2008) using LTR_retriever (Ou and Jiang, 2018) with default parameters.

2.5 Annotation of protein-coding genes

Protein-coding genes were annotated as outlined by Wang et al. (2021). First, *E. glaucum*, *M. acuminata*, *M. balbisiana*, *Z. officinale*, and *Oryza sativa* L. protein sequences (Jain et al., 2019) were aligned with the *T. dealbata* genome using TBLASTN (version 2.2.31+, parameters, E-value < 1e-5) (Camacho et al., 2009), and a homology-based prediction was performed using GeneWise (version 2.4.1) with default parameters (Birney et al., 2004). Second, the *de novo* and genome-guided RNA-seq assemblies were combined for transcriptome-based prediction using the program to assemble spliced alignment (PASA; version 2.3.3) with default parameters (Haas et al., 2003). Third, a *de novo* prediction of gene models was performed using AUGUSTUS (version 3.2.3) (Stanke et al., 2006) with high-confidence gene model-trained parameters (exon number > 2 and CDS length > 1200 bp) selected from the PASA results. Finally, all the predicted gene models were integrated into a final gene set using EvidenceModeler (version 1.1.1) with default parameters (Haas et al., 2008). The final protein-coding gene set was functionally annotated using the publicly available protein databases, including Swiss-Prot, TrEMBL (Bairoch and Apweiler, 2000), InterPro (Hunter et al., 2009), and KEGG (Moriya et al., 2007), as described by Wang M. et al., (2022). Gene ontology (GO) terms were then assigned based on InterPro entries.

2.6 Annotation of non-coding RNAs

Non-coding RNAs (ncRNAs), including ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), microRNAs (miRNAs), and small nuclear RNAs (snRNAs) were annotated using the *de novo* and homology-based methods. The rRNAs were predicted by aligning the assembly against the *Arabidopsis thaliana* rRNA sequences using BLASTN (version 2.2.31+, parameters, E-value < 1e-5). The tRNAs were predicted using tRNAscan-SE (version 1.4) (Lowe and Eddy, 1997), while snRNAs and miRNAs were predicted using Infernal (version 1.1.3) with default parameters (Nawrocki and Eddy, 2013).

2.7 Phylogenetic analysis and divergence time estimation

A phylogenetic analysis was performed based on the protein sequences of *T. dealbata* and four Zingiberales species, including *E. glaucum*, *M. acuminata*, *M. balbisiana*, and *Z. officinale* (haplotype A), with *O. sativa* as the outgroup species. The gene family clustering was performed using OrthoMCL (version 2.0.9) with default parameters (Li et al., 2003). Single copy genes (SCGs) in the six species were identified based on the clustering results. In addition, the SCGs protein sequences were aligned using MAFFT (version 7.313, parameters, LINSI) (Katoh et al., 2002). Finally, a maximum likelihood phylogenetic tree was reconstructed from the alignments of concatenated SCGs using RAxML (version 8.0.17, parameters, PROTGAMMAILGX, n = 500) (Stamatakis, 2014). The divergence time of *T. dealbata* was estimated using MCMCTREE in the PAML (version 4.9e, parameters, independent rates, F84 model) (Yang, 2007) based on the divergence between *O. sativa* and *E. glaucum* (103.2–117.1 million years ago, MYA) from the TimeTree database (Kumar et al., 2017) as the calibration point. Subsequently, the gene family expansions and contractions per species were detected using CAFE (version 3.1) with default parameters (De Bie et al., 2006).

2.8 Whole genome duplication (WGD) analysis

Recent WGD events within the *T. dealbata* genome were analyzed by comparing *T. dealbata* and *M. acuminata* protein sequences using MCScanX (version 1.1) with default parameters (Wang et al., 2012). Next, the synonymous substitution rate (*Ks*) was calculated per collinear gene pair within and between the two species using "add_ka_and_ks_to_collinearity.pl". Synteny blocks shared between *T. dealbata* and *M. acuminata* and in *T. dealbata* were visualized using TBtools (version 1.120) (Chen et al., 2020). Different gene duplication types were detected using DupGen_finder with default parameters (Qiao et al., 2019).

2.9 Selection analysis

Three species, including *T. dealbata*, *M. acuminata*, and *O. sativa* were selected for selection analysis. The coding sequences of SCGs among the three species were aligned using MAFFT and trimmed by Gblocks (version 0.91b) with default parameters (Castresana, 2000). Finally, selection analysis was performed based on the branch site model using Codeml in PAML (version 4.9e). The LRT p-value was calculated using Chi2 in PAML.

3 Results

3.1 Genome sequencing and assembly

A total of 21.30 Gb Illumina reads were generated for genome survey analysis (Table S1). The *T. dealbata* genome was estimated to be 256.15 Mb in size, with no evidence of polyploidy (Figure S1). In addition, 36.99 Gb (144.40× genome coverage) HiFi reads were generated for *de novo* genome assembly, which yielded 145 contigs with a total length of 260.54 Mb, which is very close to the estimated genome size. After removing the low-quality contigs with low read depth or high GC content (Figure S2), a Hi-C scaffolding analysis on 39.44 Gb Hi-C reads (153.97× genome coverage), yielded 251.92 Mb sequences that were anchored to eight pseudo-chromosomes (Figure S3).

The final genome assembly (Figure 1) was 255.05 Mb in length, with contig and super-scaffold N50 of 29.80 and 30.83 Mb, respectively (Table 1, Table S2). Five of the eight pseudo-chromosomes were completely assembled without any gap; two had one gap, while one pseudo-chromosome had two gaps (Table S3). Approximately 99.78% of the Illumina reads were mapped back to the *T. dealbata* genome, with a 10-fold minimum genome coverage of 99.87% (Figure S4). The genome assembly had an overall BUSCO score of 97.52% (Table S4). Approximately 96.34% of RNA-seq reads could be successfully aligned to the genome (Table S5). In addition, we found that all pseudo-chromosomes of the *T. dealbata* genome contained (TTAGGG)n telomeres at both ends (Table S6) and centromeres in the central region (Table S7). Overall, these data implied the *T. dealbata* genome was of high quality and completeness.

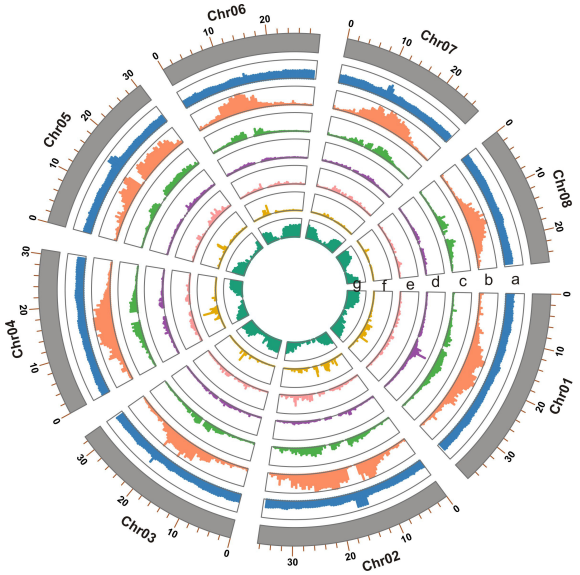


FIGURE 1
Genome features of the *T. dealbata* genome indicated in tracks from a to e: **(A)** GC content; **(B)** repeat content; **(C)** LTR density; **(D)** density of DNA transposons; **(E)** density of long interspersed nuclear elements; **(F)** density of short interspersed nuclear elements; and **(G)** gene density. All features are presented in non-overlapping windows of 500 kb.

TABLE 1 Summary statistics for the *T. dealbata* assembly.

Genomic feature	Value
Estimated genome size (Mb)	256.15
Length of genome assembly (Mb)	255.05
Number of scaffolds	78
Longest scaffold (Mb)	37.82
Scaffold N50 (Mb)	30.83
Number of contigs	84
Longest contig (Mb)	37.82
Contig N50 (Mb)	29.80
Number of pseudo-chromosomes	8
Sequences anchored to pseudo-chromosomes (%)	98.77
Number of gaps	4
Numbers of gene models	24,780
Mean transcript length (bp)	3,352.15
Mean coding sequence length (bp)	1,289.94
Total size of repeat sequences (Mb)	100.35

3.2 Genome features

A total of 100.35 Mb of repeat sequences representing 39.34% of the *T. dealbata* genome were predicted from the high-quality assembly (Table S8). The repeat content of *T. dealbata* was lower than that of other sequenced Zingiberales species (Figure S5). The LTR-RTs were the most abundant repeat class, with 31.54 Mb (12.36%) of the genome and a *Gypsy* to *Copia* ratio of 4.4:1. Within the repeat-masked genome, 24,780 high confidence protein-coding genes covering 92.50% of the complete BUSCO genes were predicted (Table S4).

In addition, 24,020 (96.93%) gene models were assigned to known functions using at least one of the protein databases, with 15,611 (63.00%) assigned to GO terms (Table S9). At the same time, 13,679 ncRNAs with a total length of 2.09 Mb, including 5,358 rRNAs, 7,647 tRNAs, 152 miRNAs, and 522 snRNAs, were identified (Table S10).

3.3 Genome evolution

A total of 21,717 *T. dealbata* genes were classified into 13,613 families, 2,289 (16.81%) of which were located in the single-copy orthogroups across the six plant species (Table S10). A phylogenetic tree reconstructed based on the SCGs revealed that *T. dealbata* had the closest genetic relationship with *Z. officinale* (Figure 2). The divergence between *T. dealbata* and *Z. officinale* was estimated to be around 55.41 MYA. In addition, 48 and 52 gene families were significantly ($P < 0.01$) expanded and contracted in the *T. dealbata* genome, respectively, and 309 gene families containing 1,161 genes were specific to *T. dealbata* (Table S11). The 439 genes within the significantly expanded gene families were highly enriched in GO

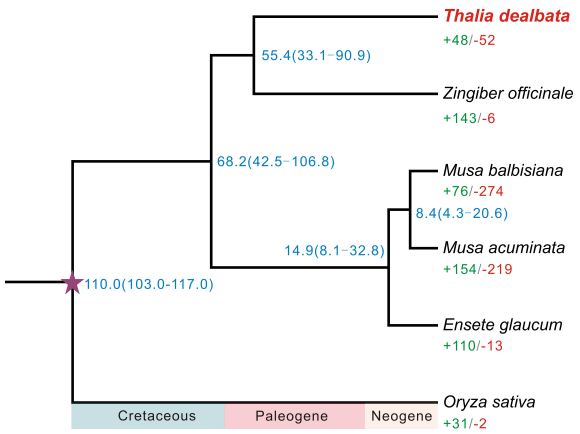


FIGURE 2
Phylogenetic tree of six plant species and evolution of gene families. The fully supported species tree was reconstructed based on 2,289 SCGs across the six species. Blue number beside each node shows the estimated divergence time of each node with 95% confidence intervals. The purple pentagram represent the calibration point. Green and red numbers indicate the numbers of significantly ($P < 0.01$) expanded and contracted gene families per species.

terms related to "glutathione metabolic process", "response to wounding", "photosynthesis, light reaction", and "defense response" (Figure S6), which possibly contributes to *T. dealbata* adaption to wetland environments. In addition, the *T. dealbata* specific genes were functionally enriched in "intracellular transport", "response to hormone", "cell wall modification", and "glycerolipid biosynthetic process" (Figure S7).

Furthermore, the WGD analysis revealed no recent WGD events in the *T. dealbata* genome (Figure 3A), although most (14,734; 59.46%) of the *T. dealbata* genes were classified as the WGD-derived genes (Table S12). However, the distribution of synonymous substitution rate (*Ks*) showed that *T. dealbata* and *M. acuminata* shared an WGD event in their common ancestor. This ancient WGD event was also supported by the 2:2 relationship of the synteny blocks between *T. dealbata* and *M. acuminata* (Figure S8) and the 1:1 relationship of the synteny blocks within *T. dealbata* (Figure S9).

We detected a total of 764 intact LTR-RTs in the *T. dealbata* genome, all of which were inserted after the split of *T. dealbata* from *M. acuminata* (Figure S10). The *T. dealbata* genome contained significantly more intact LTR-RTs than the *M. acuminata* (764 vs 278), although *T. dealbata* had a smaller genome size than *M. acuminata* (255 Mb vs 473 Mb).

Finally, 3,863 SCGs were identified among *T. dealbata*, *M. acuminata*, and *O. sativa*, of which 1,017 genes were positively selected in *T. dealbata*. These positively selected genes were mainly related to "RNA modification", "translation", "lipid metabolic process", "RNA methylation", and "plant organ development" (Figure 3B).

4 Discussion and conclusion

In this study, we performed deep sequencing and chromosome-level genome assembly of *T. dealbata*, an emergent wetland plant belonging to Marantaceae from the order Zingiberales. Based on high coverage PacBio and Hi-C reads, we assembled a near-complete genome assembly of *T. dealbata*, the first reported assembly in Marantaceae. This *T. dealbata* assembly has considerably high completeness and continuity, with most of the pseudo-chromosomes were completely assembled. The high quality of this genome indicated the advantages of PacBio HiFi sequencing in constructing highly continuous genome assemblies with long and accurate reads (Nurk et al., 2020; Wang M. et al., 2022). However, there are still one to two gaps in three pseudo-chromosomes, which might be caused by the species-specific complex repetitive regions in *T. dealbata* genome. Further integration of ultra-long reads and other high-throughput sequencing data will make it possible to generate a telomere-to-telomere (T2T) genome for *T. dealbata*.

Using Hi-C scaffolding analysis, we anchored the genome sequences of *T. dealbata* into eight pseudo-chromosomes, showing the different chromosome number from a previous study (2n = 2x = 12; Suessenguth, 1921). The 3D-DNA software does not require *a priori* chromosome number as input, and the Hi-C contact map shows a clear pattern of eight chromosome interaction (Figure S3). Thus, we speculated that there was some mistake in the relatively old research of Suessenguth (1921). Future karyotype analysis of *T. dealbata* can verify the validity of our speculation.

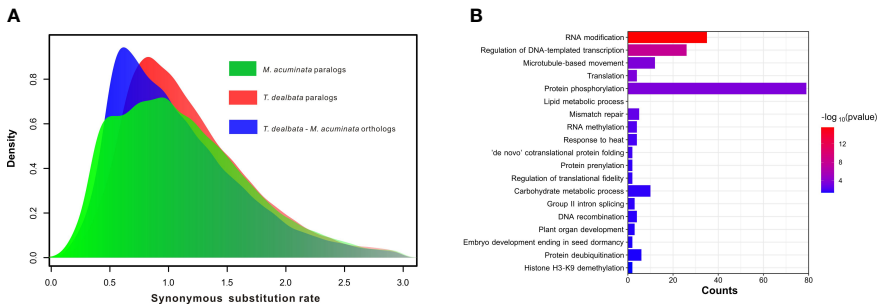


FIGURE 3
WGD event and positive selection in *T. dealbata*. **(A)** Results of recent WGD analysis based on the paralog and ortholog gene pairs within and between *T. dealbata* and *M. acuminata*. **(B)** GO enrichment of the 1,017 positively selected in *T. dealbata*..

The high-quality *T. dealbata* genome has lead to accurate structural annotation of pretein-coding genes and ncRNAs, enabling us to gain further insights into the evolutionary history of *T. dealbata* and related species. We found that *T. dealbata* had the closest genetic relationship with *Z. officinale*, which was consistent with the close relationship between Marantaceae and Zingiberaceae that was revealed by two previous studies (Sass et al., 2016; Carlsen et al., 2018). These two species shared an ancient WGD event with *M. acuminata* in their common ancestor, which was follwed by diploidization events that involved substantial genome reshuffling and gene losses. The early divergence among these species provided a long enough period to allow sufficient divergence in genomic characteristics and adaptation strategies in *T. dealbata* and *Z. officinale*. The identified expanded, contracted, and unique gene families together with a number of positively selected genes in *T. dealbata* genome are possibly responsible to the adaptation of *T. dealbata* to wetland environment.

Overall, the high-quality *T. dealbata* genome assembly presented in this study will provide a valuable genomic resource for the study of plant adaptation to wetland environments and the evolutionary analysis of Marantaceae and Zingiberales. We look forward more genetic and genomic analysis and functional studies of this interesting wetland plant in the future.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

MT, JH, LW, and HL designed and supervised the study. MT and JD collected the samples and extracted the genomic DNA and RNA. MT, XM, and PG performed genomic data analysis. MT drafted the manuscript, and YB revised this manuscript. All authors contributed to the article and approved the submitted version.

Funding

The authors declare that this study received funding from Kunming Novo Medical Laboratory Co., Ltd. The funder had the following involvement in the study: designed and supervised the study, technical guidance, genome sequencing and assembly, analysis, interpretation of data and the writing of this article. All authors agreed to submit it for publication.

This research was also funded by the research on the collection and identification of forage resources and productive cultivation techniques in Yunnan.

Acknowledgments

We thank the colleagues of College of Landscape Architecture and Horticulture, Yunnan Agricultural University for their generous help, College of Animal Science and Technology, Yunnan Agricultural University for providing a well-founded experimental platform, and Kunming Novo Medical Laboratory Co., Ltd. for financial support.

