




Meliaceae genomes provide insights into wood development and limonoids biosynthesis

Gaofeng Cui^{1,2,3,†} , Yun Li^{1,2,†}, Xin Yi^{1,2,†} , Jieyu Wang⁴, Peifan Lin³, Cui Lu³, Qunjie Zhang³, Lizhi Gao^{5,*} and Guohua Zhong^{1,2,*} 

¹College of Plant Protection, South China Agricultural University, Guangzhou, China

²Key Laboratory of Natural Pesticide & Chemical Biology, Ministry of Education, South China Agricultural University, Guangzhou, China

³Institution of Genomics and Bioinformatics, South China Agricultural University, Guangzhou, China

⁴Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China

⁵Engineering Research Center for Selecting and Breeding New Tropical Crop Varieties, Ministry of Education, College of Tropical Crops, Hainan University, Haikou, China

Received 29 August 2022;

revised 20 November 2022;

accepted 25 November 2022.

*Correspondence (Tel/fax +86-0898-66273817; email lgaogenomics@163.com (L. G.); Tel/fax +86-020-85280308; email guohuazhong@scau.edu.cn (G. Z.)

[†]These authors contributed equally to this work as co-first authors.

Summary

Meliaceae is a useful plant family owing to its high-quality timber and its many limonoids that have pharmacological and biological activities. Although some genomes of Meliaceae species have been reported, many questions regarding their unique family features, namely wood quality and natural products, have not been answered. In this study, we provide the whole-genome sequence of *Melia azedarach* comprising 237.16 Mb with a contig N50 of 8.07 Mb, and an improved genome sequence of *Azadirachta indica* comprising 223.66 Mb with a contig N50 of 8.91 Mb. Moreover, genome skimming data, transcriptomes and other published genomes were comprehensively analysed to determine the genes and proteins that produce superior wood and valuable limonoids. Phylogenetic analysis of chloroplast genomes, single-copy gene families and single-nucleotide polymorphisms revealed that Meliaceae should be classified into two subfamilies: Cedreloideae and Melioideae. Although the Meliaceae species did not undergo additional whole-genome duplication events, the secondary wall biosynthetic genes of the woody Cedreloideae species, *Toona sinensis*, expanded significantly compared to those of *A. indica* and *M. azedarach*, especially in downstream transcription factors and cellulose/hemicellulose biosynthesis-related genes. Moreover, expanded special oxidosqualene cyclase catalogues can help diversify Sapindales skeletons, and the clustered genes that regulate terpene chain elongation, cyclization and modification would support their roles in limonoid biosynthesis. The expanded clans of terpene synthase, O-methyltransferase and cytochrome P450, which are mainly derived from tandem duplication, are responsible for the different limonoid classes among the species. These results are beneficial for further investigations of wood development and limonoid biosynthesis.

Keywords: Meliaceae, phylogenetic relationship, tandem duplication, wood development, limonoid biosynthesis.

Introduction

Meliaceae (Sapindales), with 56 genera and 575–650 species, are common canopy and understory trees in tropical and subtropical forests and pervade rain forests, mangrove swamps and semidesert regions (He *et al.*, 2022; Kubitzki, 2011; Mabberley *et al.*, 1995; Pennington and Styles, 1975). Although the family Meliaceae exhibits a diversity of vegetative morphology and growth properties, it is best known for its high-quality timber and many limonoids. The graceful and woodworm-proof mahogany furniture had higher market share compared to oak and walnut furniture in Europe (Kubitzki, 2011; Mabberley *et al.*, 1995; Pennington and Styles, 1975). Additionally, Meliaceae limonoids are unique and show insect antifeedant and growth-regulating properties, medicinal effects in humans and other animals, and antifungal, bactericidal and antiviral activities (Tan and Luo, 2011; Tundis *et al.*, 2014).

Many genera of Meliaceae are economically important to the timber industry, and some are the most sought after worldwide. For example, the famous mahogany mainly derived from

neotropical *Swietenia macrophylla*. Asiatic toon, *Toona ciliata*, and *Chukrasia tabularis* have been the most desirable timber in India and Australia. Other famous species include the neotropical *Cedrela odorata* and the African genera *Entandrophragma* (sapele and utile), *Khaya* (African mahogany) and *Lovoa* (Nigerian golden walnut). In the Old World, *Azadirachta indica* and *Melia azedarach* have also been grown as shade or avenue trees as some Melioideae trees, notably of *Azadirachta* and *Dysoxylum* species, have sulphur-containing volatiles with onion- or garlic-like smells (Kubitzki, 2011; Mabberley *et al.*, 1995). The most widely grown and important timber type is Cedreloideae. The wood anatomy of Meliaceae, which is well documented owing to its economic significance, has exhibited different characteristics of the secondary xylem of the subfamily Melioideae and Swietenioideae (now included in Cedreloideae) (Oyediji Amusa *et al.*, 2020; Riesco Muñoz *et al.*, 2019).

One characteristic of the Sapindales order is the synthesis of nortriterpenoids derived from tetracyclic triterpenes, which are known as protolimonoids (Kubitzki, 2011). The term 'limonoids'

originated from the bitterness of lemon or other citrus. Limonoids mainly occur in plants of the Meliaceae and Rutaceae families. The bitterness of Meliaceae bark has long been known and used in medicine and pharmaceutical applications (Bao *et al.*, 2016; Tan and Luo, 2011; Tundis *et al.*, 2014). For example, a traditional Chinese medicine formulation containing toosendanin, a limonoid from *M. azedarach* and *M. toosendan*, displayed anti-botulism effects and has been used as an anthelmintic vermifuge against ascarids (Lian *et al.*, 2020). The neem tree, *A. indica*, one of the most famous limonoid-producing plants, produces special limonoids, such as azadirachtin, that are bioactive against approximately 400 species from more than 10 important insect orders, including antifeedant, growth and development inhibition, repellent, gastric toxicity and sterilization (Saleem *et al.*, 2018). Limonoids are significant chemotaxonomic markers of Meliaceae, Rutaceae and Simaroubaceae; the chemotaxonomy significances of Meliaceae limonoids mainly focused on the subfamily Swietenioideae and Melioideae (Fernandes Da Silva *et al.*, 2021; Tan and Luo, 2011). For example, almost all citrus limonoids belong to the ring A,D-seco limonoids, which are found only in the *Toona*, *Cedrela* and *Dysoxylum* genera in Meliaceae. Ring C-seco limonoids possess the highest bioactivities and originate from the *Azadirachta* and *Melia* genera. The ring B-seco limonoids are found only in the *Turraea* and *Toona* genera (Fernandes Da Silva *et al.*, 2021; Tan and Luo, 2011).