Conflict of interest

Authors HL and LW were employed by the company Kunming Novo Medical Laboratory Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1183361/full#supplementary-material

References

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45–48. doi: 10.1093/nar/28.1.45

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and genomewise. *Genome Res.* 14, 988–995. doi: 10.1101/gr.1865504

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinf.* 10, 421. doi: 10.1186/1471-2105-10-421

Carlsen, M. M., Fér, T., Schmickl, R., Leong-Škorničková, J., Newman, M., and Kress, W. J. (2018). Resolving the rapid plant radiation of early diverging lineages in the tropical zingiberales: pushing the limits of genomic data. *Mol. Phylogenet. Evol.* 128, 55–68. doi: 10.1016/j.ympev.2018.07.020

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552. doi: 10.1093/oxfordjournals.molbev.a026334

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18, 170–175. doi: 10.1038/s41592-020-01056-5

Cronk, J. K., and Fennessy, M. S. (2016). Wetland plants: biology and ecology. *CRC press*. doi: 10.1201/9781420032925

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271. doi: 10.1093/bioinformatics/btl097

Deng, N., Liu, C., Tian, Y., Song, Q., Niu, Y., and Ma, F. (2021). Complete chloroplast genome sequences and codon usage pattern among three wetland plants. *Agron. J.* 113, 840–851. doi: 10.1002/agj2.20499

D'hont, A., Denoeud, F., Aury, J. M., Baurens, F. C., Carreel, F., Garsmeur, O., et al. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 488, 213–217. doi: 10.1038/nature11241

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytoch. Bull.* 19, 11–15.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). *De novo* assembly of the *Aedes aegypti* genome using Hi-c yields chromosome-length scaffolds. *Science* 356, 92–95. doi: 10.1126/science.aal3327

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., et al. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-c experiments. *Cell Syst.* 3, 95–98. doi: 10.1016/j.cels.2016.07.002

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). *LTRharvest*, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinf.* 9, 1–14. doi: 10.1186/1471-2105-9-18

Gray, M. J., Hagy, H. M., Nyman, J. A., and Stafford, J. D. (2013). "Management of wetlands for wildlife," in *Wetland techniques*. Eds. J. Anderson and C. Davis (Dordrecht: Springer). doi: 10.1007/978-94-007-6907-6_4

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, J. R.K., Hannick, L. I., et al. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666. doi: 10.1093/nar/gkg770

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9, R7. doi: 10.1186/gb-2008-9-1-r7

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res.* 37, D211–D215. doi: 10.1093/nar/gkn785

Jain, R., Jenkins, J., Shu, S., Chern, M., Martin, J. A., Copetti, D., et al. (2019). Genome sequence of the model rice variety KitaakeX. *BMC Genomics* 20, 1–9. doi: 10.1186/s12864-019-6262-4

Katoh, K., Misawa, K., Kuma, K.-I., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* 34, 1812–1819. doi: 10.1093/molbev/msx116

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

Li, H. L., Wu, L., Dong, Z., Jiang, Y., Jiang, S., Xing, H., et al. (2021). Haplotype-resolved genome of diploid ginger (*Zingiber officinale*) and its unique gingerol biosynthetic pathway. *Hortic. Res.* 8, 189. doi: 10.1038/s41438-021-00627-7

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964. doi: 10.1093/nar/25.5.955

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35, W182–W185. doi: 10.1093/nar/gkm321

Nawrocki, E. P., and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29, 2933–2935. doi: 10.1093/bioinformatics/btt509

Nurk, S., Walenz, B. P., Rhie, A., Vollger, M. R., Logsdon, G. A., Grothe, R., et al. (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* 30, 1291–1305. doi: 10.1101/gr.263566.120

Ou, S., and Jiang, N. (2018). LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 176, 1410–1422. doi: 10.1104/pp.17.01310

Pezeshki, S. R. (2001). Wetland plant responses to soil flooding. *Environ. Exp. Bot.* 46, 299–312. doi: 10.1016/S0098-8472(01)00107-1

Podlevsky, J. D., Bley, C. J., Omana, R. V., Qi, X., and Chen, J. J. L. (2007). The telomerase database. *Nucleic Acids Res.* 36, D339–D343. doi: 10.1093/nar/gkm700

Price, A. L., Jones, N. C., and Pevzner, P. A. (2005). *De novo* identification of repeat families in large genomes. *Bioinformatics* 21, i351–i358. doi: 10.1093/bioinformatics/bti1018

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., et al. (2019). Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biol.* 20, 38. doi: 10.1186/s13059-019-1650-2

Sass, C., Iles, W. J., Barrett, C. F., Smith, S. Y., and Specht, C. D. (2016). Revisiting the zingiberales: using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. *PeerJ* 4, e1584. doi: 10.7717/peerj.1584

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–W439. doi: 10.1093/nar/gkl200

Suessenguth, K. (1921). Bemerkungen zur meiotischen und somatischen kernteilung bei einigen monokotylen. *Flora oder Allgemeine Botanische Zeitung* 114, 313–328. doi: 10.1016/S0367-1615(17)31551-3

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinf.* 25, 4–10. doi: 10.1002/0471250953.bi0410s05

Wang, M., Huang, J., Liu, S., Liu, X., Li, R., Luo, J., et al. (2022). Improved assembly and annotation of the sesame genome. *DNA Res.* 29, dsac041. doi: 10.1093/dnares/dsac041

Wang, J., Lu, X., Zhang, J., Ouyang, Y., Wei, G., and Xiong, Y. (2020). Rice intercropping with alligator flag (*Thalia dealbata*): a novel model to produce safe cereal grains while remediating cadmium contaminated paddy soil. *J. Hazard. Mater.* 394, 122505. doi: 10.1016/j.jhazmat.2020.122505

Wang, Z., Miao, H., Liu, J., Xu, B., Yao, X., Xu, C., et al. (2019). *Musa balbisiana* genome reveals subgenome evolution and functional divergence. *Nat. Plants* 5, 810–821. doi: 10.1038/s41477-019-0452-6

Wang, Z., Rouard, M., Biswas, M. K., Droc, G., Cui, D., Roux, N., et al. (2022). A chromosome-level reference genome of *Ensete glaucum* gives insight into diversity and chromosomal and repetitive sequence evolution in the musaceae. *GigaScience* 11, giac027. doi: 10.1093/gigascience/giac027

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). *MCScanX*: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49–e49. doi: 10.1093/nar/gkr1293

Wang, M., Tong, S., Ma, T., Xi, Z., and Liu, J. (2021). Chromosome-level genome assembly of sichuan pepper provides insights into apomixis, drought tolerance, and alkaloid biosynthesis. *Mol. Ecol. Resour.* 21, 2533–2545. doi: 10.1111/1755-0998.13449

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yue, J., Chen, Q., Wang, Y., Zhang, L., Ye, C., Wang, X., et al. (2023). Telomere-to-telomere and gap-free reference genome assembly of the kiwifruit *Actinidia chinensis*. *Hortic. Res.* 10, uhac264. doi: 10.1093/hr/uhac264

Zedler, J. B., and Kercher, S. (2005). Wetland resources: status, trends, ecosystem services, and restorability. *Annu. Rev. Environ. Resour.* 30, 39–74. doi: 10.1146/annurev.energy.30.050504.144248

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 14 August 2023
DOI 10.3389/fpls.2023.1068796

OPEN ACCESS

EDITED BY
Kai-Hua Jia,
Shandong Academy of Agricultural Sciences, China

REVIEWED BY
Liang Xiao,
Beijing Forestry University, China
Juan Li,
Shanxi Normal University, China

*CORRESPONDENCE
Jinhui Chen
jinhuichen@hainanu.edu.cn

RECEIVED 13 October 2022
ACCEPTED 17 May 2023
PUBLISHED 14 August 2023

CITATION
Chen J, Liu M, Meng X, Zhang Y, Wang Y, Jiao N and Chen J (2023) Multiomics studies with co-transformation reveal microRNAs via miRNA-TF-mRNA network participating in wood formation in *Hevea brasiliensis*. *Front. Plant Sci.* 14:1068796. doi: 10.3389/fpls.2023.1068796

Multiomics studies with co-transformation reveal microRNAs via miRNA-TF-mRNA network participating in wood formation in *Hevea brasiliensis*

Jinhui Chen[1,2]*, Mingming Liu[1,3], Xiangxu Meng[1,2], Yuanyuan Zhang[4,5], Yue Wang[1,2], Nanbo Jiao[1,2] and Jianmiao Chen[1,3]

[1]Sanya Nanfan Research Institute of Hainan University, Hainan Yazhou Bay Seed Laboratory/Key Laboratory of Genetics and Germplasm Innovation of Tropical Special Forest Trees and Ornamental Plants, Ministry of Education, School of Forestry, Hainan University, Sanya, China, [2]Engineering Research Center of Rare and Precious Tree Species in Hainan Province, School of Forestry, Hainan University, Haikou, China, [3]School of Tropical Crops, Hainan University, Haikou, China, [4]Rubber Research Institute, Chinese Academy of Tropical Agricultural Sciences, Haikou, Hainan, China, [5]State Centre for Rubber Breeding, Haikou, Hainan, China

Introduction: MicroRNAs (miRNAs) are small endogenous non-coding RNAs that play an important role in wood formation in plants. However, the significance of the link between miRNAs and their target transcripts in wood formation remains unclear in rubber tree (*Hevea brasiliensis*).

Methods: In this study, we induced the formation of reaction wood by artificially bending rubber trees for 300 days and performed small RNA sequencing and transcriptome deep sequencing (RNA-seq) to describe the complement of miRNAs and their targets contributing to this process.

Results and discussion: We identified 5, 11, and 2 differentially abundant miRNAs in normal wood (NW) compared to tension wood (TW), in NW relative to opposite wood (OW), and between TW and OW, respectively. We also identified 12 novel miRNAs and 39 potential miRNA-mRNA pairs with different accumulation patterns in NW, TW, and OW. We noticed that many miRNAs targeted transcription factor genes, which were enriched in KEGG pathways associated with phenylpropanoid biosynthesis, phenylalanine metabolism, and pyruvate metabolism. Thus, miRNA-TF-mRNA network involved in wood formation via tension wood model were constructed. We validated the differential accumulation of miRNAs and their targets by RT-qPCR analysis and overexpressed miRNA in *Nicotiana benthamiana* with its potential target gene. These results will provide a reference for a deep exploration of growth and development in rubber tree.

KEYWORDS

reaction wood, phenylpropanoid biosynthesis pathway, lignin biosynthesis, *Hevea brasiliensis*, miRNA

Introduction

Rubber tree (*Hevea brasiliensis*) is the main source of natural rubber (NR). Once latex production is no longer economically viable, rubber trees are also used as timber. The wood from rubber trees has in fact become the main export commodity in southeast Asia (Nair, 2010). Its delicate color and outstanding physical performance make it an excellent option for flooring and home furnishings. Due to its high potential commercial value, increasing timber yield and quality has become a key point of biotechnology in the rubber tree industry (Priyadarshan, 2017). However, two major limitations hinder a more widespread use of rubber wood: (i) The extent of unlignified or partially lignified tension wood fiber, which is not easily digested by enzymes, is high, and the proportion of normal fibers is low (Mellerowicz and Gorshkova, 2012; Gritsch et al., 2015); and (ii) it has high sensitivity to biodegradation owing to low levels of phenolic compounds with biocidal activity (Pramod et al., 2019; Thaochan et al., 2020). The biosynthesis of lignin and polyphenolic derivatives in living trees, particularly in rapidly growing woody plants such as rubber trees, contributes to wood quality and durability (Kuyyogsuy et al., 2018; Pramod et al., 2019; Thaochan et al., 2020).

Lignin is an abundant biopolymer that is essential for plant cell wall integrity and stem strength (Shen et al., 2012). The biosynthesis of lignin monomers begins with phenylalanine deamination, leading to the production of three monolignin alcohols: coniferyl, sinapyl, and *p*-coumaryl alcohols (Eudes et al., 2012). Several genes that participate in lignin biosynthesis and modulate lignin levels have been identified in dicots (Sibout et al., 2005; Weng et al., 2010). For instance, knocking down *4-coumarate:CoA ligase* (*4CL*) expression in hybrid poplar (*Populus tremula* × *P. alba*) resulted in a sharp decrease in lignin content and markedly altered wood chemical composition and wood metabolism (Voelker et al., 2010). The *PtrWND* (*wood-associated NAC domain*) genes were shown to induce the expression of wood biosynthetic genes, including associated structural genes and transcription factors, resulting in ectopic deposition of lignin in poplars (Zhong et al., 2010). Despite the above progress in our understanding of lignin biosynthesis, much remains to be investigated in terms of transcriptional and post-transcriptional regulation. Recently, regulation of wood formation via non-coding RNAs (ncRNAs) and microRNAs (miRNAs) has received increasing attention (Lu et al., 2013).

miRNAs are post-transcriptional modulators of gene function by promoting the cleavage of their complementary target messenger RNAs (mRNAs) and/or imposing translation repression (Bartel, 2004; Zhang et al., 2019). Several studies have illustrated the vital roles played by miRNAs in wood formation (Li et al., 2020; Yu et al., 2020). For instance, transcript levels for 17 of the 29 *LACCASE* (*PtrLAC*) genes in the black cottonwood (*P. trichocarpa*) genome decreased in *P. trichocarpa* trees overexpressing miR397a, in turn leading to a reduction of lignin levels (Lu et al., 2013). Similarly, the overexpression of *miR319a* in *Populus tomentosa* in seedlings resulted in delayed secondary growth and decreased xylem production (Hou et al., 2020). In particular, *miR165b* guides the development of pith secondary cytoderm is by restraining the *AtHB15* (*HOMEOBOX 15*) expression domain (Du and Wang, 2015). However, how wood anatomical features like container shape and thickness are genetically governed is not very clear. These traits are important for cell wall composition and overall tree performance (Quan et al., 2018; Quan et al., 2019).

To explore the molecular basis of changes in transcript levels caused by miRNAs during wood formation, we sequenced small RNAs from three wood parts: normal wood (NW)which is from the stem of the tree at breast height, tension wood (TW) which is upper side of the bending trunk, and opposite wood (OW) which is lower side of the bending trunk. We then assigned predicted target genes to these wood-abundant miRNAs by identifying genes whose transcript levels were inversely correlated with miRNA abundance. We complemented this approach by using a bioinformatics method for miRNA target prediction and constructed the resulting miRNA–mRNA interaction network. The small RNAs found in this study are good candidates for miRNAs that are involved in wood formation. They may also help with the development of functional markers for molecular breeding in rubber trees and other tropical plants to help change the composition of lignin or physical characteristics.

Materials and methods

Plant materials and microscopy observations

The rubber trees (clone Reyan 7-33-97) used in this study were grown in the experimental greenhouse of Hainan University (Danzhou, Hainan, China; 109°29′25″ E, 19°30′40″ N) at the end of June 2016. To probe the genes involved in the formation of reaction wood, three rubber trees of similar age and with trunks of similar diameter (about 2 cm) were bent at a 30° angle for 300 days and were selected as experimental materials to force the formation of reaction wood (Figure S1). The bending was applied starting at 9 a.m. on August 17th, 2020, and ended at 9 a.m. on June 13rd, 2021, at which point the wood samples were rapidly processed. The wood quality from the collected samples was assessed by scanning electron microscopy (SEM, Phenom proX, the Netherlands) in Center for Analytical Instrumentation (Hainan University), and the resulting images were processed in ImageJ software to measure the area of the gelatinous (G) layer. The SEM images indicated that the TW (tension wood) reaction wood had an extremely dense cementitious layer (G-layer; Figure S2A) that was not present in NW (normal wood) or OW (opposite wood; Figures S2B, C). These observations were consistent with earlier findings (Sujan et al., 2015). Therefore, xylem samples from the collected trees were selected for further analysis.

Samples were collected for NW, TW, and OW from the same individuals to allow for direct comparison in an identical genetic background. Briefly, the bark above the sampling area was removed to expose the inner wood layers. A sharp razor blade was then used to collect TW (upper side) and OW (lower side) from the same branch (Li et al., 2013). The control for stem xylem tissue was NW and was collected about 1 m above the ground, before the bending point. All collected samples were about 2 cm × 1 cm and 4–5 mm in

depth. Samples were harvested in the morning, quickly frozen in liquid nitrogen, and stored at –80°C until use.

RNA extraction and qualification

Total RNA was extracted from nine samples (NW1, NW2, NW3, OW1, OW2, OW3, TW1, TW2, and TW3) using a modified cetyl trimethyl ammonium bromide (CTAB) method (Chang et al., 1993). Each tree counted as one biological replicate (NW1-3, TW1-3, and OW1-3). Traces of genomic DNA were removed with RNase-free DNase I digestion (Takara, Beijing, China). RNA degradation and DNA contamination were assessed by electrophoresis on 1% (w/v) agarose gels. RNA concentration, quality, and integrity were estimated on a NanoPhotometer spectrophotometer (IMPLEN, CA, USA) and an RNA Nano 6000 Chip on a Bioanalyzer 2100 (Agilent Technologies, CA, USA). Only RNA samples with an $OD_{260/280}$ ratio of 1.9–2.2, an $OD_{260/230}$ ratio ≥ 2.0, and RNA integrity number (RIN) values > 6.8 were processed for further experiments.

Small RNA library construction and sequencing

Three micrograms of total RNA per sample was used to construct sequencing libraries with a NEBNext multiplex small RNA library prep kit for Illumina (NEB, USA) following the manufacturer's protocol. Briefly, the NEB 3′ SR adapter was ligated to the 3′ end of miRNAs, siRNAs (small interfering RNAs), and piRNAs (piwi-interacting RNAs). The SR real-time primer was then annealed to the 3′ SR adapter to initiate double-stranded DNA (dsDNA). The 5′ end adapter was connected to the 5′ ends of miRNAs. First-strand cDNA synthesis with M-MuLV Reverse Transcriptase (RNase H-). The libraries were amplified by 35 PCR cycles with index (X) primer, SR primer for Illumina, and LongAmp Taq 2X master mix. The PCR products were separated on 8% polyacrylamide gel for 80 min at 100 V. DNA fragments responding to 140-160 bp (the length of sRNAs with 5′ and 3′ adapters) were purified from the gel and eluted in 8 μL of elution buffer. Library quality and titer were evaluated using a DNA high sensitivity chip and Agilent Bioanalyzer 2100 instrument. A TruSeq SR Cluster Kit v3-cBot-HS (Illumina) was used for cluster formation on a cBot cluster generation system following the manufacturer's instructions. Libraries were sequenced on an Illumina HiSeq 2,500 platform as 50-bp single-end reads.

Identification of known miRNAs and novel miRNAs

After adapter trimming from the raw reads, any clean reads shorter than 18 nt or low-quality reads (reads with N bases > 10% and reads with a 3′ end with Q < 20 [$Q = -10Log_{10}^{error\_ratio}$]) were discarded. The remaining clean reads were mapped to the rubber tree reference genome (Liu et al., 2020) with Bowtie2 allowing no mismatch (Langmead et al., 2009). Reads derived from rRNAs, protein-coding genes, snRNAs, tRNAs, snoRNAs, and repeat sequences were removed by filtering the clean reads against the Rfam database and RepeatMasker (Tempel, 2012; Kalvari et al., 2018). Known miRNAs were identified with miRBase 20.0 (Meng et al., 2018); potential miRNAs and their secondary structures were determined using the miRDeep2 algorithm (Friedländer et al., 2012) and sRNA-tools-cli (http://srna-workbench.cmp.uea.ac.uk/). The formation of a typical hairpin structure was used as a criterion to identify novel miRNAs from the transcripts showing no match to known miRNAs. Unannotated small RNAs were assessed with miRDeep2 (Friedländer et al., 2012) and miREvo (Wen et al., 2012) using minimum free energy, possible Dicer cleavage sites, and the secondary structures of small RNA tags. Custom scripts were applied to count all candidate miRNAs and estimate the basic deviation between each position of all identified miRNAs and the first position of identified miRNAs of a given length. The secondary structure of the novel_28 mature sequence was predicted and compared to other miRNA families from rubber tree and other species.

Prediction of the target genes of miRNAs

The RNA-seq data were from our previous study (Meng et al., 2021). The online tool psRNATarget (https://www.zhaolab.org/psRNATarget/) was used to predict miRNA targets with default parameters with expectation ≤ 3 (Dai et al., 2018). A gene was deemed to be a putative target for a miRNA when its transcript levels showed a negative Pearson's correlation coefficient with miRNA abundance (correlation < –0.8, *P*-value < 0.05). miRNA abundance was estimated with the formula (Zhou et al., 2010): Normalized abundance = mapped read count/total reads * 1,000,000.

KEGG enrichment analysis of co-expressed target genes

The predicted co-expressed target genes were subjected to KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway. KEGG (http://www.genome.jp/kegg/) pathway enrichment analysis was carried out with KOBAS software (Mao et al., 2005).

Construction for miRNA–mRNA interaction networks and tissue-specific expression analysis

The Pearson's correlation coefficients between miRNAs and transcription factor genes were calculated using expression values from this study for miRNAs and from our previous study (Meng et al., 2021) for mRNAs or between transcription factor genes and other genes to identify miRNAs, transcription factor genes, and their co-expressed genes. The resulting interaction network was built in R (version 4.0.1) and visualized with Cytoscape (version 3.8.1).

Plasmid construction

According to the target sites (Figure 1A), to overexpress novel_28 (*hbr-miR482c*), the *pre-miR482c* sequence was amplified from genomic DNA isolated from normal wood tree with primers containing BamH I and Sac I restriction sites. The resulting PCR product was digested with BamH I and Sac I and ligated into the vector pBI121 downstream of the cauliflower mosaic virus (CaMV) 35S promoter (Shanghai Generay Biotech; Figure 1B). The full-length coding sequence of *HbrCAD1* (GH714_013930) was also cloned into pBI121 (GH714_013930-pBI121; Figure 1B). Both constructs were transformed into *Agrobacterium* (*Agrobacterium tumefaciens*) strain GV3101. The primer sequences used for cloning are listed in Table S1.

Verification of interaction between *hbr-miR482c* and its target

Agrobacterium-mediated transient expression in *Nicotiana benthamiana* leaves (English, 1997) was used to assess the targeting of *HbrCAD1* transcripts by hbr-miR482c. Agrobacterium cultures harboring the *hbr-miR482c* or HbrCAD1 construct were resuspended in infiltration buffer, mixed, and infiltrated into six *N. benthamiana* leaves (Figure 1C) from 3-week-old plants. As negative controls, Agrobacterium cultures a mix of cultures harboring pB121 and the 35S::HbrCAD1 construct were infiltrated in N. benthamiana leaves.

Validation of miRNA expression and that of their target genes by RT-qPCR

A GoScript™ Reverse Transcription kit (Promega, USA) was used for first-strand cDNA synthesis from total RNA of the nine samples collected in this study. Six differentially expressed miRNAs (novel_47, novel_67, novel_28, novel_165, novel_177, and novel_101) and six target genes (*ALDO1*, *PEPcase*, *CAD1*, *RF2*, *PKc_like*, *GT*) of miRNA-mRNA correlation network were chosen (Figure S3). Gene-specific primers were designed for the target genes with Primer Premier v5 software (Table S1). qPCR was performed with TB Green Premix Ex Taq II (Tli RNase H Plus; Takara, Beijing, China) for the target genes according to the manufacturer's instructions. PCR conditions were as follows: denaturation at 94°C for 2 min, then 40 cycles of 95°C for 5 s and 60°C for 30 s. *Ubiquitin* was used as internal reference for normalization of expression data of rubber tree samples (Meng et al., 2022), and *β-actin* was used for *N. benthamiana* (Nawaz et al., 2019). For miRNA, a miRNA RT-qPCR Detection Kit (Aidlab, Beijing, China) was used following the manufacturer's instructions. PCR conditions were as above. *U6* transcripts were used as internal reference for normalization (Zeng et al., 2009). A melting curve was performed from 60°C to 95°C to confirm the specificity of the amplicons. Relative expression levels of miRNAs and their target genes were estimated by the $2^{-\triangle\triangle Ct}$ method (Schmittgen and Livak, 2008). Three technical replicates were analyzed per sample.

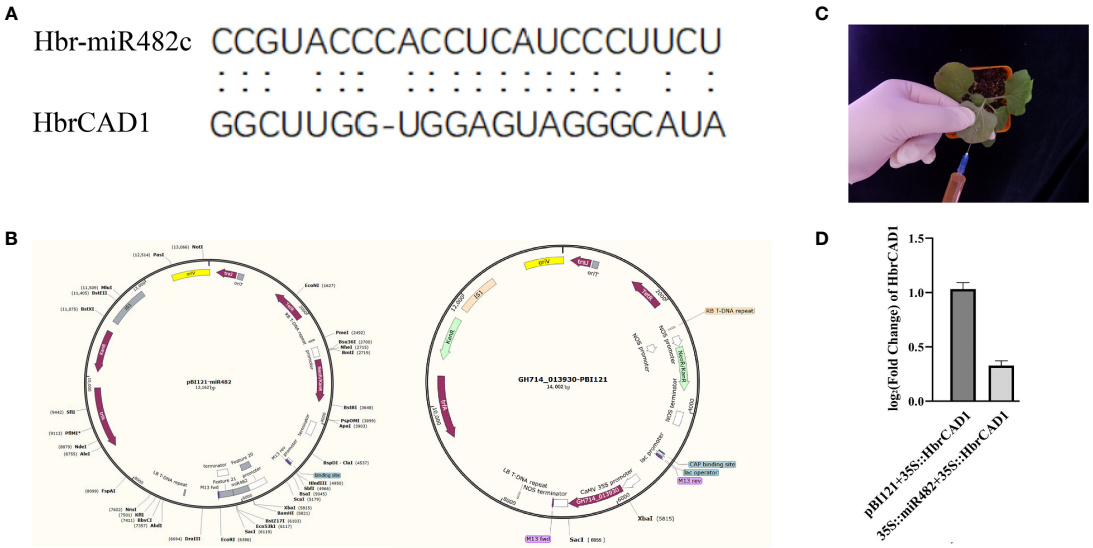


FIGURE 1
Validation of the targeting of *HbrCAD1* by its predicted miRNA novel_28 in *Nicotiana benthamiana*. *β-actin* was selected as internal reference; data are means ± standard error of three independent biological replicates. **(A)** Alignment of *hbr-miR482c* and *HbrCAD1*. **(B)** The plasmids pBI121-miR482c and GH714-013930-pBI121 used for the assay. **(C)** Principle of transient infiltration of *N. benthamiana* leaves with an Agrobacterium cell suspension. **(D)** Relative *HbrCAD1* expression levels in different samples. Data are shown as Log2 (fold-change), with the expression of *HbrCAD1* from the pair pBI121 + *35S:HbrCAD1* set to 1.

10.3389/fpls.2023.1068796

TABLE 1 Summary of reads from small RNA sequencing.

Sample	Raw reads	Clean reads	Total sRNA	Mapped sRNA	Mapping ratio (%)
NW1	10,891,608	10,601,788	8,002,438	7,378,835	92.21%
NW2	14,495,357	14,353,830	11,160,172	9,992,003	89.53%
NW3	12,182,526	11,943,571	9,115,657	8,359,523	91.71%
TW1	10,151,849	9,858,994	6,329,046	5,936,751	93.80%
TW2	12,222,618	11,860,937	8,982,250	8,115,462	90.35%
TW3	13,032,187	12,697,455	9,391,865	8,629,289	91.88%
OW1	13,364,645	12,973,816	10,208,783	9,422,696	92.30%
OW2	11,209,112	10,839,323	6,835,678	6,268,583	91.70%
OW3	12,370,905	12,040,308	8,412,275	7,664,797	91.11%

Results

Overview of small RNA sequencing from reaction wood

We constructed nine small RNA libraries from different wood tissues (NW, TW, and OW) to explore the role of miRNAs in wood development in rubber trees. We obtained between 10.15 and 14.50 million raw reads after sequencing. We removed low-quality reads and removed adapters to yield 9.86 to 14.35 million clean reads with a length ranging from 18 to 30 nucleotides (nt) (Table 1).

We aligned the clean reads against the rubber tree reference genome with Bowtie2 and used RSEM software to estimate read counts per gene model. We successfully mapped from 5,936,751 to 9,992,003 reads to the rubber tree genome, or 89.53%-93.80% of all clean reads across the nine small RNA libraries (Table 1). We then removed all reads mapping to tRNA (transfer RNA), snoRNA (small nucleolar RNA) and snRNA (small nuclear RNA) loci. We focused on the remaining unannotated reads. We observed a peak for 21-nt sRNAs, representing >14% of all candidates unannotated sRNAs in the libraries, with a second peak for 24-nt sRNAs (from 10% to 13%; Figure 2).

Annotation of known and novel miRNAs expressed in reaction wood

We compared the unannotated sRNAs identified above to all miRNAs or their *pre-miRNA*s deposited in miRBase to identify known and novel miRNAs. We identified 22 (NW), 22 (TW), and 23 (OW) known miRNAs belonging to 21 miRNA families (Figure 3A). Of these, 19 miRNAs were shared across the NW, TW, and OW libraries (Figure 3B; Table S2). We determined that the first base of these known miRNAs tended to be a uracil (U) for 18–22-nt miRNAs (Figure S4).

The known miRNA *hbr-miR166*, *hbr-miR396*, *hbr-miR408*, and *hbr-miR482* families were each represented by two members, while the remaining miRNA families only had one member (Figure S5). Furthermore, two miRNAs (*hbr-miR6170* and *hbr-miR6171*) were specific to OW (Figure S5), suggesting that each wood tissue is associated with a slightly different set of miRNAs.

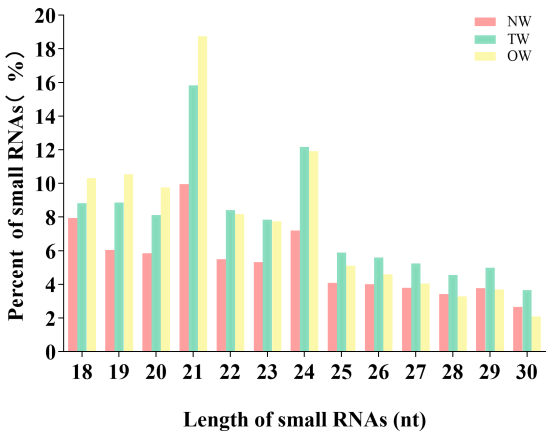


FIGURE 2
Length distribution of miRNA sequences sequenced from *Hevea brasiliensis* reaction wood.

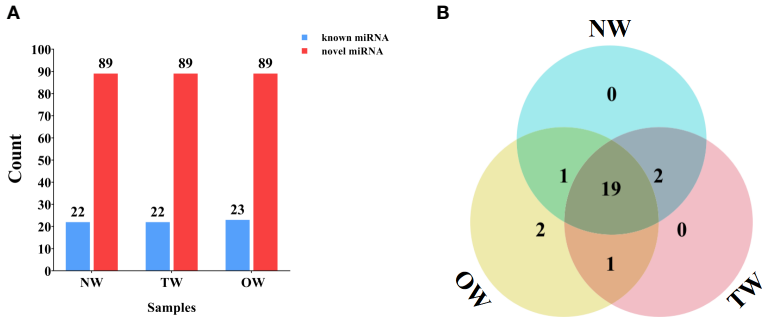


FIGURE 3
Numbers of miRNAs identified in different xylem tissues. **(A)** Known and novel miRNAs identified from small RNA sequencing of NW, TW, and OW from 300-day rubber tree reaction wood. **(B)** Venn diagram showing the extent of overlap between known miRNAs in each tissue type (NW, TW, and OW) from 300-day rubber tree reaction wood.

The sequences failed to match known miRNAs or *pre-miRNA*s were used to predict novel miRNA resulting in 89 novel miRNAs with typical hairpin structure (Figure 3, S6). The first base of these novel rubber tree miRNAs (18–30 nt in length) also largely started with a uracil (Figure S7). Furthermore, we cloned and assigned novel_28 to the miR482 family through BLAST search against other species in the miRBase database. We then predicted the secondary structure of novel_28 and compared its mature sequence to that of other *hbr-miR482* family members, which revealed novel_28 is a new member of the *hbr-miR482* family (Figures S6, S8). Thus, we renamed novel_28 as hbr-miR482c.

Differentially abundant miRNAs

To investigate how miRNAs might regulate wood formation, we quantified the abundance of all miRNAs as TPMs (transcripts per million), followed by a comparison of miRNA abundance between the tree wood types (|Fold change|≥2 and $P \leq 0.05$). The OW vs. NW comparison yielded 11 differentially abundant miRNAs, 5 for the TW vs. NW comparison, and 2 for the TW vs. OW comparison. Of the five differentially abundant miRNAs between TW and NW samples, three were more abundant in TW and two were more abundant in NW tissues. Similarly, 5 miRNAs were more accumulated in OW and 6 were more accumulated in NW tissues. Finally, one miRNA was more abundant in each TW and OW tissue (Table S3).

Prediction of miRNA target genes

We predicted miRNA targets through psRNATarget (Dai et al., 2018), which identifies transcripts with complementary sequences to those of miRNA candidates. We independently calculated Pearson's correlation coefficients between the abundance of each miRNA and the transcript levels of all rubber tree transcripts obtained from a previous study (Meng et al., 2021). We considered genes as candidate targets when their transcript levels were negatively correlated (<–0.8) with miRNA abundance. In the TW vs. NW comparison, we determined that 4 novel miRNAs have the potential to target 10 genes, but no clear target could be identified for the novel_86 miRNA. Similarly, we identified 31 targets for 8 of the differentially abundant miRNAs in the OW vs. NW comparison, with no predicted targets for the other three novel miRNAs (Table S4). Further, we identified 11 target genes for the 2 differentially abundant miRNAs from the comparison between TW and OW. We selected those potential targets with a negative correlation of –0.8 or below with their associated miRNA, resulting in 39 putative targets for 12 novel miRNAs (cor ≤ –0.8 and $P \leq 0.05$) (Table S5).

Identification of genes co-expressed with transcription factor genes involved in rubber tree reaction wood formation

In the above list of target genes, we noticed that 53.85% (or 21 of 39) encode transcription factors (TFs; Figure S9). To define the putative targets of the encoded TFs, we measured the Pearson's correlation coefficients (PCCs) between TF genes and all other rubber tree genes, using RNA-seq data available at the NCGC (National Genomics Data Center) under the accession numbers CRA004241 and CRA004243. A KEGG pathway enrichment analysis showed that these co-expressed targets are largely related to phenylpropanoid biosynthesis, phenylalanine metabolism, and fatty acid biosynthesis (Figure S10). These results suggest a central role for these TFs during reaction wood formation under prolonged mechanical stress.

Differentially expressed mRNA-miRNA pairs related to wood formation

We then explored how miRNA-mediated adjustment of transcripts affected reaction wood development. Here, we focused on target genes co-expressed with those TF genes that were associated with phenylalanine metabolism or carotenoid biosynthesis and constructed the underlying interaction network (Figure S10, 4). Our above analysis predicted that *hbr-miR482c* modulates the transcript levels of *HbrCAD1*, which is related to phenylpropanoid biosynthesis. Similarly, novel_67 might modulate the transcript levels of genes related to lignin biosynthesis by

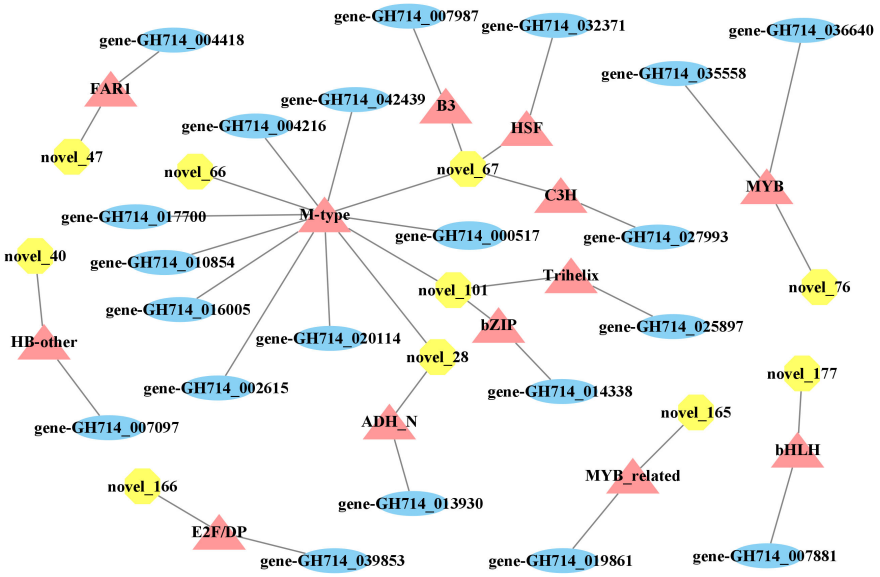


FIGURE 4
miRNA-transcription factor-mRNA networks associated with wood formation. Yellow octagons represent miRNAs, orange triangles represent transcription factor genes, and blue ellipses represent genes.

targeting C3H-type zinc finger TF family members (Figure 4); novel_93 was predicted to modulate the transcript levels of genes related to phenylalanine metabolism. In this network, *hbr-miR482c* and novel_76 each targeted two genes involved in wood formation. The C3H, FAR1 (FAR-RED IMPAIRED RESPONSE 1), bZIP, and MYB TF families possibly play vital functions in wood growth from our miRNA-TF-mRNA network. The target genes co-expressed with these TF genes, such as fatty acyl-CoA reductase 1 (*FAR1*, gene-GH714_004418), cinnamyl-alcohol dehydrogenase (*CAD1*), and p-coumarate 3-hydroxylase (*C3H*, gene-GH714_027993), were involved in carotenoid biosynthesis. We thus propose that the miRNA-TF-mRNA regulatory network described here may play a significant role in adjusting the molecular events necessary for reaction wood development in rubber tree. The expression levels of these genes and associated miRNAs are shown in Figure 5.

We wished to confirm the transcript levels of miRNAs and their target genes estimated from the RNA-seq datasets by a reverse transcription quantitative PCR (RT-qPCR) analysis of the same RNA samples. We observed a generally comparable trend in the accumulation of most miRNAs and their target genes between RT-qPCR and RNA-seq data in the OW, NW, and TW samples, although the fold-change (FC) values from RNA-seq did not precisely match the expression values obtained by RT-qPCR (Figure S3). These results indicate the dependability of the sequencing data.