In addition to these features, the affinities with other Sapindales families, relationships within the family, karyology and palaeobotany of Meliaceae are important issues in understanding its developmental biology and evolutionary trajectory (Muellner *et al.*, 2008, 2011; Muellner-Riehl *et al.*, 2016). Despite the importance of the Meliaceae family, genetic research on this family is not well-developed. The low genome contiguity of *A. indica*, which was obtained with HiSeq short Illumina reads (Krishnan *et al.*, 2012, 2016; Kuravadi *et al.*, 2015), limited its utility for downstream genomic research, and the chromosome-level genome of *Toona sinensis* provided limited information on evolutionary history and genetic variation (Ji *et al.*, 2021). Recently, a chromosome-level genome of *A. indica* was assembled to reveal terpene biosynthesis, and the results revealed that most *A. indica*-specific terpene synthase (TPS) genes and cytochrome P450 (CYP) genes were located on chromosome 13 (Du *et al.*, 2022). However, many more questions related to Meliaceae features have not been settled and require attention. Overall, the high-quality genomes of *A. indica* and *M. azedarach* and the genome resequencing of many other related species would help clarify the phylogenetic relationship of Meliaceae members. A wide and complete investigation would help to uncover the genetic mechanisms behind the special features of the Meliaceae family, including timber quality and limonoid biosynthesis.

Results

Improved *A. indica* genome sequencing and assembly of the *M. azedarach* genome

Based on the results of k-mer counting, both *A. indica* and *M. azedarach* had a small genome with lower heterozygosity (approximately 0.3%, k-mer = 17). The estimated genome sizes were 267.28 Mb for *A. indica* and 270.17 Mb for *M. azedarach* (Figure S1A,B). The Oxford Nanopore Technologies (ONT) platform generated a total of 50.27 Gb and 35.99 Gb subreads of *A. indica* and *M. azedarach* respectively. After filtering, 19.12 Gb

and 16.37 Gb of clean bases were used to generate the primary genome assembly (Table S1, Figure S1C,F). After self-correcting and polishing, the analyses yielded a final assembly of 232.68 Mb with a contig N50 value of 8.91 Mb for *A. indica*, covering 87.05% of the estimated genome and an assembly of 239.23 Mb with a contig N50 of 8.07 Mb for *M. azedarach*, covering 88.54% of the estimated genome (Table 1 and Table S3). The Benchmarking Universal Single-Copy Orthologs (BUSCO) results showed 99.07% completeness for *A. indica* and 96.05% completeness for *M. azedarach*. In Core Eukaryotic Gene Mapping Approach (CEGMA) estimation, both of the two genome assemblies had 231 complete core genes (93.15%), suggesting completeness (Table S2). The assemblies were further assessed using RNA sequencing transcripts and Illumina short reads. The mapping rates and average coverage depth of the third-generation data of the two genome assemblies reached 97.82% and 200.17× for *A. indica* and 95.15% and 134.43× for *M. azedarach*. Furthermore, a total of 41.93 Gb and 43.30 Gb clean bases were obtained from the Illumina sequencing platform for HiC-based assembly of *A. indica* and *M. azedarach* respectively (Table S4). After paired-end mapping and valid interaction pair filtering, 44.18% and 45.97% of the clean data were used for contig clustering. Finally, 96.13% and 99.14% of the assembled genome were anchored to 14 pseudo-chromosomes comprising 223.66 Mb in *A. indica* and 14 chromosomes comprising 237.16 Mb in *M. azedarach* (Figure S1D,G). The chromosome lengths of *A. indica* ranged from 12.42 Mb to 21.50 Mb, while that of *M. azedarach* ranged from 14.01 Mb to 22.81 Mb (Table S5).

Gene repeats and annotations

In the repeat annotation, a total of 71.36 Mb and 84.01 Mb transposon elements (TEs), comprising 30.67% and 35.12% of the whole-genome size, were identified in the *A. indica* and *M. azedarach* genome respectively (Table S7). Among all the classifications of TEs, long terminal repeat retrotransposons (LTR-RTs) constituted the largest portion (17.42% and 23.28% respectively), followed by DNA transposons (10.11% and 8.53% respectively). Accompanied by tandem repeats, and simple repeats, total repeats reached 33.11% and 37.47% of the two genomes, respectively, while 0.03% and 0.04% of the two genomes were identified as non-coding RNA (ncRNA), including miRNA, snRNA and spliceosomal RNA respectively (Table S6).

Protein-coding genes were annotated using a combination of RNA-seq data and *ab initio*-bases. Genome annotation yielded 23 087 and 21 983 protein-coding genes with 97.53% and 97.45% BUSCO completeness in *A. indica* and *M. azedarach* respectively (Table S8). 94.01% and 94.82% of the genes had related transcripts in the Non-Redundant Protein Sequence Database (NR), and most of them could be classified using Gene Ontology (GO) terms, Cluster of Orthologous Groups of proteins (COG/KOG) terms, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and the SwissProt database (Figure S1E,H). In total, 21 799 and 20 925 proteins were annotated (Table S9). As expected, the density of coding genes increased towards the telomeres with the opposite trend for repeated elements (Figure 1). Our results indicated that an improved chromosome-level *A. indica* genome, whose contig N50 increased almost 400-fold compared to the reference genome, and a new chromosome-level genome of the Meliaceae species, *M. azedarach*, which is widely distributed in China and Old World tropics, was obtained with high base accuracy, high continuity, high

Table 1 Statistics for published Meliaceae genomes

Material	Ain (Guangzhou)	Maz (Kunming)	Ain (India)	Ain (Azalnd)	Ain (GKVK)	Tsi (Fuyang)	Ain (Hainan)
Year	This study	This study	2012	2016	2015	2020	2022
Sequencing platform	Illumina, ONT, HiC	Illumina, ONT, HiC	Illumina, IonTorrent	Illumina, PacBio	Illumina, Roche/454	Illumina, ONT, HiC	Illumina, PacBio, HiC
Coverage	78.48×, 200.17×	68.56×, 134.43×	13–50×, 0.5×	13–50×, 5–9.7×	21×	67–117×, 146×	118×, 256×
Denovo	NextDenovo	NextDenovo	SOAP denovo	SOAP denovo2	Velvet	smartdenovo	RACON, Pilon
Genome size/Mb	232.68	239.23	364	182.93–308.83	267	596	281
contig N50/bp	8 907 986	8 068 821	740	3491	15 948	1 525 641	6 039 544
Longest Contig/bp	19 556 515	18 913 001	10 111	-	241 170	-	15 111 501
Longest Scaffold/bp	21 501 397	22 810 243	3 641 215	12 211 325	-	32 278 227	-
Number of Scaffolds	82	98	9714	21 743	-	-	70
Number of genes	23 087	21 983	20 169	32316.77	44 495	34 345	25 767
Average gene length/bp	3101.7	3308.55	1695.95	-	876 (CDS)	3959	2837
Total repeat element/bp	77 049 478	89 633 253	47 427 034	54 375 206	86.9 Mb	385 217 481	115 181 900
Repeat element Ratio/%	33.11	37.47	13.03	24.15	32.44	64.56	40.99
Level	Chromosome	Chromosome	Draft	Draft	Draft	Chromosome	Chromosome

Ain for *A. indica*, Maz for *M. azedarach*, Tsi for *T. sinensis*.

degree of genome coverage and more accurate gene structure annotation compared to previous attempts.

Phylogenetic and whole-genome duplication analyses

To reconstruct the phylogeny of Meliaceae, another nine species representing important lineages of Meliaceae were sampled and sequenced using the Illumina platform (Figure 2a and Table S10). After k-mer evaluation, the genome sizes ranged from 239.25 Mb to 867.80 Mb and the heterozygosity ranged from 0.34% to 1.18%. Moreover, genome repeat length and genome uniqueness were also assessed, but these indices had little relevance between the subfamilies. To investigate the genetic diversity and evolutionary history of Meliaceae species, phylogenetic relationships for the whole family were reconstructed using single-copy genes, chloroplast genomes and single nucleotide polymorphisms (SNP). The divergence times were then estimated using MCMCTree with calibrations of fossil evidence (Figure 2b and Figure S2). All three phylogenetic trees displayed congruent topology and divided Meliaceae into two clades. Clade one contained species that belonged to Cedreloideae (*Khaya senegalensis*, *Swietenia macrophylla*, *T. sinensis*, *Chukrasia tabularis* and *C. tabularis velutina*), whereas clade two contained Melioideae species (*M. azedarach*, *M. toosendan*, *A. indica*, *Aphanamixis grandifolia*, *Aglaiia duperreana* and *Trichilia conaroides macrocarpa*). The divergence time of Cedreloideae and Melioideae was approximately 22.16 million years ago (Mya), and that of *A. indica* and *M. azedarach* was approximately 6.45 Mya. In addition, Meliaceae was relatively closer to the family Rutaceae, which exhibited divergence at approximately 52.52 Mya, followed by Sapindaceae (approximately 53.51 Mya) and Anacardiaceae (approximately 72.06 Mya). However, the divergence times of Sapindales families were indistinct because the estimated divergence time of Sapindaceae was close to that of the Rutaceae. Thus, more evidence is needed to distinguish the topology of Sapindales.

To evaluate the historical whole-genome duplication (WGD) events of *A. indica* and *M. azedarach*, four-fold synonymous third-codon transversion rates (4DTv) and synonymous substitution rates (K_s) were analysed (Figure 2c and Figure S3). Self-comparison of the two genomes was performed using MCScanX,

and the distribution of K_s values revealed a potential WGD event for *A. indica*, *M. azedarach* and other angiosperm species, indicating that *A. indica* and *M. azedarach* underwent a single WGD in close proximity. The K_s distributions of other Meliaceae species were also calculated, and two notable peaks were identified (Figure 2d). This result indicates that all Meliaceae species share a single WGD event, and the old peak may be an ancient WGD event. Further studies should be conducted to confirm that this WGD event specifically occurred in Meliaceae or another upper clade.

The expansion and contraction gene families of the 11 species were estimated using CAFE. A total of 649 and 550 genes expanded in *A. indica* and *M. azedarach* genomes, respectively, while the contracted genes were 4143 and 4239 respectively. Significantly expanded families ($P \leq 0.05$) were clustered in the GO items and KEGG pathways (Figure S4). Among the expanded genes of *A. indica*, the most enriched molecular function was catalytic activity (GO:0003824), which mainly resulted from the expansion of oxidoreductase (GO:0016747/0016705), transferase (GO:0016491/0016705) and terpene synthase (GO:0010333) activities. As for the expanded genes of *M. azedarach*, the most enriched molecular function was catalytic activity, but the expanded genes were classified into oxidoreductase and serine-type endopeptidase activities (GO:0004252). These expanded genes may be involved in specific phenotypes and environmental adaptability.

Expansion of secondary wall biosynthesis-related genes may associate with wood development in Cedreloideae

Cellulose, hemicellulose, lignin and other substances accumulate continuously in the secondary wall, and the xylem's main and hard parts in woody plants are important factors in determining the yield and quality of wood (Liu et al., 2022; Xie et al., 2022). Thus, the main gene families involved in the formation of these three compounds were checked between the Cedreloideae species, *T. sinensis*, and the Melioideae species, *A. indica* and *M. azedarach*, to identify the ancestral mechanism for timber formation in the woody Cedreloideae species, which may further improve tree breeding (Figure 3a,b and Table S12).

Cellulose accounts for approximately 20% of plant primary cell walls and approximately 50% of secondary cell walls by weight