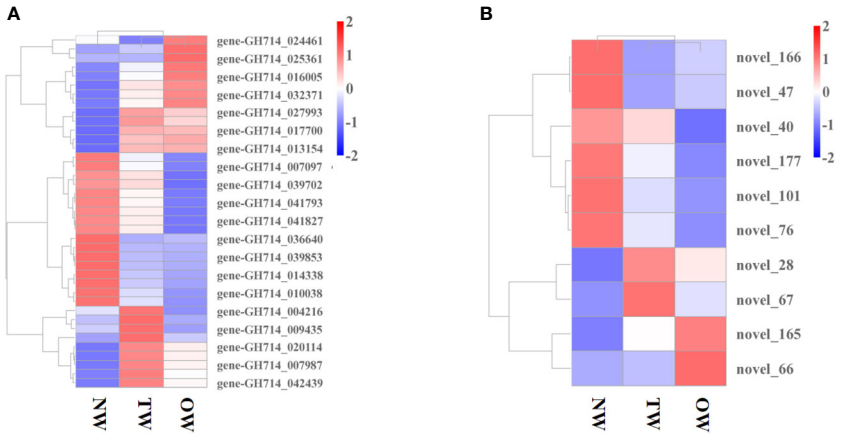


FIGURE 5
Expression profile of miRNAs and genes co-expressed with transcription factor genes. **(A)** Heatmap representation of the expression levels of genes co-expressed with transcription factor genes in different tissues. **(B)** Heatmap representation of the expression profile of miRNAs in TW, OW, and NW. Columns represent three different tissue types and rows represent different transcripts. Each square represents a transcript and the color indicates the level of expression; red represents up-regulation and blue represents down-regulation.

HbrCAD1 transcripts are cleaved by *hbr-miR482c*

miRNAs can cause the cleavage of their target transcripts between the 10th and 11th nucleotides of their target site (Rhoades et al., 2002). In this study, we predicted from bioinformatics analysis that *hbr-miR482c* may influence lignin biosynthesis by targeting *HbrCAD1* transcripts (Figure 4). To test this hypothesis experimentally, we co-expressed the precursor of *hbr-miR482c* and HbrCAD1, driven by the strong cauliflower mosaic virus (35S) promoter, into *N. benthamiana* leaves by Agrobacterium-mediated transient infiltration. As negative controls, we co-infiltrated the construct expressing the *HbrCAD1* and the empty effector vector (pBI121). We observed that *HCAD1* transcript levels decrease significantly when the *hbr-miR482c* precursor is co-expressed (Figure 1D). Thus we suspected that *HbrCAD1* transcripts might be the target gene of *hbr-miR482c*.

Discussion

A model for reaction wood formation in rubber tree and other tree species

The scanning electron microscopy of the xylem samples from reaction wood had confirmed that the tension wood had the remarkably thick gelatinous layer (Figure S2). These results were supported by earlier findings and confirmed a tension wood model were constructed in our study (Sujan et al., 2015). We wished to identify genes important for wood formation based on their expression models in various wood tissues of rubber tree. We drew heatmaps and constructed an interaction network to reveal connections between genes related to phenylpropanoid biosynthesis.

Enriched KEGG terms underscored the involvement of phenylpropanoid biosynthesis, phenylalanine metabolism, and pyruvate metabolism in reaction wood formation. Indeed, our results showed that the genes co-expressed with miRNA-targeted TF genes were related to metabolism and physiological functions that all contribute to reaction wood development. Similar consequences have been reported in previous work, lending some support to our current study (Li et al., 2013; Lv et al., 2021). Future work should pay close attention to exploring the functions of noncoding RNAs and their candidate target genes predicted from our RNA-seq results. In particular, genes linked to wood development can now be identified from their expression patterns across wood growth regions.

Integrated analysis of the miRNA-TF-mRNA network

Important transcription factors associated with secondary growth have been identified, such as members of the C3H and MYB TF families (Demura and Fukuda, 2007; Zhong and Ye, 2009). Zhang et al. reported that miRNAs may be connected to tension wood development by regulating secondary cell wall biosynthesis in Moso bamboo (*Phyllostachys edulis*) (Zhang et al., 2018). In this study, we determined that the transcript levels of 21 TF genes from 13 families changed over the course of wood bending (Figure S9). Of these, TF genes from the *C3H* and *MYB* families may help adjust the expression of genes related to phenylpropanoid biosynthesis and possibly play a significant role in the wood development in rubber tree. Previous research revealed that overexpressing the MYB TF genes *PtoMYB216*, *PtoMYB74*, and *PtoMYB92* from *P. tomentosa* induced the expression of genes related to lignin biosynthesis, resulting in thicker xylem cell walls, more xylem layers, ectopic lignin deposition, and enhanced lignin contents by 13–50% (Qiaoyan et al., 2013; Li et al., 2015; Li et al., 2018). Analogously, compared to the control, the expression levels of lignin biosynthetic genes, lignification ability, xylem volume, and lignin levels of *C3H* overexpressed in *Arabidopsis* all increased (Fornalé et al., 2015). We also observed that C3H-type zinc finger TF family members possibly take part in wood development by regulating the transcript levels of *GLUCOSYLTRANSFERASE* (*GT*) (gene-GH714_027993), which is associated with cellulose biosynthesis. This observation underscores the significance of *C3H* family members during plant development.

miRNA regulaged key genes in wood formation pathway

The proteins that are thought to catalyze glucan-chain elongation in cellulose and callose biosynthesis are processive GTs that belong to GT2 family. In *Arabidopsis*, 10 to 12 GT2 family members form CESA (cellulose synthase catalytic subunit) and callose synthase (Yang et al., 2016). Moreover, GTs are a vital component of wood development. For instance, in *Arabidopsis*, CesA8 participates in secondary cell wall formation, as its loss-of-function mutation produced plants with a delicate stem phenotype (Taylor et al., 1999; Taylor et al., 2000; Taylor et al., 2003). Similarly, a mutation in rice *CesA8* caused a dramatic decrease in the cellulose content of secondary cell walls, resulting in a brittle culm phenotype (Zhang et al., 2009; Song et al., 2013). Meng et al. found that *HbrCesA8*, which is related to cellulose biosynthesis, may participate in reaction wood formation in rubber tree (Meng et al., 2021). These findings support our miRNA-mRNA model for rubber tree wood formation. Our multiomics-based approach produced a post transcription network of which several nodes are regulated by novel_67, which possibly affects some target genes during wood development.

The expression levels of *CAD*, *4CL*, peroxidase (*POD*), and *CAFFEATE O-METHYLTRANSFERASE* (*COMT*) are specifically linked to lignin composition (Do et al., 2007; Wagner et al., 2009; Vanholme et al., 2010; Chanoca et al., 2019). Here, we showed that *HbrCAD1* transcript levels are less abundant in OW tissues compared to NW, suggesting that HbrCAD1-catalyzed reactions might contribute less to wood formation in OW relative to NW tissues. Previous studies have shown that repressing *PAL*, *CAD*, or other enzymes of the lignin biosynthetic pathway may cause decreased lignin content (Chanoca et al., 2019). Furthermore, in *Arabidopsis*, studies have indicated that single or double knockout mutants in *POD*s

representatively resulted in a small but significant decrease in lignin content and altered lignin biosynthesis in inflorescence stalks, indicating that modulating *POD* expression levels may affect lignin composition (Herrero et al., 2013; Barros et al., 2015). In general, we hypothesize that the expression pattern of *HbrCAD1* contributes to establishing the difference in lignin levels across tissue types.

Conclusions

Based on the small RNA data obtained from wood tissues collected from rubber tree, we identified 114 miRNAs (25 known and 89 novel) present in 300-day reaction wood. We also established a network linking miRNAs, their putative TF target genes, and the genes that are co-expressed with these TF genes in the context of cellulose biosynthesis. Finally, we revealed the interaction landscape of these three regulatory layers in adjusting reaction wood growth and validated the network in wood formation of rubber trees. In summary, we described target genes associated with wood development in rubber tree and studied their post-transcriptional regulation. These results will provide the theoretical basis to clarify miRNA-mediated post-transcriptional mechanisms during wood growth and development in rubber trees.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://bigd.big.ac.cn/gsa, CRA007612 https://bigd.big.ac.cn/gsa, CRA004818.

Author contributions

JHC designed the research. MML, XM, YZ, YW, NJ and JMC performed the research. All authors analyzed and interpreted the data. JHC and MML wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (31960321), the PhD Scientific Research and Innovation Foundation of Sanya Yazhou Bay Science and Technology City (HSPHDSRF-2023-12-006) and the Scientific Research Fund Project of Hainan University (KYQD (ZR)1830).

Acknowledgments

We thanks Center for Analytical Instrumentation, Hainan University, for their help in our experiments. This research work is also supported by High-performance Computing Platform of YZBSTCACC.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1068796/full#supplementary-material

References

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell.* 116 (2), 281–297. doi: 10.1016/s0092-8674(04)00045-5

Barros, J., Serk, H., Granlund, I., and Pesquet, E. (2015). The cell biology of lignification in higher plants. *Ann. Bot.* 115, 1053–1074. doi: 10.1093/aob/mcv046

Chang, S., Puryear, J., and Cairney, J. (1993). A simple and efficient method for isolating RNA from pine trees. *Plant Mol. Biol. Rep.* 11 (2), 113–116. doi: 10.1007/BF02670468

Chanoca, A., de, V. L., and Boerjan, W. (2019). Lignin engineering in forest trees. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00912

Dai, X., Zhuang, Z., and Zhao, P. X. (2018). PsRNATarget: a plant small RNA target analysis server, (2017 Release). *Nucleic Acids Res.* 46 (W1), W49–W54. doi: 10.1093/nar/gky316

Demura, T., and Fukuda, H. (2007). Transcriptional regulation in wood formation. *Trends Plant Sci.* 12, 64–70. doi: 10.1016/j.tplants

Do, C. T., Pollet, B., Thévenin, J., Sibout, R., Denoue, D., Barrière, Y., et al. (2007). Both caffeoyl coenzyme a 3-o-methyltransferase 1 and caffeic acid O-methyltransferase 1 are involved in redundant functions for lignin, flavonoids and sinapoyl malate biosynthesis in *Arabidopsis*. *Planta.* 226 (5), 1117–1129. doi: 10.1007/s00425-007-0558-3

Du, Q., and Wang, H. (2015). The role of HD-ZIP III transcription factors and miR165/166 in vascular development and secondary cell wall formation. *Plant Signal Behav.* 10 (10), e1078955. doi: 10.1080/15592324.2015.1078955

English, J. J. (1997). Requirement of sense transcription for homology-dependent virus resistance and trans -inactivation. *Plant J.* 12, 597–603. doi: 10.1046/j.1365-313X.1997.d01-13.x

Eudes, A., George, A., Mukerjee, P., Kim, J. S., Pollet, B., Benke, P. I., et al. (2012). Biosynthesis and incorporation of side-chain-truncated lignin monomers to reduce lignin polymerization and enhance saccharification. *Plant Biotechnol. J.* 10 (5), 609–620. doi: 10.1111/j.1467-7652.2012.00692.x

Fornalé, S., Rencoret, J., Garcia-Calvo, L., Capellades, M., Encina, A., Santiago, R., et al. (2015). Cell wall modifications triggered by the down-regulation of coumarate 3-hydroxylase-1 in maize. *Plant Sci.* 236, 272–282. doi: 10.1016/j.plantsci.2015.04.007

Friedländer, M. R., MacKowiak, S. D., Li, N., Chen, W., and Rajewsky, N. (2012). MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* 40 (1), 37–52. doi: 10.1093/nar/gkr688

Gritsch, C., Wan, Y., Mitchell, R. A. C., Shewry, P. R., Hanley, S. J., and Karp, A. (2015). G-Fibre cell wall development in willow stems during tension wood induction. *J. Exp. Bot.* 66, 6447–6459. doi: 10.1093/jxb/err339

Herrero, J., Fernández-Pérez, F., Yebra, T., Novo-Uzal, E., Pomar, F., Pedreño, M. Á., et al. (2013). Bioinformatic and functional characterization of the basic peroxidase 72 from arabidopsis thaliana involved in lignin biosynthesis. *Planta* 237 (6), 1599–1612. doi: 10.1007/s00425-013-1865-5

Hou, J., Xu, H., Fan, D., Ran, L., Li, J., Wu, S., et al. (2020). Mir319a-targeted ptotcp20 regulates secondary growth *via* interactions with ptowox4 and ptownd6 in populus tomentosa. *New Phytol.* 228 (4), 1354–1368. doi: 10.1111/nph

Kalvari, I., Nawrocki, E. P., Argasinska, J., Quinones-Olvera, N., Finn, R. D., Bateman, A., et al. (2018). Non-coding RNA analysis using the rfam database. *Curr. Protoc. Bioinf.* 62 (1), e51. doi: 10.1002/cpbi.51

Kuyyogsuy, A., Deenamo, N., Khompatara, K., and Ekchaweng, K. (2018). Chitosan enhances resistance in rubber tree (*Hevea brasiliensis*), through the induction of abscisic acid (ABA). *Physiol. Mol. Plant P.* 102, 67–78. doi: 10.1016/j.pmpp.2017.12.001

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10 (3), R25. doi: 10.1186/gb-2009-10-3-r25

Li, H., Huang, X., Li, W., Lu, Y., Dai, X., Zhou, Z., et al. (2020). MicroRNA comparison between poplar and larch provides insight into the different mechanism of wood formation. *Plant Cell Rep.* 39 (9), 1199–1217. doi: 10.1007/s00299-020-02559-3

Li, X., Yang, X., and Wu, H. X. (2013). Transcriptome profiling of radiata pine branches reveals new insights into reaction wood formation with implications in plant gravitropism. *BMC Genomics* 14, 768. doi: 10.1186/1471-2164-14-768

Li, C., Wang, X., Ran, L., Tian, Q., Fan, D., and Luo, K. (2015). Ptomyb92 is a transcriptional activator of the lignin biosynthetic pathway during secondary cell wall formation in populus tomentosa. *Plant Cell Physiol.* 56 (12), 2436–2446. doi: 10.1093/pcp/pcv157

Li, C., Ma, X., Yu, H., Fu, Y., and Luo, K. (2018). Ectopic expression of ptomyb74 in poplar and arabidopsis promotes secondary cell wall formation. *Front. Plant Sci.* 9, 1262 doi: 10.3389/fpls.2018.01262

Liu, J., Shi, C., Shi, C. C., Li, W., Zhang, Q. J., Zhang, Y., et al. (2020). The chromosome-based rubber tree genome provides new insights into spurge genome evolution and rubber biosynthesis. *Mol. Plant* 13 (2), 336–350. doi: 10.1016/j.molp.2019.10.017

Lu, S., Li, Q., Wei, H., Chang, M. J., Tunlaya-Anukit, S., Kim, H., et al. (2013). Ptr-miR397a is a negative regulator of laccase genes affecting lignin content in *Populus trichocarpa*. *Proc. Natl. Acad. Sci. U.S.A.* 110 (26), 10848–10853. doi: 10.1073/pnas.1308936110

Lv, C., Lu, W., Quan, M., Xiao, L., Li, L., Zhou, J., et al. (2021). Pyramiding superior haplotypes and epistatic alleles to accelerate wood quality andyield improvement in poplar breeding. *Ind. Crop Prod.* 171, 113891. doi: 10.1016/j.indcrop.2021.113891

Mao, X., Cai, T., Olyarchuk, J. G., and Wei, L. (2005). Automated genome annotation and pathway identification using the KEGG orthology (KO) as a controlled vocabulary. *Bioinformatics.* 21, 3787–3793. doi: 10.1093/bioinformatics/bti430

Mellerowicz, E. J., and Gorshkova, T. A. (2012). Tensional stress generation in gelatinous fibres: a review and possible mechanism based on cell-wall structure and composition. *J. Exp. Bot.* 63, 551–565. doi: 10.1093/jxb/err339

Meng, X. X., Kong, L. S., Zhang, Y. Y., Wu, M. J., Wang, Y., Li, J., et al. (2022). Gene expression analysis revealed hbr-miR396b as a key piece participating in reaction wood formation of *Hevea brasiliensis* (rubber tree). *Ind. Crop Prod.* 177, 114460. doi: 10.1016/j.indcrop.2021.114460

Meng, X. X., Wang, Y., Li, J., Jiao, N., Zhang, X., Zhang, Y., et al. (2021). RNA Sequencing reveals phenylpropanoid biosynthesis genes and transcription factors for *Hevea brasiliensis* reaction wood formation. *Front. Genet.* 29. doi: 10.3389/fgene.2021.763841

Meng, X., Zhang, X., Li, J., and Liu, P. (2018). Identification and comparative profiling of ovarian and testicular microRNAs in the swimming crab *Portunus trituberculatus*. *Gene.* 640, 6–13. doi: 10.1016/j.gene

Nair, K. (2010). The agronomy and economy of important tree crops of the developing world. *Agron. Economy Important Tree Crops Developing World.* 30 (4), 313–351. doi: 10.1016/B978-0-12-384677-8.00008-4

Nawaz, Z., Kakar, K. U., Ullah, R., Yu, S., Zhang, J., Shu, Q. Y., et al. (2019). Genome-wide identification, evolution and expression analysis of cyclic nucleotide-gated channels in tobacco (*Nicotiana tabacum* l.). *Genomics.* 111 (2), 142–158. doi: 10.1016/j.ygeno.2018.01.010

Pramod, S., Reghu, C. P., and Rao, K. S. (2019). *Biochemical characterization of wood lignin of hevea brasiliensis. wood is good* (Singapore: Springer), 199–209. doi: 10.1007/978-981-10-3115-1_19

Priyadarshan, P. M. (2017). Refinements to hevea rubber breeding. *Tree Genet. Genomes.* 13 (1), 1–17. doi: 10.1007/s11295-017-1101-8

Qiaoyan, T., Xianqiang, W., Chaofeng, L., Wanxiang, L., Li, Y., Yuanzhong, J., et al. (2013). Functional characterization of the poplar r2r3-myb transcription factor ptomyb216 involved in the regulation of lignin biosynthesis during wood formation. *PloS One* 8 (10), e76369–. doi: 10.1371/journal.pone.0076369

Quan, M., Du, Q., Xiao, L., Lu, W., Wang, L., Xie, J., et al. (2019). Genetic architecture underlying the lignin biosynthesis pathway involves noncoding RNAs and transcription factors for growth and wood properties in *Populus*. *Plant Biotechnol. J.* 17 (1), 302–315. doi: 10.1111/pbi.12978