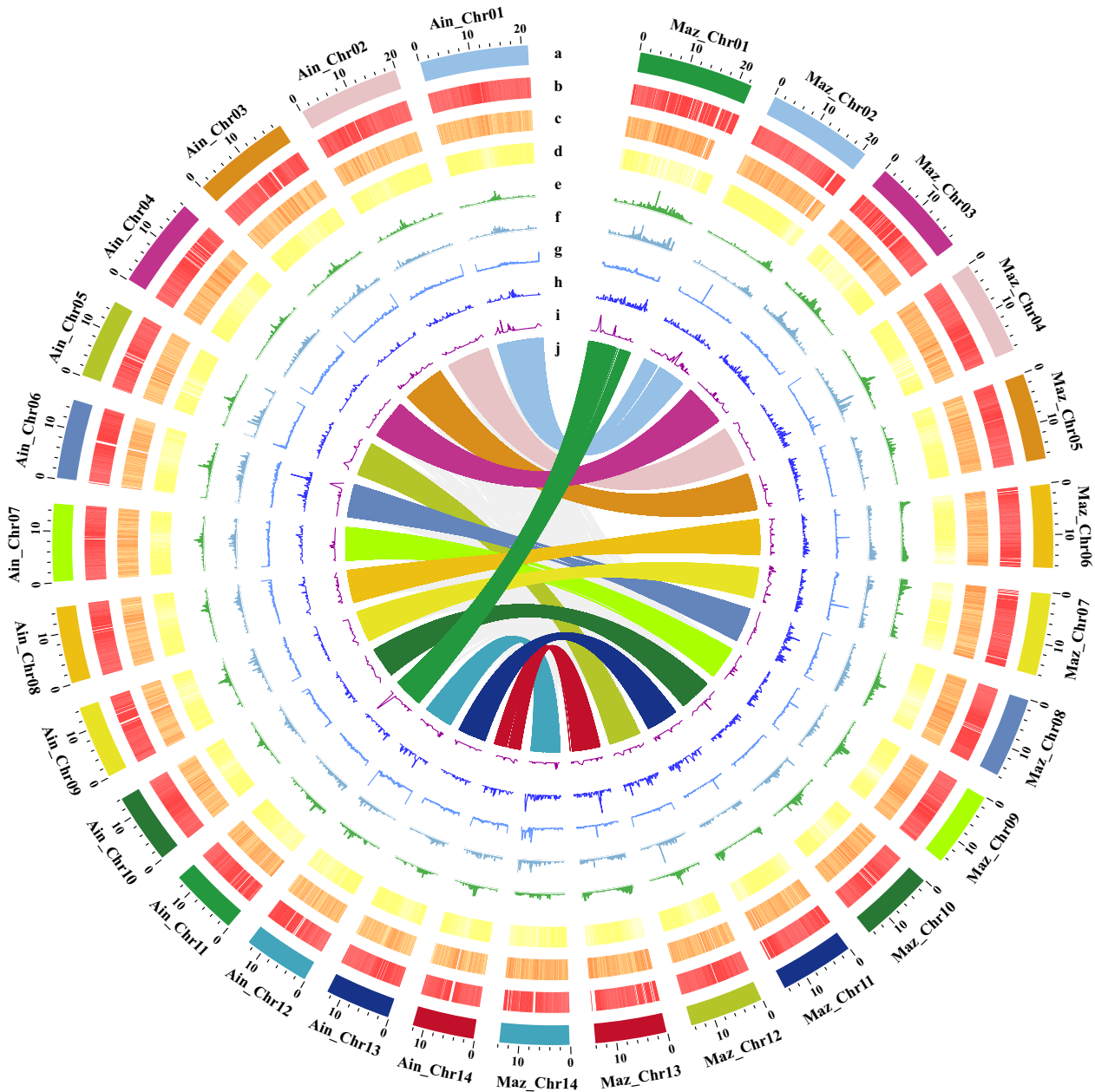


Figure 1 Genome Features of *A. indica* (Ain) and *M. azedarach* (Maz). (a) Chromosome karyotypes, (b) Gene density, (c) TRs, (d) LTR-RTs, (e) Gypsy, (f) Copia, (g) LINE, (h) SINE, (i) ncRNA, (j) Synteny between the two genomes.

(Baez *et al.*, 2022). Almost all the genes related to cellulose synthesis were increased from Melioideae to Cedreloideae; the gene numbers of cellulose synthase A (CESA), glycosyltransferase STELLO1 (STL), cellulose synthase interactive protein (CSI) and sucrose synthetases (SUS) in *T. sinensis* genome were almost double of those in *A. indica* and *M. azedarach* (Figure S5). In *Arabidopsis thaliana* cellulose biosynthesis, CESA1, CESA3 and CESA6 have been shown to act on the primary walls, while CESA4, CESA7 and CESA8 function in the secondary walls (Ender and Persson, 2011). However, only CESA4 and CESA2/5/6/9 were expanded in *T. sinensis* (Figure 3d). The synteny of STL and CSI exhibited one to two manners, while that of sucrose-phosphate synthase (SPS)/SUS was almost many-to-many manners, with the gene number expanding from four in *A. indica* to six in *T. sinensis*.

Hemicellulose helps maintain the integrity and stability of the cell wall, and can account for 25% of the secondary wall's content (Baez *et al.*, 2022; Julian and Zabolina, 2022). Hemicellulose can be divided into three classes: xylan, glucomannan and galactoglucomannan. Coniferous wood is mainly galactoglucomannan, whereas broadleaf timber and grass is mainly xylan. Glycosyltransferases (GTs) are involved in xylan synthesis, and the GT8, GT43 and GT47 gene families are involved in hemicellulose synthesis in *Arabidopsis* (Hao and Mohnen, 2014). Both glycosyltransferases underwent gene expansion, since the number of GT8, GT43 and GT47 gene families increased from 33, 4 and 39 in *A. indica* and *M. azedarach* to 65, 9 and 64 in *T. sinensis* (Figure 3d and Figure S5). Most GTs revealed one-to-many synteny between *A. indica* and *T. sinensis*, suggesting a positive

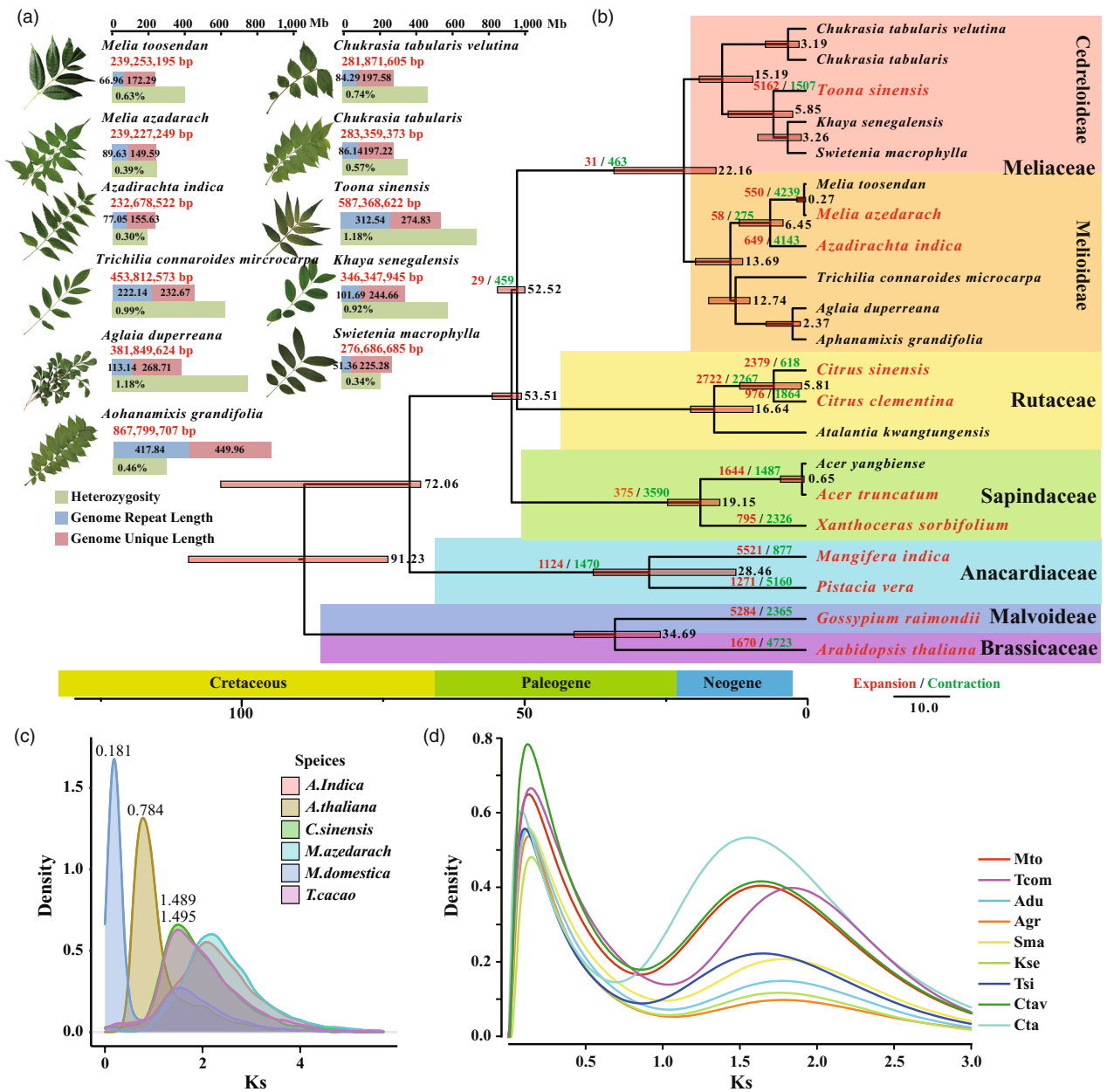


Figure 2 Phylogenetic and WGD analyses of Meliaceae. (a) Information on sequencing samples, including leaf morphology, estimated genome sizes and heterozygosity. (b) Maximum Likelihood based phylogenetic analysis of 11 Meliaceae species and 10 other plant species using chloroplast genomes with PhyML software. The divergence times (Mya) are estimated according to the fossil evidence on the TimeTree website. The numbers of expanded (red) and contracted (green) gene families shown in each branch are based on the genome data of selected species (red). (c) K_s distribution of self-syntenic genes of selected genomes. (d) K_s distribution of sequenced Meliaceae species.

role of xylan in wood formation. Among the cellulose synthase-like (CSL) genes, *CLSA* and *CSLC*, β -1,4-glucan synthase and xyloglucan 6-xylosyltransferase, are involved in the synthesis of the xyloglucan backbone. CSLD and CSLG are both thought to be Golgi-localized β -glucan synthases that polymerize the backbones of non-cellulosic polysaccharides, such as hemicelluloses, in plant cell walls (Wang *et al.*, 2001). Interestingly, the expansion pattern varied between the different subfamily species. The *CLSA*, *CLSC* and *CSLD* families showed an increase in gene numbers in *T. sinensis*, whereas only *CSLG* expanded in *A. indica* and *M. azedarach*.

Lignin is a three-dimensional polymer of phenylpropanoid alcohols (or monolignols), including p-coumaryl alcohol, coniferyl alcohol and sinapyl alcohol. These alcohols are associated with cellulose and hemicellulose and provide rigidity to plant-supporting and plant-conducting tissues (Boerjan *et al.*, 2003; Yang *et al.*, 2007; Vanholme *et al.*, 2010). The lignin biosynthesis pathways can be divided into phenylpropane synthesis (shikimic acid pathway), lignin monomer synthesis (phenylpropane metabolic pathway) and monomer polymerization (Figure 3a). After examining all enzymes related to the lignin pathway (Figure S6), there was little difference between *A. indica* and *T. sinensis*,