Quan, M., Xiao, L., Lu, W., Liu, X., Song, F., Si, J., et al. (2018). Association genetics in *Populus* reveal the allelic interactions of *Pto-MIR167a* and its targets in wood formation. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00744

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., and Bartel, D. P. (2002). Prediction of plant microRNA targets. *Cell.* 110 (4), 513–520. doi: 10.1016/s0092-8674(02)00863-2

Schmittgen, T. D., and Livak, K. J. (2008). Analyzing real-time PCR data by the comparative CT method. *Nat. Protoc.* 3, 1101–1108. doi: 10.1038/nprot.2008.73

Shen, H., He, X., Poovaiah, C. R., Wuddineh, W. A., Ma, J., Mann, D. G. J., et al. (2012). Functional characterization of the switchgrass (Panicum virgatum) R2R3-*MYB* transcription factor Pv*MYB*4 for improvement of lignocellulosic feedstocks. *New Phytol.* 193 (1), 121–136. doi: 10.1111/j.1469-8137.2011.0392

Sibout, R., Eudes, A., Mouille, G., Pollet, B., Lapierre, C., Jouanin, L., et al. (2005). *CINNAMYL ALCOHOL DEHYDROGENASE-C* and *-D* are the primary genes involved in lignin biosynthesis in the floral stem of *Arabidopsis*. *Plant Cell.* 17, 2059–2076. doi: 10.1105/tpc.105.030767

Song, X. Q., Liu, L. F., Jiang, Y. J., Zhang, B. C., Gao, Y. P., Liu, X. L., et al. (2013). Disruption of secondary wall cellulose biosynthesis alters *CAD*mium translocation and tolerance in rice plants. *Mol. Plant* 6, 768–780. doi: 10.1093/mp/sst025

Sujan, K. C., Yamamoto, H., Matsuo, M., Yoshida, M., Naito, K., and Shirai, T. (2015). Continuum contraction of tension wood fiber induced by repetitive hygrothermal treatment. *Wood Sci. Technol.* 49, 1157–1169. doi: 10.1007/s00226-015-0762-4

Taylor, N. G., Howells, R. M., Huttly, A. K., Vickers, K., and Turner, S. R. (2003). Interactions among three distinct *CesA* proteins essential for cellulose synthesis. *Proc. Natl. Acad. Sci. U. S. A.* 100, 1450–1455. doi: 10.1073/pnas.0337628100

Taylor, N. G., Laurie, S., and Turner, S. R. (2000). Multiple cellulose synthase catalytic subunits are required for cellulose synthesis in *Aarabidopsis*. *Plant Cell.* 12, 2529–2539. doi: 10.1105/tpc.12.12.2529

Taylor, N. G., Scheible, W. R., Cutler, S., Somerville, C. R., and Turner, S. R. (1999). The irregular xylem3 locus of *Aarabidopsis* encodes a cellulose synthase required for secondary cell wall synthesis. *Plant Cell.* 11, 769–779. doi: 10.1105/tpc.11.5.769

Tempel, S. (2012). Using and understanding RepeatMasker. *Methods Mol. Biol.* 859, 29–51. doi: 10.1007/978-1-61779-603-6_2

Thaochan, N., Pornsuriya, C., Chairin, T., and Sunpapao, A. (2020). Roles of systemic fungicide in antifungal activity and induced defense responses in rubber tree (*Hevea brasiliensis*) against leaf fall disease caused by *Neopestalotiopsis cubana*. *Physiol. Mol. Plant P* 111, 101511. doi: 10.1016/j.pmpp.2020.101511

Vanholme, R., Ralph, J., Akiyama, T., Lu, F., Pazo, J. R., Kim, H., et al. (2010). Engineering traditional monolignols out of lignin by concomitant up-regulation of F5H1 and down-regulation of COMT in *Arabidopsis*. *Plant J.* 64 (6), 885–897. doi: 10.1111/j.1365-313X.2010.04353.x

Voelker, S. L., Lachenbruch, B., Meinzer, F. C., Jourdes, M., Ki, C., Patten, A. M., et al. (2010). Antisense down-regulation of *4CL* expression alterslignification, tree growth, and saccharification potential of field-grownpoplar. *Plant Physiol.* 154 (2), 874–886. doi: 10.1104/pp.110.159269

Wagner, A., Donaldson, L., Kim, H., Phillips, L., Flint, H., Steward, D., et al. (2009). Suppression of 4-coumarate-CoA ligase in the coniferous gymnosperm *Pinus radiata*. *Plant Physiol.* 149 (1), 370–383. doi: 10.1104/pp.108.125765

Wen, M., Shen, Y., Shi, S., and Tang, T. (2012). MiREvo: an integrative microRNA evolutionary analysis platform for next-generation sequencing experiments. *BMC Bioinforma.* 13, 140. doi: 10.1186/1471-2105-13-140

Weng, J. K., Mo, H., and Chapple, C. (2010). Over-expression of *F5H* in COMT-deficient *Arabidopsis* leads to enrichment of an unusual lignin and disruption of pollen wall formation. *Plant J.* 64 (6), 898–911. doi: 10.1111/j.1365-313X.2010.04391.x

Yang, W., Schuster, C., Beahan, C. T., Charoensawan, V., Peaucelle, A., Bacic, A., et al. (2016). Regulation of meristem morphogenesis by cell wall synthases in *Arabidopsis*. *Curr. Biol.* 26 (11), 1404–1415. doi: 10.1016/j.cub.2016.04.026

Yu, X., Gong, H., Cao, L., Hou, Y., and Qu, S. (2020). *MicroRNA397b* negatively regulates resistance of *Malus hupehensis* to *Botryosphaeria dothidea* by modulating *MhLAC7* involved in lignin biosynthesis. *Plant Sci.* 292, 110390. doi: 10.1016/j.plantsci.2019.110390

Zeng, C., Wang, W., Zheng, Y., Chen, X., Bo, W., Song, S., et al. (2009). Conservation and divergence of microRNAs and their functions in euphorbiaceous plants. *Nucleic Acids Res.* 38, 981–995. doi: 10.1093/nar/gkp1035

Zhang, B., Deng, L., Qian, Q., Xiong, G., Zeng, D., Li, R., et al. (2009). A missense mutation in the transmembrane domain of *CESA4* affects protein abundance in the plasma membrane and results in abnormal cell wall biosynthesis in rice. *Plant Mol. Biol.* 71, 509–524. doi: 10.1007/s11103-009-9536-4

Zhang, X., Wang, W., Zhu, W., Dong, J., Cheng, Y., Yin, Z., Beahan, C., et al. (2019). Mechanisms and functions of long non-coding RNAs at multiple regulatory levels. *Int. J. Mol. Sci.* 20 (22), 5573. doi: 10.3390/ijms20225573

Zhang, H., Ying, Y. Q., Wang, J., Zhao, X. H., Zeng, W., Beahan, C., et al. (2018). Transcriptome analysis provides insights into xylogenesis formation in moso bamboo (phyllostachys edulis) shoot. *Sci. Rep.* 8 (1), 3951. doi: 10.1038/s41598-018-21766-3

Zhong, R., and Ye, Z. H. (2009). Transcriptional regulation of lignin biosynthesis. *Plant Signal. Behav.* 4, 1028–1034. doi: 10.4161/psb.4.11.9875

Zhong, R., Lee, C., and Ye, Z. H. (2010). Functional characterization of poplar wood-associated NAC domain transcription factors. *Plant Physiol.* 152, 1044–1055. doi: 10.1104/pp.109.148270

Zhou, L., Chen, J., Li, Z., Li, X., Hu, X., Huang, Y., et al. (2010). Integrated profiling of MicroRNAs and mRNAs: microRNAs located on Xq27.3 associate with clear cell renal cell carcinoma. *PloS One* 5 (12), e15224. doi: 10.1371/journal.pone.0015224

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 16 November 2023
DOI 10.3389/fpls.2023.1184112

OPEN ACCESS

EDITED BY
Kai-Hua Jia,
Shandong Academy of Agricultural Sciences, China

REVIEWED BY
Ren-Gang Zhang,
Chinese Academy of Sciences (CAS), China
Jin Hoe Huh,
Seoul National University, Republic of Korea
Sunil Kumar Sahu,
Beijing Genomics Institute (BGI), China

*CORRESPONDENCE
Amanda M. Hulse-Kemp
amanda.hulse-kemp@usda.gov

RECEIVED 11 March 2023
ACCEPTED 17 October 2023
PUBLISHED 16 November 2023

CITATION
Delorean EE, Youngblood RC, Simpson SA, Schoonmaker AN, Scheffler BE, Rutter WB and Hulse-Kemp AM (2023) Representing true plant genomes: haplotype-resolved hybrid pepper genome with trio-binning. *Front. Plant Sci.* 14:1184112. doi: 10.3389/fpls.2023.1184112

Representing true plant genomes: haplotype-resolved hybrid pepper genome with trio-binning

Emily E. Delorean[1,2], Ramey C. Youngblood[3], Sheron A. Simpson[4], Ashley N. Schoonmaker[2], Brian E. Scheffler[4], William B. Rutter[5] and Amanda M. Hulse-Kemp[1,2]*

[1]Genomics and Bioinformatics Research Unit, USDA-ARS, Raleigh, NC, United States, [2]Crop and Soil Sciences Department, North Carolina State University, Raleigh, NC, United States, [3]Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University, Starkville, MS, United States, [4]Genomics and Bioinformatics Research Unit, United States Department of Agriculture - Agriculture Research Service (USDA-ARS), Stoneville, MS, United States, [5]US Vegetable Laboratory, United States Department of Agriculture - Agriculture Research Service (USDA-ARS), Charleston, SC, United States

As sequencing costs decrease and availability of high fidelity long-read sequencing increases, generating experiment specific *de novo* genome assemblies becomes feasible. In many crop species, obtaining the genome of a hybrid or heterozygous individual is necessary for systems that do not tolerate inbreeding or for investigating important biological questions, such as hybrid vigor. However, most genome assembly methods that have been used in plants result in a merged single sequence representation that is not a true biologically accurate representation of either haplotype within a diploid individual. The resulting genome assembly is often fragmented and exhibits a mosaic of the two haplotypes, referred to as haplotype-switching. Important haplotype level information, such as causal mutations and structural variation is therefore lost causing difficulties in interpreting downstream analyses. To overcome this challenge, we have applied a method developed for animal genome assembly called trio-binning to an intra-specific hybrid of chili pepper (*Capsicum annuum* L. cv. HDA149 *x Capsicum annuum* L. cv. HDA330). We tested all currently available softwares for performing trio-binning, combined with multiple scaffolding technologies including Bionano to determine the optimal method of producing the best haplotype-resolved assembly. Ultimately, we produced highly contiguous biologically true haplotype-resolved genome assemblies for each parent, with scaffold N50s of 266.0 Mb and 281.3 Mb, with 99.6% and 99.8% positioned into chromosomes respectively. The assemblies captured 3.10 Gb and 3.12 Gb of the estimated 3.5 Gb chili pepper genome size. These assemblies represent the complete genome structure of the intraspecific hybrid, as well as the two parental genomes, and show measurable improvements over the currently available reference genomes. Our manuscript provides a valuable guide on how to apply trio-binning to other plant genomes.

KEYWORDS

haplotype, pepper, genome assembly, trio-binning, HiFi

1 Introduction

Reference genomes are now available for hundreds of plant species, providing valuable tools for researchers and plant breeders. However, there are still limitations in many of the available reference genomes. As the numbers of *de novo* reference assemblies increase and pan-genome assemblies become more widely available (Bayer et al., 2020), we are finding that individuals exhibit varying amounts of presence/absence variation (PAV), copy number variation (CNV) and structural variation (SV) (Liu et al., 2020; Wang et al., 2020; Lee et al., 2022; Tang et al., 2022; Yang et al., 2022; Zhou et al., 2022). This variation not only occurs between individuals, but also within the genome of a single individual when that individual is heterozygous. Important genetic information is often lost when sequencing reads are aligned to a single merged reference genome. Ideally, the true haplotype of each individual within a project would be available, particularly for founder parents of breeding lines.

The decreasing cost of DNA sequencing in conjunction with third generation long-read sequencing technologies has brought custom plant genome assemblies a step closer to reality, even for polyploids and species with large genomes (Kress et al., 2022; Newman et al., 2023; Sahu and Liu, 2023). However, there are still many technical hurdles involved in assembling a biologically accurate fully-phased plant genome. The typical genome assembly is a haploid representation of a diploid individual. If the individual is homozygous then a single haploid genome assembly is sufficient given that the two haploid genomes, or haplotypes, within the organism are effectively the same. If the organism is heterozygous then there becomes the chance that the resulting genome assembly is a mosaic or chimera of the individual's two haplotypes (haplotype switching). These chimeric genomic regions are not biologically accurate and may be misleading in downstream analysis, such as during candidate gene mining (Benevenuto et al., 2019). Correctly assembling each haploid genome is referred to as haplotype phasing and is one of the key challenges facing modern genome assembly methods.

Advances in computational approaches can also help overcome these genome assembly challenges, significantly decrease costs, and improve the quality of the final assemblies. The error profiles in long-read sequencing were drastically improved with the availability of circular consensus or high-fidelity (HiFi) reads, which became available in 2019 (Wenger et al., 2019). Currently there are two genome assembly softwares that support HiFi reads, Hifiasm (Cheng et al., 2021) and HiCanu (Nurk et al., 2020). Hifiasm is an intrinsically haplotype-aware assembler that builds a string graph of overlapping sequences where all haplotype information is saved as a fork (called bubbles). Hifiasm by default also generates two partially phased haplotypes assemblies (hap1/hap2).

Several approaches have been used to try and correct haplotype switching and produce an accurate fully phased genome. Prior to HiFi reads, haplotype phasing was often highly involved and relied on single nucleotide polymorphism (SNP) data or germ cell sequencing (Minio et al., 2017; Shi et al., 2019; Campoy et al., 2020; Minio et al., 2022). Falcon and Falcon-Unzip assemblers corrected the high error rate of PacBio Continuous Long Read Sequencing (CLR) and used differences in SNPs to partition haplotypes (Chin et al., 2016). Another advance in technology came with the advent of Hi-C sequencing for scaffolding, which relies on intra-chromosomal contacts. The Hi-C paired-end reads are aligned to a partially phased genome assembly to determine which pieces of the assembly, or haplotigs, belong together along a chromosome. The disadvantage of using Hi-C is the absence of inter-chromosomal information, which means that sorting chromosomes into the proper genome isn't possible and the phasing success may be lower compared to other methods (Kronenberg et al., 2021; Mao et al., 2023). While Bionano optical maps have the capability of providing haplotype phasing for humans (Seo et al., 2016), that utility is not yet available for plants.

Another method to resolve haplotype switching is trio-binning, where short reads generated from the parents are used to bin long reads generated from the offspring prior to assembly. A 'trio' refers to the combination of a mother-father-offspring. This method has been suggested for development of telomere to telomere (T2T) or gapless genome assembly efforts (Nurk et al., 2022). Excitingly, we are seeing the first T2T plant genomes being released, but these again are for small homozygous genome species like rice and did not use trio-binning (Li et al., 2021; Gladman et al., 2023). The trio-binning genome assembly method (Figure 1) relies on HiFi long read sequencing of an F_1 individual and short read sequencing of the two parent lines of the F_1 individual (Figure 1A). The short reads from the parents are broken into k-mers that are distinct in one parent line compared to the other parental line. These k-mers are then aligned to the long reads to partition the long reads into 3 sequence bins containing either 1) long reads unique to parent A, 2) long reads unique to parent B, or 3) long reads that are likely shared between parent A and B. Trio-binning has been used extensively to help phase and assemble animal genomes (Koren et al., 2018; Yen et al., 2020; Rhie et al., 2021; Yang et al., 2021; Rautiainen et al., 2023). In plants, it has recently been used for an inter-specific hybrid (Montgomery et al., 2020) and then most recently for a cross between subspecies (Huang et al., 2022), but not yet for an intra-specific cross. Intra-specific crosses, or breeding within the same species, represent most plants that researchers and breeders are working with.

In this study, we applied trio-binning to simultaneously assemble two phased genomes from an intra-specific cross of two parental chili pepper lines (*Capsicum annuum* L. cv 'HDA149' and cv 'HDA330'). Accurate assembly of the chili pepper (*Capsicum annuum* L., 2n=2x=24) genome is frustrated by its large size (3.5 Gbp) and complexity due to high rate of repetitive elements (75-80%) (Kim et al., 2014; Qin et al., 2014; Hulse-Kemp et al., 2018; Lee et al., 2022; Shirasawa et al., 2022). Our results show that this method, when combined with modern scaffolding approaches, can successfully be used to produce two high quality phased genomes that are just as contiguous as the best currently available references, produced in recent pan-genome efforts (*Capsicum annuum* L. cultivars 'Dempsey' and 'Zhangshugang') (Lee et al., 2022; Liu et al., 2023). We provide best practices that can be applied by other groups seeking to produce quality biologically accurate, haplotype-resolved reference genomes for their lines of interest.

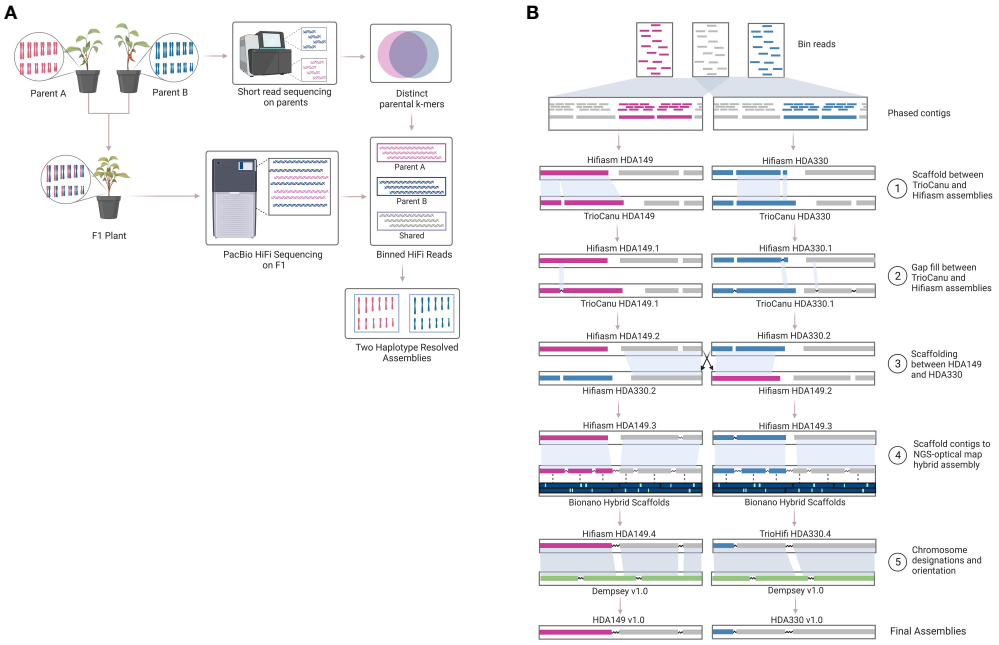


FIGURE 1
Trio-binning workflow. Overview of the trio-binning workflow used for producing haplotype-resolved biologically accurate plant genomes. **(A)** Schematic of lab-based and theoretical protocol utilized in the trio-binning workflow. **(B)** Detailed overview of the *in silico* process for assembly and scaffolding, including incorporation of Bionano optical mapping.