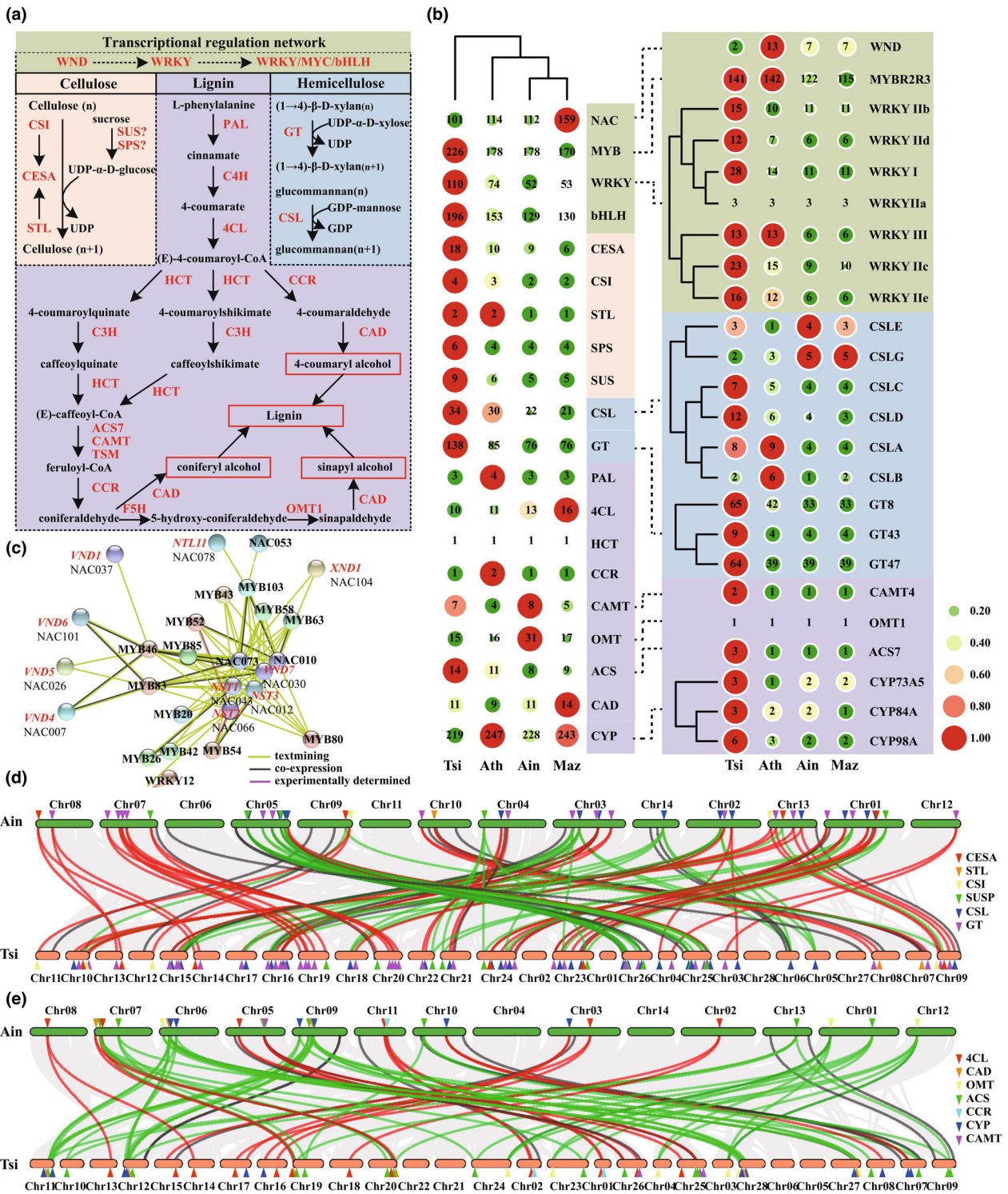


Figure 3 Expansion of secondary wall biosynthesis-related genes in Cedreloideae, *T. sinensis*. (a) Synthetic pathways of the main components of secondary wall, namely cellulose, hemicellulose and lignin. (b) Statistics of related genes in *T. sinensis* (Tsi), *A. indica* (Ain), *M. azedarach* (Maz) and *A. thaliana* (Ath) genomes. (c) Transcriptional regulation network of secondary cell walls using the *A. thaliana* factors with STRING. (d) Synteny ($e < 1e^{-05}$) of cellulose and hemicellulose biosynthesis-related genes between *T. sinensis* (Tsi) and *A. indica* (Ain) genomes. (e) Synteny ($e < 1e^{-05}$) of lignin biosynthesis-related genes between *T. sinensis* (Tsi) and *A. indica* (Ain) genomes. In d and e, the one-to-many synteny was marked with red and many-to-many synteny was marked with green.

except for the expansion of caffeoyl-CoA-O-methyltransferase 4 (CAMT4), 1-aminocyclopropane-1-carboxylate (ACC) synthase 7 (ACS7) and three CYPs: cinnamic acid 4-hydroxylase (C4H), also

known as CYP73A5), p-coumaroyl shikimate/quinate 3'-hydroxylase (C3H, CYP98A3) and coniferyl aldehyde 5-hydroxylase (F5H, CYP84A1). Although some 4-coumarate-CoA

ligases (4CLs) showed many-to-many synteny between *A. indica* and *T. sinensis*, the total gene number with stricter selection criteria showed little difference (Figure 3e). Interestingly, methyltransferase, O-methyltransferase (OMT) and CAMT all showed increased total gene number in *A. indica* genome, which may be important for plant special features.

The expression of these secondary wall biosynthetic genes is also controlled by a complex transcriptional regulation network, which is mainly divided into three levels. The first level contains a group of wood-related NAC domain transcription factors (WNDs) (Lin et al., 2017; Ohtani et al., 2011). The second level features the main switches MYB46 and MYB83, which are directly regulated by SND1 and its homologues NST1, NST2, VND6, VND7 and other MYB transcription factors (Grima-Pettenati et al., 2012; McCarthy et al., 2010). Lastly, the third layer contains other transcription factors, including MYB, WRKY and bHLH, that directly activate or inhibit the expression of key enzymes in secondary wall biosynthesis (Liu et al., 2022). We checked the transcriptional regulation network in *A. thaliana* with STRING, and the results provided similar divisions (Figure 3c and Figure S7), in which vascular-related NAC domains (VNDs) regulated MYB46 and MYB83. Additionally, many other NAC and MYB transcription factors constitute a regulatory network that interacts with other transcription factors. However, the NAC and WND transcription factors in *T. sinensis* were the fewest among the four species. Using the same procedure, we only identified two WND transcription factors in the *T. sinensis* genome and 13 members within the *A. thaliana* genome. This *T. sinensis* number was also significantly smaller than that of Melioideae species (Figure S8). Further studies are needed to explore the biological functions of these transcription factors and the incomplete annotation of *T. sinensis* genome. However, the other three groups, MYB, WRKY and bHLH of *T. sinensis* genome, were larger in number than those of the other species (Figures S9–S11). And we found that almost all WRKY subfamilies experienced expansion in *T. sinensis* genome.

Expansion of terpenoids oxidases may be responsible for limonoids biosynthesis

Structurally, limonoids are also known as tetranortriterpenoids because they are formed by the loss of four terminal carbons of the side chain in the apotrucallane or apoephane skeleton, which then cyclize to form the 17 β -furan ring (Hodgson et al., 2019; Tan and Luo, 2011). Thus, limonoid biosynthesis from a triterpene backbone can originate from the isoprenoid biosynthesis pathway, starting with squalene cyclization and altered by oxidoreductases, isomerases, methyl/acetyltransferases and hydrolases (Bhambhani et al., 2017; Wang et al., 2016, 2022). The first committed step of both mevalonate pathway (MVA) pathways and methyl-erythritol phosphate pathway (MEP)

leads to the synthesis of isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP) (Aarthy et al., 2018; Pandreka et al., 2015). Oxidosqualene cyclases (OSCs) are the primary enzymes to convert triterpenoid carbon structures into precursors of triterpenoid metabolites (Lian et al., 2020; Wang et al., 2022; Xue et al., 2012). However, only euphane and triucallane from the tetracyclic skeleton can be modified and converted to limonoids by CYPs, TPS, methyltransferases and acyltransferases.