2 Results

2.1 Plant selection and sequencing

The two *Capsicum annuum* L. (2n = 2x = 24) parental pepper plants, HDA149 and HDA330, were confirmed as double haploids with the Illumina PepperSNP16K array (Hulse-Kemp et al., 2016). As expected, each line exhibited near complete homozygosity across all SNP sites in the array; 99.7% for HDA149 and 99.5% for HDA330 (Supplemental Data 1). Compared to each other, the parents had different alleles at 26.1% of SNP sites on the array and mummer alignments of the F_1 indicated a genome-wide heterozygosity rate of 0.1168%.

We sequenced the parental plants at 45-50x depth with Illumina 150 bp paired end reads. The two parent plants were crossed to generate the F_1 hybrid plant and it was sequenced to 58x depth with PacBio HiFi long reads (Sequel IIe) over 7 SMRT cells. Depth of sequencing and genome-wide heterozygosity were calculated based on the estimated genome size of *Capsicum annuum* of 3.5 Gb (Belletti et al., 1998; Moscone et al., 2003; Hulse-Kemp et al., 2018). The resulting three DNA sequence data sets (HDA149 Illumina short reads, HDA330 Illumina short reads, and the F_1 Hifi reads) were used for trio-binning genome assembly.

2.2 Trio-binning assembly

We conducted trio-binning on the PacBio HiFi reads of the F_1 hybrid using the two assembly softwares available at the time of this research, TrioCanu (Koren et al., 2018) and Hifiasm (Cheng et al., 2021). Both softwares utilize the k-mers from the parental short reads for haplotype partitioning. TrioCanu bins the HiFi reads prior to assembly; in contrast, Hifiasm partitions haplotigs after assembly. Of the 11,822,010 total HiFi reads after filtering, TrioCanu partitioned 4,586,239 reads (38.8%) to the HDA149 specific bin, 4,505,092 (38.1%) to the HDA330 specific bin, and 2,729,826 (23.1%) to the shared bin of non-haplotype specific reads. The shared reads and corresponding haplotype binned reads were used to generate a TrioCanu assembly for each parent. The resulting assemblies, TrioCanu HDA149 and TrioCanu HDA330, were highly contiguous with N50 values of 66.53 and 86.50 Mb, and genome size values of 3.31 and 3.30 Gb (Table 1). The Hifiasm assemblies, Hifiasm HDA149 and Hifiasm HDA330, exhibited higher contiguity with N50 values of 228.06 and 177.89 Mb, but lower genome size values of 3.10 and 3.09 Gb (Table 1).

2.3 Haplotype switching

To confirm that the trio-binning assemblies were haplotype resolved, we mapped the TrioCanu binned reads onto each of the assemblies and calculated differences in alignment coverage over 1 Mb windows. Haplotype specific windows of an assembly will show high alignment coverage for the corresponding set of parent specific binned reads and low coverage for the opposite set of parent specific binned reads. Each assembly should show differences in alignment rates favoring only the reads from their specific corresponding bins if there is no haplotype switching, for example TrioCanu HDA149 should have windows of higher alignment rates only for HDA149 reads. All four assemblies, TrioCanu HDA149 (97.7%), TrioCanu HDA330 (97.7%), Hifiasm HDA149 (97.7%) and Hifiasm HDA330

TABLE 1 Experimental assembly comparison.

	Hifiasm HDA149	Hifiasm HDA330	TrioCanu HDA149	TrioCanu HDA330
Binning software	yak	yak	Canu v2.2	Canu v2.2
Assembly software	Hifiasm v0.16.1-r375	Hifiasm v0.16.1-r375	Canu v2.2	Canu v2.2
Number of contigs	364	119	5879	5914
Contig N50 (Mb)	228.056	177.885	66.526	86.496
Longest contig (Mb)	263.427	270.729	40.427	57.044
Assembly Size (Mb)	3100.1	3088.7	3306.1	3297.9

Comparison of genome assembly statistics of trio-binned assemblies generated with Hifiasm and HiCanu.

(99.3%) showed windows of higher alignment rates for their corresponding haplotype bin, indicating that the assemblies were correctly haplotype resolved (Figures 2A–D).

As a control, we also generated non-binned Hifiasm assemblies of the F_1 and calculated differences in alignment coverage of the binned reads. By default, without parental k-mers or Hi-C data, Hifiasm attempts to naively phase haplotypes and produces 3 assemblies: primary, hap1 and hap2 https://github.com/chhylp123/hifiasm, Accessed 03/09/2023). These non-binned assemblies showed haplotype switching, calculated in the same way as above, the Hifiasm hap1 had 86.8% and hap2 had 87.5% alignment rates. This was visualized as 1 Mb windows which showed alternating haplotypes of higher alignment rate (Figures 2E, F).

2.4 Assembler comparison

In total, we generated 4 trio-binned assemblies with two assemblers (Hifiasm and HiCanu). We named these assemblies 'TrioCanu HDA149', 'TrioCanu HDA330', 'Hifiasm HDA149' and 'Hifiasm HDA330' (Table 1). The Hifiasm assemblies had 16-50x fewer contigs and 2-3x higher N50 values than the TrioCanu assemblies (Table 1), but the TrioCanu assemblies were ~ 200 Mb larger in size.

We were curious if the two assemblers differed in their ability to assemble the same genomic regions. To test this, we mapped the TrioCanu assemblies against the Hifiasm assemblies and generated dotplots of the largest contigs, > 5 Mb. Overall, the two assemblers generated the same large contigs, however there were a number of sequence regions where one was able to assemble through while the other was not. In HDA149, Hifiasm was able to assemble through 3 regions that TrioCanu was not (Figure 3A). In HDA330, Hifiasm assembled though 7 regions that TrioCanu did not, and TrioCanu assembled through 1 region that Hifiasm did not (Figure 3B). Given that in these 11 regions one assembler performed better than the other, we decided to leverage this information during our scaffolding workflow, described in the next section.

2.5 Scaffolding and quality assessment of assemblies

The assemblies were highly contiguous owing to the PacBio HiFi reads, but we were interested to see if our data could generate chromosome scale *de novo* assemblies. To achieve this, we built an iterative scaffolding workflow that first used homology scaffolding between the different assemblers followed by gap-filling with RagTag software (Figure 1B and Supplemental Table 1). For example, Hifiasm HDA330 was scaffolded against and gap-filled using TrioCanu HDA330. Step 1 increased scaffold N50 values from 177.8 to 247.4 Mb for Hifiasm HDA330, from 86.5 to 232.1 Mb for TrioCanu HDA330, and from 66.5 to 231.1 Mb for TrioCanu HDA149 (Supplemental Data 2). As expected, Hifiasm HDA149 assembly N50 values did not increase in this step as there were no regions that TrioCanu had assembled better than Hifiasm in this haplotype (Figure 3A). However, during gap-filling, contig N50 values did increase for all 4 assemblies. For the third step, we anchored the assemblies using the other haplotype. For example, Hifiasm HDA330 was used to scaffold Hifiasm HDA149. Improvements were made in scaffold N90 values, from 134.5 Mb to 243.7 Mb in TrioCanu HDA149, from 131.9 to 189.9 Mb in Hifiasm HDA149, from 178.1 to 237.8 Mb in TrioCanu HDA330, and 177.9 to 189.9 Mb in Hifiasm HDA330. Although N50 and N90 values had improved considerably, the majority of each assembly (>3.0 Gb) was still not yet captured in 12 scaffolds representing each of 12 chromosomes (Supplemental Data 2, page "Scaffolding Statistics").

Next, we merged the Bionano optical map of the F_1 sample with each of the four contig level assemblies. Optical mapping gave scaffold N50 values between 182.6 – 210.9 Mb, however, during conflict resolution, the Bionano Saphyr software also made between 26 – 47 cuts to the contigs of our assemblies (Supplemental Data 2, page "Scaffolding Statistics"). This substantially lowered the contig N50 values by 2-3x. To retain contig integrity, we anchored the Step 3 scaffolded assemblies onto their respective Bionano-Hybrid assembly using RagTag (Figure 3). This brought our assemblies closer to full chromosome scale, with ~ 3.0 Gb of the assemblies being captured in 13 scaffolds for TrioCanu HDA149 and Hifiasm HDA149 and in 14 scaffolds for TrioCanu HDA330 and Hifiasm HDA330.

Final assembly of pseudomolecules were oriented and given chromosome designations through RagTag homology scaffolding to the previously published pepper assembly, Dempsey v1.0 (Figure 1B and Supplemental Table 1). During this step, we found that using Dempsey v1.0 allowed us to anchor distal ends of chr5 and chr11 in our assemblies. Our final Hifiasm assemblies which had the best assembly statistics (Table 2) were chosen as the final reference assemblies, HDA149v1.0 and HDA330v1.0, are available through NCBI

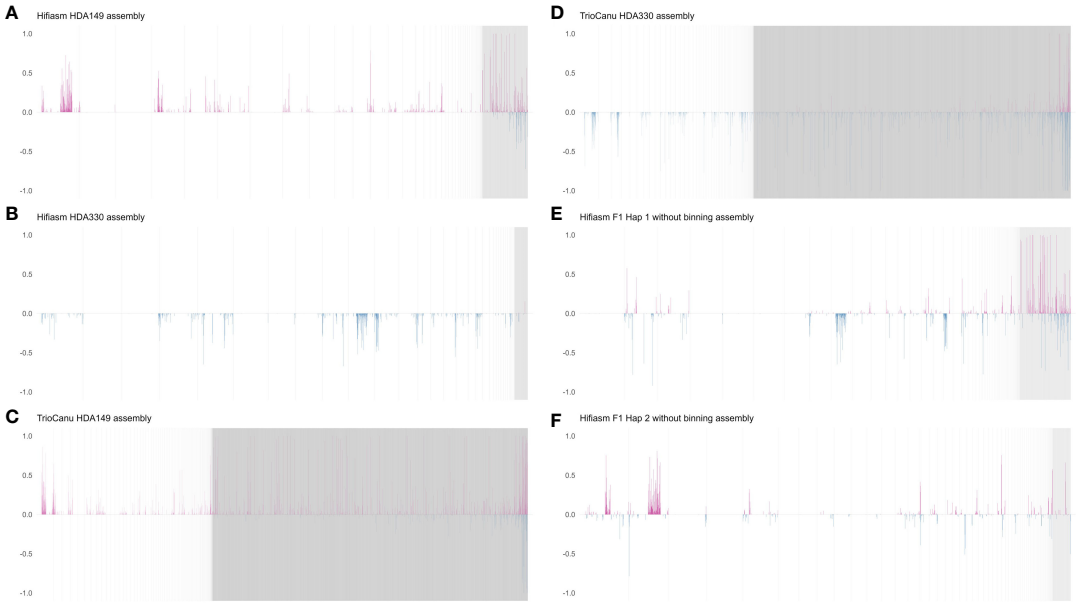


FIGURE 2
Haplotype switching. Haplotype switching was illustrated by aligning TrioCanu binned HiFi reads of parent A (HDA149) and parent B (HDA330) to each contig level genome assembly. The x-axis shows 1 Mb windows across contigs. The contigs were arranged from longest to shortest. Vertical gray lines show the boundaries of contigs. The y-axis shows the difference in percent coverage of the binned reads over a 1 Mb window of the given assembly. Higher coverage of HDA149 is shown in pink and higher coverage of HDA330 is shown in blue. **(A)** Hifiasm HDA149 assembly with trio-binning. **(B)** Hifiasm HDA330 assembly with trio-binning. **(C)** TrioCanu HDA149 assembly with trio-binning. **(D)** TrioCanu HDA330 assembly with trio-binning. **(E)** Hifiasm haplotype 1 assembly in default run mode, without parental k-mers for trio-binning. **(F)** Hifiasm haplotype 2 assembly in default run mode, without parental k-mers for trio-binning.

JAVHYQ000000000 and JAVHYR000000000, respectively and at the SolGenomics database (https://solgenomics.net/ftp/genomes/Capsicum_annuum/C.annuum_F1_HDA149_x_HDA330).

Our final TrioCanu assemblies, HDA149alt-v1.0 and HDA330alt-v1.0, are available through USDA Ag Data Commons (Supplemental Table 2, https://data.nal.usda.gov/dataset/triobinning-capsicum-annuum-genome-assemblies).

As measurements of assembly quality, we examined gaps in the Hifiasm assemblies, repeat content and telomere repeats (Figures 4A, B). We saw that generally gaps occurred toward the telomeric regions of the chromosomes, coinciding with the fact that most chromosomes were completely captured in a single contig as seen in chr7 of HDA330 or nearly completely captured as seen in chr6, chr8, chr9, chr10 and chr12 of HDA330. Strong telomere repeat peaks were detected in 12 of the 24 chromosome arms of HDA330. Peaks in long-terminal repeat (LTR) content did not mandate gaps in the assemblies, as seen clearly in chr1, chr7 and chr10 of HDA330. Similar results were seen for HDA149.

The final Hifiasm assemblies, HDA149v1.0 and HDA330v1.0, had minimal large scale structural variation (Figure 5A). Divergences in percent sequence identity were observed on several chromosomes, in particular chr9, chr7, and chr1. These results were expected given that HDA149 and HDA330 were developed as resistance gene introgressions into the Yolo Wonder background (Hendy et al., 1985). Compared to Dempsey, there was a large inversion on chr11 in both assemblies (Figures 5B, C). Overall, trio-binning with Hifiasm produced haplotype level assemblies with substantially higher contig N50 values of 228 Mb compared to 18 Mb for Dempsey and 35.4 Mb for Zhangshugang (Table 2). The trio-binning assemblies also had higher long-terminal repeat assembly index (LAI) scores of 8.98 and 9.00 compared to 7.70 for Dempsey and 8.19 for Zhangshugang (Table 2) (Lee et al., 2022; Liu et al., 2023). Our assemblies captured 77.0 and 92.5 Mb more of the total *C. annuum* genome. Additionally, the HDA330 assemblies reported a slight increase in genic space coverage as estimated with BUSCO (Table 2).

3 Discussion

We generated two high quality, fully haplotype phased *de novo* pepper (*Capsicum annuum* L.) genome assemblies using trio-binning, as evidenced by LAI scores of 8.98 and 9.00. These assemblies accurately represent the haploid genomes within a single diploid intra-specific hybrid plant, making a 9.6-9.8% improvement on completeness based on genome size estimates and are 6.4X more contiguous at the contig level with over 90% of the assembly sequence (Contig L90, Table 2) included in the first 10 contigs of each of the haplotypes produced in this study. High quality pepper genome assemblies such as the two presented here and those already available create a valuable community resource for in-depth analysis of genome evolution, structural variation, and haplotype specific gene clusters. Of particular interest in crop breeding are resistance gene clusters that are often haplotype specific with little or no recombination due to significant structural variation (Jiao and Schneeberger, 2020; Vaughn et al., 2022). A single reference

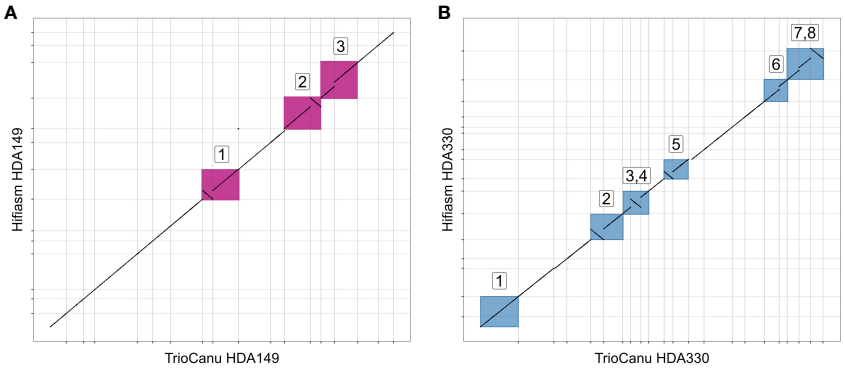


FIGURE 3
Utility of reciprocal scaffolding of assemblies from alternate software. Dotplots show alignments between largest contigs of TrioCanu and Hifiasm assemblies. Opportunities to improve contiguity through iterative scaffolding are highlighted in boxes that are numbered and shown in pink for HDA149 **(A)** or blue for HDA330 **(B)**.

TABLE 2 Final assembly statistics.

	HDA149v1.0	HDA330v1.0	Zhangshugang	Dempsey	UCD10X
Contig number	359	112	91	532	134,101
Scaffold number	239	51	601	121	81,378
Contig N50 (Mb)	228.1	228.6	35.4	18.3	0.1
Contig L50	7	7	25	51	6,631
Scaffold N50 (Mb)	266.0	281.3	259.7	260.5	227.2
Contig N90 (Mb)	131.9	171.4	19.4	9.7	0.1
Contig L90	10	10	49	98	13,035
Scaffold N90 (Mb)	254.0	254.6	253.2	249.5	219.1
Assembly Size (Mb)	3,100.6	3,118.8	3,023.8	3,053.5	3,124.3
% of Estimated Genome Size	88.6%	88.25%	86.39%	86.7%	89.3%
% of Assembly Placed in Chromosomes	99.6%	99.8%	99.9%	99.7%	83.2%
Busco Completeness (%)	97.4%	98.4%	97.1%	97.7%	96.5%
LAI	8.98	9.00	8.19	7.70	6.79
Source	This study	This study	Liu et al., 2023	Lee et al., 2022	Hulse-Kemp et al., 2018
Sequencing technology	PacBio HiFi	PacBio HiFi	PacBio CLR and Illumina short reads	PacBio CLR and Illumina short reads	10x Genomics Linked-Reads
Scaffolding technology	Bionano Optical Maps and RagTag homology based scaffolding	Bionano Optical Maps and RagTag homology based scaffolding	Phase Genomics Hi-C	Dovetail Hi-C, Bionano Optical Maps and four genetic maps	Four genetic maps, three transcriptome maps, and one genomic map

Comparison of assembly statistics between our trio-binned final assemblies, HDA149v1.0 and HDA330v1.0 and three previously published assemblies, Zhangshugang v1.0, Dempsey v1.0 and UCD10x v1.0.