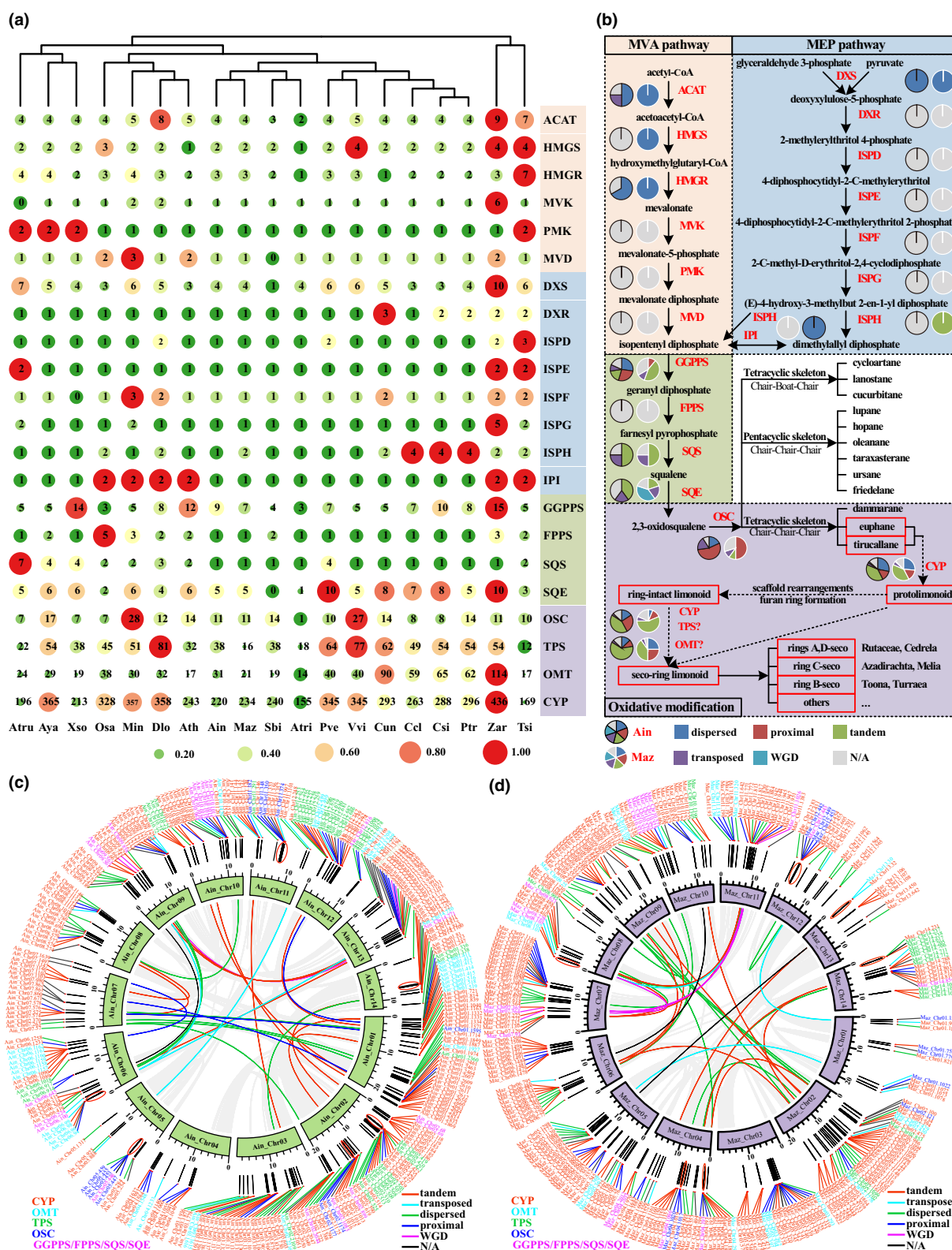
To determine the mechanisms that enrich for limonoids in Meliaceae, we analysed all genes involved in the limonoid biosynthesis pathway of the main Sapindales taxonomy (Table S11). First, we focused on the copy number variations of these gene families (Figure 4a and Table S13). There were few differences between the MVA and MEP pathway factors among these genomes, except for *T. sinensis* and *Z. armatum* for their potential duplication events. Although some species or families possessed more gene repetition, including phosphomevalonate kinase (PMK) of Sapindaceae species and 4-hydroxy-3-methylbut-2-enyl diphosphate reductase (ISPH) of Rutaceae, the majority of these genes shared similar numbers in different families, suggesting conservation of basic pathways (Figure S12). The gene numbers varied with chain elongation, cyclization and modification, especially for geranylgeranyl pyrophosphate synthase (GGPPS), squalene epoxidase (SQE), OSC, TPS, OMT and CYP gene families (Figure S13). We noticed that some of these terpenoid oxidases did not have large gene families in the Meliaceae species. Thus, limonoids in Meliaceae are not dependent on the gene dosage effect but on other mechanisms.

We then focused on the comparison within Meliaceae, especially for compounds with higher bioactivities, such as ring C-seco limonoids, which mainly originated from the Azadirachta and Melia genera. We found that gene duplication related to the limonoid biosynthesis pathway in *A. Indica* and *M. azedarach* was mainly due to tandem duplication and proximal duplication (Figure 4b). Among them, terpenoid oxidase expansion may be responsible for limonoid biosynthesis. WGD had the least contribution to gene duplication, suggesting that the two species did not undergo additional polyploid events. We then checked the locations of these genes on the chromosomes of *A. Indica* and *M. azedarach* (Figure 4c,d). The majority of these terpenoid oxidases are clustered within 10 gene gaps to facilitate their roles. For example, GGPPS clustered at Chr10 of *A. Indica* and *M. azedarach*. Four out of five GGPPS of *M. azedarach* were derived from tandem duplication, while three of five GGPPS of *A. Indica* resulted from proximal duplication. Many OSCs were derived from proximal duplication, and they clustered with SQE cyclization at Ain_Chr5 and Maz_Chr12. Moreover, a cluster consisting of two OSC and four CYP at Ain_Chr11 was not found in the corresponding Maz_Chr01. As

Figure 4 Tandem duplication of terpenoids oxidases may be responsible for limonoid biosynthesis in *A. indica* and *M. azedarach*. (a) Statistics of limonoids biosynthesis-related genes in selected genomes. (b) Limonoid synthetic pathways, which are divided into MVA and MEP pathway, cyclization stage and oxidative modification. (c) Gene positions and duplication sources of related genes in *A. indica*. (d) Gene positions and duplication sources of related genes in *M. azedarach*. Ain, *A. indica*; Ath, *Arabidopsis thaliana*; Atri, *Amborella trichopoda*; Atru, *Acer truncatum*; Aya, *Acer yangbiense*; Ccl, *Citrus clementina*; Csi, *C. sinensis*; Cun, *C. unshiu*; Dlo, *Dimocarpus longan*; Maz, *M. azedarach*; Min, *Mangifera indica*; Osa, *Oryza sativa*; Ptr, *Poncirus trifoliata*; Pve, *Pistacia vera*; Sbi, *Sclerocarya birrea*; Tsi, *T. sinensis*; Vvi, *Vitis vinifera*; Xso, *Xanthoceras sorbifolium*; Zar, *Zanthoxylum armatum*. AACT, acetyl-CoA acetyltransferase; HMGS, 3-hydroxy-3-methylglutaryl CoA synthase; HMGR, 3-hydroxy-3-methylglutaryl CoA reductase; MVK, mevalonate kinase; MVD, Mevalonate 5-diphosphate decarboxylase; DXS, 1-deoxy-D-xylulose-5-phosphate synthase; DXR, 1-deoxy-D-xylulose-5-phosphate reductoisomerase; ISPD, 2-C-methyl-D-erythritol 4-phosphate cytidyltransferase; ISPE, 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase; ISPF, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; ISPG, 4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase; IPI, isopentenyl diphosphate isomerase.

for TPS, there were two large clusters at Ain_Chr12 and Ain_Chr02, the majority of which resulted from tandem duplication. However, there was only one cluster at Maz_Chr14, corresponding to the chromosome of Ain_Chr12. The OMT of *A. indica* was significantly higher than that of *M. azedarach*,

resulting in two large OMT clusters at Ain_Chr01 and Ain_Chr06. These results were mainly attributed to tandem replication. Considering the main limonoid differences, extra tandem replications might contribute to the biosynthesis of azadirachtin and other related compounds.