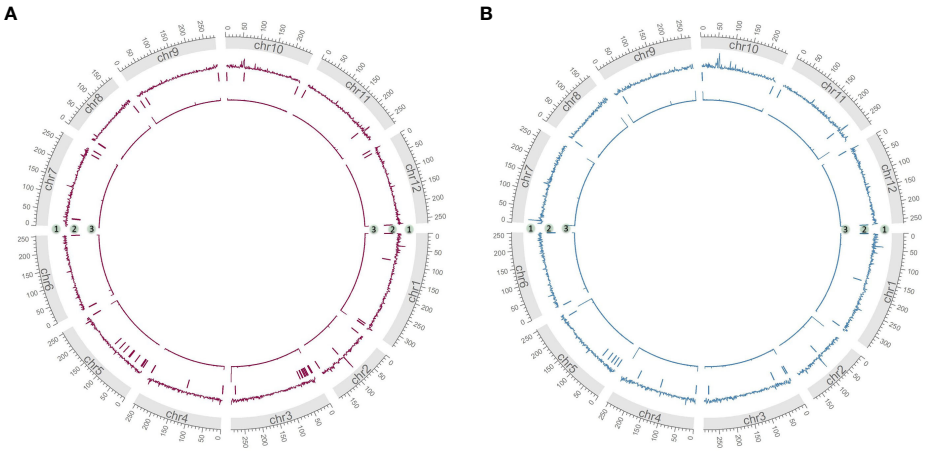


FIGURE 4
Characterization of developed assemblies. Circos plots of final Hifiasm assemblies HDA149v1.0 **(A)** and HDA330v1.0 **(B)** show long terminal repeat content across 1 MB windows in track 1, gap locations in track 2 and telomere repeat peaks across 1 kb windows in track 3.

genome inhibits characterization of these resistance genes and hinders reliable molecular marker development. In this new era of project specific high quality genome assemblies, researchers can now easily capture these important haplotype specific regions.

Generating *de novo* assemblies of an F_1 individual is a powerful tool for biparental mapping, experimental population studies, and breeding. These assemblies capture the complete landscape of sequence diversity segregating in that population, which is often difficult to discern when using a generic reference genome. An excellent example highlighting the improved ability of having the complete landscape of sequence diversity for detecting causative loci for traits of interest was recently published in melon (Vaughn et al., 2022). However, separately assembling two parental long read assemblies is more costly and potentially more error prone than trio-binning. Another benefit of trio-binning is that its ability to partition haplotypes increases with increasing heterozygosity of the individual, as shown in outbred individuals such as humans (Koren et al., 2018) and *Arctia plantaginis* (Yen et al., 2020), and in subspecies F_1 hybrids such as *Bos taurus taurus* x *Bos taurus indicus* (Koren et al., 2018) and *Amaranthus tuberculatus* x *Amaranthus hybridus* (Montgomery et al., 2020). The utility of Bionano optical maps have been extensively demonstrated by the Telomere-to-Telomere (T2T) Consortium for human genomes (Mc Cartney et al., 2022) and here we showed that Bionano optical maps can also be used in conjunction with trio-binning in plants.

We found benefit from utilizing multiple assembly softwares and the best solution was to use components from both software. The algorithms perform differently in different parts of the genomes and can complement each other through scaffolding techniques (Figure 3). Future development of these two softwares may improve usability and results that may negate the strategy we found performed the best. But if time and compute resources allow, it may also be beneficial for others to generate assemblies from both softwares so that iterative scaffolding can exploit the differences in assembler software and improve genome contiguity. Integration of other techniques such as HiC may also help to improve assembly of missing components into final scaffolds and has been shown to enable some haplotype-based assembly (Cheng et al., 2022), but would require additional cost and is difficult to obtain high-quality data in many plants.

Careful consideration, and likely direct comparison, of available assembly software should be made when generating *de novo* assemblies. Approximately 200 Mb additional of sequence was

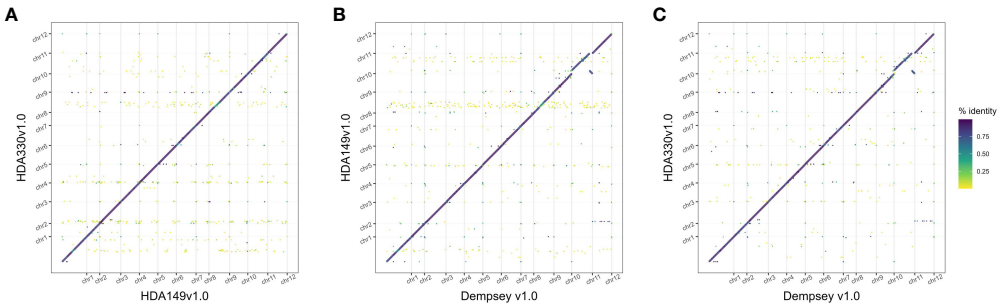


FIGURE 5
Comparison of final Hifiasm assemblies. Dotplots of assembly by assembly alignments of **(A)** HDA149v1.0 to HDA330v1.0, **(B)** HDA149v1.0 to Dempsey v1.0, and **(C)** HDA330v1.0 to Dempsey v1.0. Gridlines show boundaries of chromosomes (x-axis) and color indicates percent identity of the alignment.

10.3389/fpls.2023.1184112

captured into the final chromosomes of the TrioCanu assemblies compared to the Hifiasm assemblies (Figure 6). These additional sequences appeared repetitive given that they aligned across the Hifiasm assembly (Figures 6A, B). The additional regions were highly fragmented as shown by the number of gaps in the assemblies (Figures 6C, D, track 2) and likely contain assembly errors given the decrease in HiFi read coverage (Figures 6C, D, track 3). TrioCanu better assembled telomeric regions (Figures 6C, D, track 5), these results suggest that Hifiasm may be collapsing repetitive regions compared to TrioCanu. Resolution of complex repetitive regions have been achieved through a combination of several technologies and softwares for the human genome (Nurk et al., 2022). Cost and time of achieving a telomere-to-telomere genome assembly must be weighed against the research needs of each project.

This work shows that *de novo* assemblies using trio-binning as developed in this study are now relatively inexpensive and easy to generate even for intra-specific hybrids (in this case, ~$15,000 cost at the time of data generation for the raw reagent cost of sequencing and Bionano) and becoming even more feasible with continual drops in sequencing costs. Plant researchers should consider using trio-binning with the methods outlined here in future studies to represent the true biology of their plants to obtain haplotype-resolved genomes.

4 Materials and methods

4.1 Hybrid development and identification

Two *Capsicum annuum* L. double haploid lines were selected as parents for generating a controlled cross, HDA149 and HDA330 (Hendy et al., 1985; Thies and Ariss, 2009). A single individual of each parental line was used to make a cross HDA149 x HDA330. Young leaf tissue (two to three unfurled leaves) from both parental lines was extracted using a DNeasy Plant kit (Qiagen, Hilden, Germany). DNA was quantitated using a Nanodrop spectrophotometer (Thermofisher Scientific, Waltham, MA, USA). Each plant was genotyped using the PepperSNP16K array (Hulse-Kemp et al., 2016). As parents were confirmed to be double haploids using the array, uniform F_1 individuals were utilized downstream in combination with the two single parental plants to represent a trio (mother-father-offspring).

4.2 Parental sequencing

The double haploid parents, HDA149 and HDA330 were sequenced with short read sequencing. TruSeq PCR-free libraries

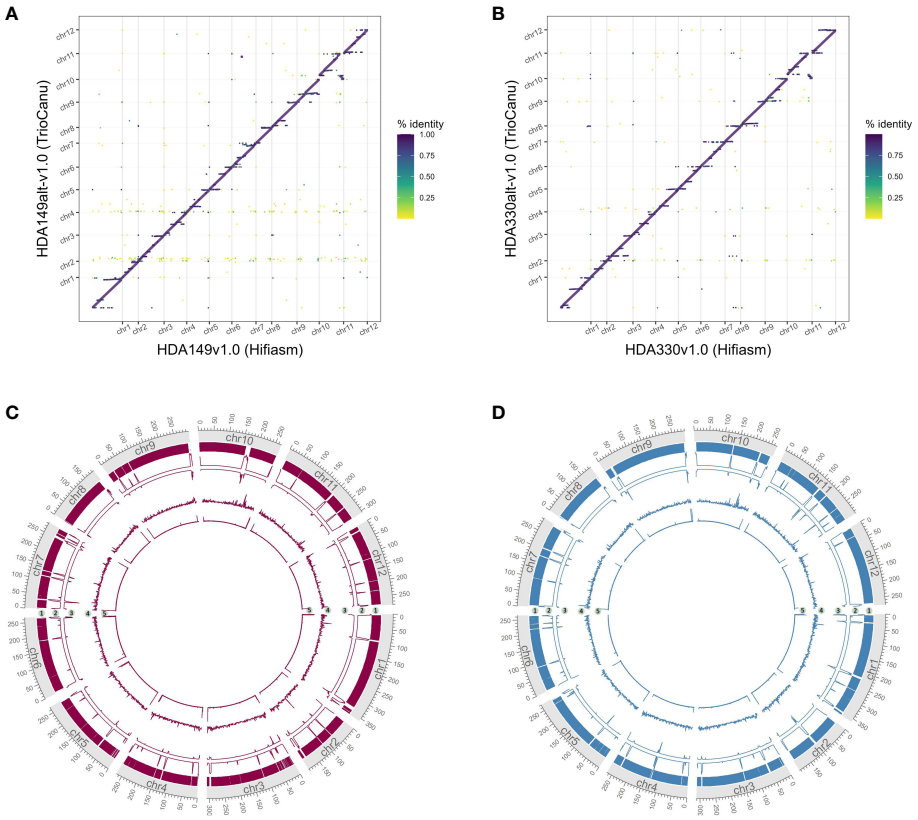


FIGURE 6
Comparison of final TrioCanu and Hifiasm assemblies. Dotplots of assembly by assembly alignments of **(A)** HDA149alt-v1.0 (TrioCanu) to HDA149v1.0 (Hifiasm), **(B)** HDA330alt-v1.0 (TrioCanu) to HDA330v1.0 (Hifiasm). Gridlines show boundaries of chromosomes (x-axis) and color indicates percent identity of the alignment. Circos plot of final TrioCanu assemblies HDA149alt-v1.0 **(C)** and HDA330alt-v1.0 **(D)** show regions of shared sequence to the corresponding final Hifiasm assembly in track 1, the number of gaps across 1 MB windows in track 2, the HiFi read alignment coverage across 1 MB windows in track 3, long terminal repeat content across 1 MB windows in track 4, and telomere repeat peaks across 1 kb windows in track 5.

were prepared and the samples were run on the Illumina Novaseq 6000 (Illumina, San Diego, CA, USA) which generated 150 bp paired end reads. The raw sequencing coverage of HDA149 was 49.3x and coverage of HDA330 was 45.3x. Raw sequencing data is available through the NCBI sequence read archive under SRR21710630 (HDA149) and SRR21710629 (HDA330).

The quality of the Illumina sequencing data was checked with FastQC version 0.11.9 (Andrews, 2010). Fastp v0.21.0 (Chen et al., 2018) was used to trim the first 12 bp, as well as remove any poly-g tails and adapter sequences. A minimum length of 50 bp was also required for a read to pass quality filtering. The resulting coverage for HDA149 was 45.0x and 41.4x for HDA330.

4.3 Hybrid sequencing

Uniform F_1 hybrid individuals (HDA149 x HDA330) were grown in greenhouse conditions then dark treated for 48 hours, unexpanded leaf tissue was flash frozen in liquid nitrogen. Nuclei were isolated from 1 gram of young leaf tissue using the Bionano Prep Plant Tissue DNA Isolation kit (Bionano Genomics, San Diego, CA). Subsequently, high molecular weight (HMW) genomic DNA was extracted from the nuclei using the Circulomics Nanobind Plant Nuclei Big DNA Kit (Pacific Biosciences, Menlo Park, CA). HMW DNA was sheared with the Covaris g-TUBE (Woburn, MA) to target fragments near 15Kb. Sheared HMW DNA was used to prepare a PacBio SMRTbell library and size selected using the BluePippin (Sage Science, Beverly, MA). The library was sequenced on 7 cells of the Sequel IIe (Pacific Biosciences, Menlo Park, CA), generating 11,829,089 PacBio HiFi reads equivalent to 58x coverage of the 3.5 Gb *C. annuum* L. genome. The PacBio HiFi reads are available through the NCBI sequence read archive under BioProject PRJNA884326. HiFiAdapterFilt identified and removed ~0.05% of PacBio HiFi reads that had adapter contamination (https://github.com/sheinasim/HiFiAdapterFilt, accessed 2023).

4.4 Bionano optical mapping

Optical mapping was done with ultra HMW genomic DNA of the F_1. Briefly, specific genomic sequences were fluorescently labeled with the Direct Label Enzyme-1 of the Bionano Prep Direct Label and Stain (DLS) kit (Bionano Genomics, San Diego, CA, USA) and imaged using the Saphyr system (Bionano Genomics, San Diego, CA, USA).

4.5 Genome assembly

Genomes were assembled using the two trio-binning softwares available as of 2022, TrioCanu (Koren et al., 2018) and Hifiasm (Cheng et al., 2021). TrioCanu and Hifiasm differ in their trio-binning approach, with TrioCanu binning the HiFi reads prior to assembly and Hifiasm binning contigs after assembly. Both rely on distinct parental k-mers from short reads to bin the long reads or contigs. Briefly, TrioCanu takes the trimmed parental short reads as input under the '-haplotype' option and finds haplotype-distinct 21-mers with the k-mer counting software meryl (Rhie et al., 2020). The alignment of the haplotype-distinct 21-mers to HiFi reads is used to determine to which bin a given HiFi read belongs, i.e. to parental haplotype HDA149 or HDA330. If a haplotype cannot be confidently assigned to either haplotype then the read is placed in an 'unknown' fasta bin. Given that HiFi reads have low sequencing errors, these unknown reads are primarily the shared sequences between the two haplotypes. TrioCanu results in three fasta files, parental haplotype 1, parental haplotype 2 and unknown haplotype (shared). Alternatively, Hifiasm applies a similar principle to assembled contigs (haplotigs) to partition into the corresponding parental haplotype.

TrioCanu v2.2 does not yet directly support genome assembly with HiFi reads, but HiCanu (Nurk et al., 2020) does. Therefore, binning and assembly were run in two steps, the first with TrioCanu and the second with HiCanu. HiFi read binning was accomplished with 'canu -p binned_reads -d binned_reads -haplotypeHDA149 illumina-HDA149*fq.gz -haplotypeHDA330 illumina-HDA330*fq.gz -pacbio HiFi-reads*fq.gz'. TrioCanu stops after binning because the HiFi reads appear to be corrected CLR reads. In the second step, assemblies are made with 'canu -p TrioCanu_HDA149_assembly -d TrioCanu_HDA149_assembly genomeSize=3.5g -pacbio-hifi binned_reads/haplotype/haplotype-HDA149.fasta.gz binned_reads/haplotype/haplotype-unknown.fasta.gz'. The same script was run for HDA330, but with the corresponding HDA330 binned reads. For simplicity, we refer to these assemblies as TrioCanu-HDA149 and TrioCanu-HDA330.

Parental k-mers for Hifiasm are first generated with yak v0.1(r56) (https://github.com/lh3/yak, accessed 2023) using 'count -k31 -b37' settings. Yak does not support multiple input files so it is necessary to first concatenate all sequence files for a parent into a single file. The yak dumps are supplied to Hifiasm version 0.16.1-r375 for assembly of the HiFi reads into the two haplotypes with the command 'hifiasm -1 HDA149.yak -2 HDA330.yak HiFi-reads*fq.gz'. We refer to the resulting assemblies Hifiasm HDA149 and Hifiasm HDA330.

For comparison, we also generated a Hifiasm assembly without trio-binning because the software attempts to partition haplotigs even without parental k-mers to inform it. The script was 'hifiasm HiFi-reads*fq.gz'. We called these assemblies Hifiasm-F1-Hap1 and Hifiasm-F1-Hap2.

4.6 Scaffolding

The four assemblies were scaffolded with the F_1 Bionano optical map and designated as Hifiasm-HDA149.BN, Hifiasm-HDA330.BN, TrioCanu-HDA149.BN and TrioCanu-HDA330.BN. Scaffolding of contigs was accomplished in 5 steps using an iterative scaffolding workflow (Figure 1B, Supplemental Table 1). Step 1 leveraged differences in the two assembly softwares by reciprocally scaffolding TrioCanu assemblies to Hifiasm assemblies using Ragtag v2.1.0 'scaffold' option and default parameters (Alonge et al., 2022). Step 2 patched gaps in assemblies using Ragtag 'patch' with minimap2 as the aligner. Step 3 scaffolded the alternative parent haplotype

assemblies against each other using Ragtag 'scaffold'. Step 4 scaffolded the assemblies to the corresponding Bionano contig-assembly hybrid scaffolds using Ragtag 'scaffold'. Step 5 scaffolded assemblies to Dempsey v1.0 (Lee et al., 2022) to order and orient chromosomes using RagTag 'scaffold'. The resulting assemblies became the publicly released versions.

4.7 Analysis and visualization

Genome statistics were retrieved using stats.sh of the BBTools suite version 38.79 (Bushnell, 2022). BUSCO v5.2.2 in genome mode with the embryophyta_odb10 database was used for calculating completeness scores (Manni et al., 2021). Long-terminal repeat assembly index (LAI) values were found using the LAI software with default parameters, analysis was run locally as well as using the webportal at https://bioinformatics.um6p.ma/PlantLAI/lai-pipeline (accessed 8/2023) (Ou et al., 2018; Mokhtar and Allali, 2022; Mokhtar et al., 2023).

Trio-binning and scaffolding workflow figures (Figure 1) were made in BioRender.com.