We further reconstructed the phylogeny for OSC, TPS, OMT and CYP and analysed the information. A diverse array of triterpenoid skeletons are directly cyclized from 2,3-oxidosqualene by members of the OSC family, which had expanded greatly by lineage-specific duplication in plants (Lian *et al.*, 2020; Wang *et al.*, 2022). There was only one OSC gene in *Amborella trichopoda*, belonging to the cycloartenol synthase clade, which showed different expansions in other angiosperm lineages (Figure 5a). Many members appeared, and they can be roughly divided into cycloartenol synthase, lanosterol synthase and pentacyclic triterpene synthase-like enzymes. Although some Sapindales species may have expanded in related catalogues, the most important were two unknown groups consisting only of Sapindales OSCs. MaOSC1, here indicated as Maz_Ch01.1022, was shown to produce tirucalla-7,24-dien-3 β -ol, the precursor of limonoids from *A. indica*, *M. azedarach* and *C. sinensis* (Hodgson *et al.*, 2019). Thus, these two groups contribute to the skeletons of special triterpenoids from Sapindales species.

Based on its structure and catalytic mechanisms, TPS can be classified into seven subfamilies (Li *et al.*, 2021; Zhou and Pichersky, 2020). As shown in Figure 5b, almost all TPS clades appeared to expand from 18 *A. trichopoda* TPS to approximately 332 TPS in the other nine species. The TPS-*e/f* and TPS-*a* subfamilies increased dramatically in *O. sativa*, whereas all TPS subfamilies expanded in Sapindales species. Interestingly, Meliaceae and Rutaceae TPSs split into different branches of the TPS-*a* and TPS-*b* subfamilies, suggesting a division between the main terpenes. In Meliaceae, increased TPS genes mainly resulted from the tandem duplication of *A. indica*, and most TPS of *T. sinensis* were in the TPS-*c* subfamilies.

The numbers of CAMT and OMT were significantly increased in *A. indica* genome, which led us to focus on their roles in neem special features. OMTs, characterized by the Methyltransf_2 (PF00891) domain, were divided into eight clades based on protein structure and similarity to *Arabidopsis* proteins (Figure 5c). The OMT genes of *O. sativa* mainly clustered into other OMT clades, which mainly consisted of trans-resveratrol di-O-methyltransferase, probable O-methyltransferase 3 and isoflavone 4'-O-methyltransferase. Rutaceae members expanded at CAMT, anthranilate N-methyltransferase (ANMT) and other OMTs, while Meliaceae members increased at acetylserotonin O-methyltransferase (ASMT), xanthohumol 4'-O-methyltransferase (OMT2) and indole glucosinolate O-methyltransferase (IGMT). The special cluster of neem OMT belonged to the OMT2 clade, which exhibits low substrate selectivity and is involved in prenylated phenolic natural product biosynthesis, which commonly appears at the side chains or rings of terpenes (Nagel *et al.*, 2008).

Finally, the characteristics and evolution of the CYP450 gene families were investigated among the main limonoids, including *A. indica*, *M. azedarach*, *T. sinensis* and *C. sinensis* (Figure 6). More than 1100 CYP sequences from the genomes were aligned to construct the phylogenetic tree, and they were further clustered into nine clans. Clan74, Clan711, Clan710, Clan51 and Clan97 are single-family CYP clans. Clan71 is the largest CYP clan, followed by Clan85, Clan86 and Clan72. Clan71 comprised almost half of all the genes, and those of *C. sinensis* experienced significant expansion since the gene numbers of CYP82, CYP705, CYP76, CYP79, CYP71A-like, CYP71A and CYP83 were several times larger than that of the Meliaceae species. These results were in accordance with a previous conclusion that Clan71 produced many gene duplications at an accelerated rate (Zheng *et al.*, 2019). Moreover, the Rutaceae CYPs seemed to increase at

CYP704, CYP90 and Clan97, while Meliaceae CYPs showed expansion in other clans. For example, all three Meliaceae CYP74 genes were much more abundant than those in Rutaceae. CYP88 and CYP716 were expanded in *M. azedarach*, while CYP94B, CYP707, CYP77 and CYP14 were mainly increased in *T. sinensis*. These CYPs might benefit the adaptability of these two species, since they are the only Meliaceae species that are widely distributed in temperate zones and would help enrich limonoid structures. As reported previously, CYP71CD and CYP71BQ in *C. sinensis* and *M. azedarach* help oxidize tirucalla-7,24-dien-3 β -ol, resulting in spontaneous hemiacetal ring formation and producing protolimonoid melianol (Hodgson *et al.*, 2019). Thus, the most increased genes for *A. indica*, CYP71B, were suggested to be associated with the biosynthesis of the species-specific compound, azadirachtin.

Discussion

With improved sequencing technologies, more plant and animal species have had their whole genomes sequenced (Mei *et al.*, 2022; Sun *et al.*, 2022). Because of its economic value, different sequencing technologies have been used to obtain the draft genome of *A. indica* (Table 1), with contig N50 lengths ranging from 740 to 15 948 bp and genome sizes from 182.93 to 364 Mb (Krishnan *et al.*, 2012, 2016; Kuravadi *et al.*, 2015). However, low genome contiguity limits its use for downstream genomic research. In 2020, a genome assembly of the Chinese mahogany, *T. sinensis*, was reported, of which 28 chromosomes comprised 596 Mb with a contig N50 value of 1.5 Mb (Ji *et al.*, 2021). More recently, a chromosome-level genome assembly of *A. indica*, of which 14 chromosomes were estimated to be 281 Mb with a contig N50 value of 6 Mb, was reported to identify terpene biosynthesis (Du *et al.*, 2022). However, we noticed that the repeat element library of neem genomes varied among different versions. In the recently reported neem genome, 40.99% (approximately 115 Mb) of the sequences were identified as repetitive sequences, and unclassified elements accounted for 14.28% (Du *et al.*, 2022). In our study, the total number of repeat elements was 77.05 Mb, and the unclassified elements were only 1.24% of the sequences. Historically, the neem tree was introduced to China by the Academician Shanhuan Zhao in 1983, and it was first planted in the Insecticidal Herbarium Garden of South China Agricultural University, Guangzhou. It has been successfully introduced in Xuwen, Guangdong, and Wanning, Hainan (Xu *et al.*, 2017). The heterozygosity ratio of neem from Hainan was estimated to be 0.896%, and the genome size was estimated to be 165 Mb based on the 21-mer depth distribution of Illumina short reads. Our plant sample was collected from the original neem tree in China, and the heterozygosity was lower than 0.30%, which provided the basis for us to obtain a better genome assembly. In our opinion, more information is needed to reveal the evolutionary pathways and feature adaptability of neem trees when comparing samples collected from different regions and countries