Haplotype switching (Figure 2) in assemblies was determined by aligning TrioCanu binned reads (HDA149 and HDA330) to the given assembly with minimap2 version 2.24-r1122 (Li, 2018). Assembly files were indexed with samtools version 1.9 (Danecek et al., 2021) and 1 Mb windows were made with bedtools version 2.30.0 'makewindows'(Quinlan and Hall, 2010). Read coverage for each window was calculated with bedtools 'coverage'. The difference in percent coverage of each window for HDA149 and HDA330 binned reads was calculated in RStudio version 2022.07.2 + 576 (RStudio Team, 2020) with R version 4.2.1 (R Core Team, 2022) and plotted with ggplot2 version 3.3.6 (Wickham, 2016). Large positive differences in coverage meant HDA149 reads covered more of the window than HDA330 reads did and therefore that the region is haplotype specific to HDA149.

Sample specific telomere repeats were identified with the 'explore' function of Telomere Identification Toolkit, tidk (https://github.com/tolkit/telomeric-identifier, accessed 2023) on the final genome assemblies with a minimum string length of 5 and maximum length of 12. The top hits 'AAAAATAGTAG' and 'TTAGGG' were searched in the final genome assemblies with default settings of the tidk 'search' function. LTRharvest v2.9.4 of Genome Tools (Gremme et al., 2013), with specifications of '-minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1 -similar 85 -vic 10 -seed 20 -seqids yes' was used to annotate long terminal repeat retrotransposons in the final assemblies. Seqkit v0.10.1 (Shen et al., 2016) 'locate' was used to find gaps in assemblies by searching for strings of Ns. LTR content over 1Mb windows, telomere repeat counts over 1 kb windows, and gap locations were visualized with the circlize package (Gu et al., 2014) in R.

Dotplots in Figures 3, 5, 6 were generated using minimap2 and modified dotplotly code (https://github.com/tpoorten/dotPlotly, accessed 2022) in RStudio with R. The dotplotly code uses the R packages dplyr version 1.0.10 (Wickham et al., 2023) and ggplot2 (Wickham, 2016). Scripts for figures can be found at the Github repository https://github.com/USDA-ARS-GBRU/Pepper_TrioBinning/.

Genome wide heterozygosity was calculated as the number of single nucleotide polymorphism and insertion/deletion positions from unique alignments between the final Hifiasm assemblies HDA149v1.0 and HDA330v1.0. Alignments were made with the default settings of nucmer in mummer v4.0.0rc1 (Marçais et al., 2018). Variant positions of unique alignments were called from the delta file with 'show-snps -C' of mummer v4.0.0rc1.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

EED, WBR, AMH-K conceived the project. RCY, SAS, BES performed sequencing and generated raw data. EED analyzed data and wrote the manuscript. ANS and RCY participated in data analysis. WBR and AMH-K supervised the project. All authors contributed to the article and approved the submitted version.

Funding

Support was provided by USDA-ARS research project numbers 6080-22000-031-000D and 6066-21310-005-00-D. This research used resources provided by the SCINet project of the USDA Agricultural Research Service, ARS project number 0500-00093-001-00-D. Support for EED was provided by NSF Postdoctoral Fellowship Award Number: 2010930.

Acknowledgments

The authors would like to thank the staff at USDA-ARS US Vegetable Laboratory for helping to maintain the pepper plants. We thank Catherine Wram for preliminary review and suggestions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1184112/full#supplementary-material

SUPPLEMENTARY DATA SHEET 1
Genotyping data from parental samples on the PepperSNP16K array.

SUPPLEMENTARY DATA SHEET 2
Detailed assembly statistics.

SUPPLEMENTARY TABLE 1
Scaffolding workflow details.

SUPPLEMENTARY TABLE 2
Final assembly availability links.

References

Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., et al. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* 23, 1–19. doi: 10.1186/S13059-022-02823-7

Andrews, S. (2010) *FastQC: a quality control tool for high throughput sequence data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., and Edwards, D. (2020). Plant pan-genomes are the new reference. *Nat. Plants* 6 (8), 914–920. doi: 10.1038/s41477-020-0733-0

Belletti, P., Marzachì, C., and Lanteri, S. (1998). Flow cytometric measurement of nuclear DNA content in Capsicum (Solanaceae). *Plant System. Evol.* 209, 85–91. doi: 10.1007/BF00991526

Benevenuto, J., Ferrão, L. F. V., Amadeu, R. R., and Munoz, P. (2019). How can a high-quality genome assembly help plant breeders? *Gigascience* 8, 1–4. doi: 10.1093/GIGASCIENCE/GIZ068

Bushnell, B. (2022) *BBTools*. Available at: sourceforge.net/projects/bbmap/.

Campoy, J. A., Sun, H., Goel, M., Jiao, W.-B., Folz-Donahue, K., Wang, N., et al. (2020). Gamete binning: chromosome-level and haplotype-resolved genome assembly enabled by high-throughput single-cell sequencing of gamete genomes. *Genome Biol.* 21, 306. doi: 10.1186/s13059-020-02235-5

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/BIOINFORMATICS/BTY560

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18 (2), 170–175. doi: 10.1038/s41592-020-01056-5

Cheng, H., Jarvis, E. D., Fedrigo, O., Koepfli, K. P., Urban, L., Gemmell, N. J., et al. (2022). Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* 40 (9), 1332–1335. doi: 10.1038/s41587-022-01261-x

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13 (12), 1050–1054. doi: 10.1038/nmeth.4035

Core Team, R. (2022). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing). Available at: https://www.R-project.org/.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *Gigascience* 10, 1–4. doi: 10.1093/GIGASCIENCE/GIAB008

Gladman, N., Goodwin, S., Chougule, K., Richard McCombie, W., and Ware, D. (2023). Era of gapless plant genomes: innovations in sequencing and mapping technologies revolutionize genomics and breeding. *Curr. Opin. Biotechnol.* 79, 102886. doi: 10.1016/J.COPBIO.2022.102886

Gremme, G., Steinbiss, S., and Kurtz, S. (2013). GenomeTools: A comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 10, 645–656. doi: 10.1109/TCBB.2013.68

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize implements and enhances circular visualization in R. *Bioinformatics* 30, 2811–2812. doi: 10.1093/BIOINFORMATICS/BTU393

Hendy, H., Pochard, E., Dalmasso, A., and Bongiovanni, M. (1985). Transmission héréditaire de la résistance aux nématodes Meloidogyne Chitwood (Tylenchida) portée par 2 lignées de Capsicum annuum L. : étude de descendances homozygotes issues d'androgenèse. *Agronomie* 5, 93–100. doi: 10.1051/AGRO:19850201

Huang, Y., Wang, H., Zhu, Y., Huang, X., Li, S., Wu, X., et al. (2022). THP9 enhances seed protein content and nitrogen-use efficiency in maize. *Nature* 612 (7939), 292–300. doi: 10.1038/s41586-022-05441-2

Hulse-Kemp, A. M., Ashrafi, H., Plieske, J., Lemm, J., Stoffel, K., Hill, T., et al. (2016). A HapMap leads to a Capsicum annuum SNP infinium array: A new tool for pepper breeding. *Hortic. Res.* 3. doi: 10.1038/HORTRES.2016.36

Hulse-Kemp, A. M., Maheshwari, S., Stoffel, K., Hill, T. A., Jaffe, D., Williams, S. R., et al. (2018). Reference quality assembly of the 3.5-Gb genome of Capsicum annuum from a single linked-read library. *Hortic. Res.* 5. doi: 10.1038/S41438-017-0011-0

Jiao, W. B., and Schneeberger, K. (2020). Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.* 11 (1), 1–10. doi: 10.1038/s41467-020-14779-y

Kim, S., Park, M., Yeom, S. I., Kim, Y. M., Lee, J. M., Lee, H. A., et al. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nat. Genet.* 46 (3), 270–278. doi: 10.1038/ng.2877

Koren, S., Rhie, A., Walenz, B. P., Dilthey, A. T., Bickhart, D. M., Kingan, S. B., et al. (2018). *De novo* assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* 36 (12), 1174–1182. doi: 10.1038/nbt.4277

Kress, W. J., Soltis, D. E., Kersey, P. J., Wegrzyn, J. L., Leebens-Mack, J. H., Gostel, M. R., et al. (2022). Green plant genomes: What we know in an era of rapidly expanding opportunities. *Proc. Natl. Acad. Sci.* 119, e2115640118. doi: 10.1073/pnas.2115640118

Kronenberg, Z. N., Rhie, A., Koren, S., Concepcion, G. T., Peluso, P., Munson, K. M., et al. (2021). Extended haplotype-phasing of long-read *de novo* genome assemblies using Hi-C. *Nat. Commun.* 12 (1), 1–10. doi: 10.1038/s41467-020-20536-y

Lee, J. H., Venkatesh, J., Jo, J., Jang, S., Kim, G. W., Kim, J. M., et al. (2022). High-quality chromosome-scale genomes facilitate effective identification of large structural variations in hot and sweet peppers. *Hortic. Res.* 9. doi: 10.1093/HR/UHAC210

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/BIOINFORMATICS/BTY191

Li, K., Jiang, W., Hui, Y., Kong, M., Feng, L. Y., Gao, L. Z., et al. (2021). Gapless indica rice genome reveals synergistic contributions of active transposable elements and segmental duplications to rice genome evolution. *Mol. Plant* 14, 1745–1756. doi: 10.1016/J.MOLP.2021.06.017

Liu, F., Zhao, J., Sun, H., Xiong, C., Sun, X., Wang, X., et al. (2023). Genomes of cultivated and wild Capsicum species provide insights into pepper domestication and population differentiation. *Nat. Commun.* 14 (1), 1–14. doi: 10.1038/s41467-023-41251-4

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., et al. (2020). Pan-genome of wild and cultivated soybeans. *Cell* 182, 162–176.e13. doi: 10.1016/J.CELL.2020.05.023

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., and Zdobnov, E. M. (2021). 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. doi: 10.1093/MOLBEV/MSAB199

Mao, J., Wang, Y., Wang, B., Li, J., Zhang, C., Zhang, W., et al. (2023). High-quality haplotype-resolved genome assembly of cultivated octoploid strawberry. *Hortic. Res.* 10. doi: 10.1093/HR/UHAD002

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PloS Comput. Biol.* 14. doi: 10.1371/JOURNAL.PCBI.1005944

Mc Cartney, A. M., Shafin, K., Alonge, M., Bzikadze, A. V., Formenti, G., Fungtammasan, A., et al. (2022). Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat. Methods* 19 (6), 687–695. doi: 10.1038/s41592-022-01440-3

Minio, A., Cochetel, N., Vondras, A. M., Massonnet, M., and Cantu, D. (2022). Assembly of complete diploid-phased chromosomes from draft genome sequences. *G3 Genes Genom. Genet.* 12. doi: 10.1093/G3JOURNAL/JKAC143

Minio, A., Lin, J., Gaut, B. S., and Cantu, D. (2017). How single molecule real-time sequencing and haplotype phasing have enabled reference-grade diploid genome assembly of wine grapes. *Front. Plant Sci.* 8. doi: 10.3389/FPLS.2017.00826

Mokhtar, M. M., Abd-Elhalim, H. M., and El Allali, A. (2023). A large-scale assessment of the quality of plant genome assemblies using the LTR assembly index. *AoB Plants* 15, 1–8. doi: 10.1093/AOBPLA/PLAD015

Mokhtar, M. M., and Allali, A. E. L. (2022). PltRNAdb: Plant transfer RNA database. *PloS One* 17, e0268904. doi: 10.1371/JOURNAL.PONE.0268904

Montgomery, J. S., Giacomini, D., Waithaka, B., Lanz, C., Murphy, B. P., Campe, R., et al. (2020). Draft Genomes of Amaranthus tuberculatus, Amaranthus hybridus, and Amaranthus palmeri. *Genome Biol. Evol.* 12, 1988–1993. doi: 10.1093/GBE/EVAA177

Moscone, E. A., Baranyi, M., Ebert, I., Greilhuber, J., Ehrendorfer, F., and Hunziker, A. T. (2003). Analysis of nuclear DNA content in capsicum (Solanaceae) by flow cytometry and feulgen densitometry. *Ann. Bot.* 92, 21. doi: 10.1093/AOB/MCG105

Newman, C. S., Andres, R. J., Youngblood, R. C., Campbell, J. D., Simpson, S. A., Cannon, S. B., et al. (2023). Initiation of genomics-assisted breeding in Virginia-type peanuts through the generation of a *de novo* reference genome and informative markers. *Front. Plant Sci.* 13. doi: 10.3389/FPLS.2022.1073542

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., et al. (2022). The complete sequence of a human genome. *Sci. (1979)* 376, 44–53. doi: 10.1126/science.abj6987

Nurk, S., Walenz, B. P., Rhie, A., Vollger, M. R., Logsdon, G. A., Grothe, R., et al. (2020). HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* 30, 1291–1305. doi: 10.1101/GR.263566.120

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* 46, e126. doi: 10.1093/NAR/GKY730

Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., et al. (2014). Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proc. Natl. Acad. Sci.* 111, 5135–5140. doi: 10.1073/pnas.1400975111

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/BIOINFORMATICS/BTQ033

Rautiainen, M., Nurk, S., Walenz, B. P., Logsdon, G. A., Porubsky, D., Rhie, A., et al. (2023). Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* 2023, 1–9. doi: 10.1038/s41587-023-01662-6

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592 (7856), 737–746. doi: 10.1038/s41586-021-03451-0

Rhie, A., Walenz, B. P., Koren, S., and Phillippy, A. M. (2020). Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* 21, 1–27. doi: 10.1186/S13059-020-02134-9

RStudio Team (2020). *RStudio: Integrated Development for R* (Boston, MA: RStudio, PBC). Available at: http://www.rstudio.com/.

Sahu, S. K., and Liu, H. (2023). Long-read sequencing (method of the year 2022): The way forward for plant omics research. *Mol. Plant* 16, 791–793. doi: 10.1016/J.MOLP.2023.04.007

Seo, J. S., Rhie, A., Kim, J., Lee, S., Sohn, M. H., Kim, C. U., et al. (2016). *De novo* assembly and phasing of a Korean human genome. *Nature* 538 (7624), 243–247. doi: 10.1038/nature20098

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A Cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS One* 11. doi: 10.1371/JOURNAL.PONE.0163962

Shi, D., Wu, J., Tang, H., Yin, H., Wang, H., Wang, R., et al. (2019). Single-pollen-cell sequencing for gamete-based phased diploid genome assembly in plants. *Genome Res.* 29, 1889–1899. doi: 10.1101/GR.251033.119

Shirasawa, K., Hosokawa, M., Yasui, Y., Toyoda, A., and Isobe, S. (2022). Chromosome-scale genome assembly of a Japanese chili pepper landrace, Capsicum annuum "Takanotsume." *DNA Res* 30. doi: 10.1093/DNARES/DSAC052

Tang, D., Jia, Y., Zhang, J., Li, H., Cheng, L., Wang, P., et al. (2022). Genome evolution and diversity of wild and cultivated potatoes. *Nature* 606 (7914), 535–541. doi: 10.1038/s41586-022-04822-x

Thies, J. A., and Ariss, J. J. (2009). Comparison between the N and Me3 genes conferring resistance to the root-knot nematode (Meloidogyne incognita) in genetically different pepper lines (Capsicum annuum). *Eur. J. Plant Pathol.* 125, 545–550. doi: 10.1007/S10658-009-9502-7

Vaughn, J. N., Branham, S. E., Abernathy, B., Hulse-Kemp, A. M., Rivers, A. R., Levi, A., et al. (2022). Graph-based pangenomics maximizes genotyping density and reveals structural impacts on fungal resistance in melon. *Nat. Commun.* 13 (1), 1–14. doi: 10.1038/s41467-022-35621-7

Wang, X., Gao, L., Jiao, C., Stravoravdis, S., Hosmani, P. S., Saha, S., et al. (2020). Genome of Solanum pimpinellifolium provides insights into structural variants during tomato breeding. *Nat. Commun.* 11 (1), 1–11. doi: 10.1038/s41467-020-19682-0

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 37 (10), 1155–1162. doi: 10.1038/s41587-019-0217-9

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (Verlag New York: Springer). Available at: https://ggplot2.tidyverse.org.

Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. (2023) *dplyr: A Grammar of Data Manipulation*. Available at: https://github.com/tidyverse/dplyr.

Yang, T., Liu, R., Luo, Y., Hu, S., Wang, D., Wang, C., et al. (2022). Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics. *Nat. Genet.* 54 (10), 1553–1563. doi: 10.1038/s41588-022-01172-2

Yang, C., Zhou, Y., Marcus, S., Formenti, G., Bergeron, L. A., Song, Z., et al. (2021). Evolutionary and biomedical insights from a marmoset diploid genome assembly. *Nature* 594 (7862), 227–233. doi: 10.1038/s41586-021-03535-x

Yen, E. C., McCarthy, S. A., Galarza, J. A., Generalovic, T. N., Pelan, S., Nguyen, P., et al. (2020). A haplotype-resolved, *de novo* genome assembly for the wood tiger moth (Arctia plantaginis) through trio binning. *Gigascience* 9, 1–12. doi: 10.1093/GIGASCIENCE/GIAA088

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., et al. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* 606 (7914), 527–534. doi: 10.1038/s41586-022-04808-9

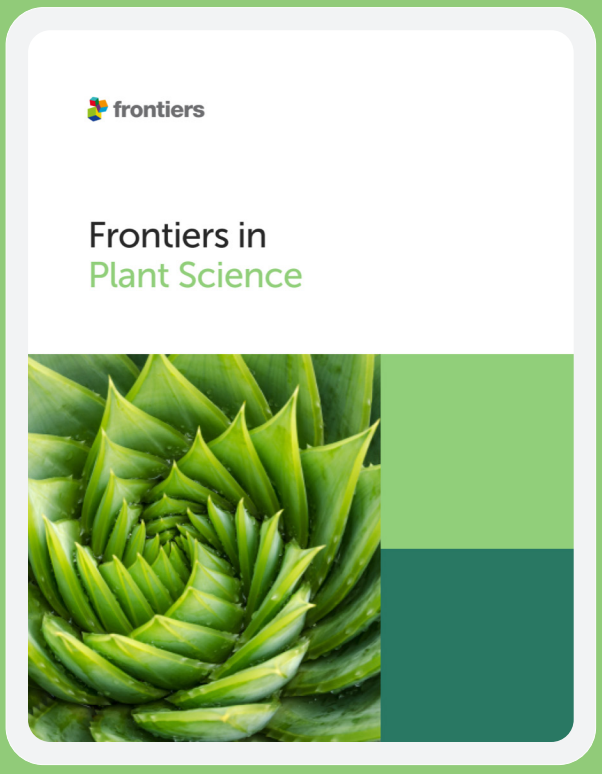
Frontiers in
Plant Science

Cultivates the science of plant biology and its applications

The most cited plant science journal, which advances our understanding of plant biology for sustainable food security, functional ecosystems and human health.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact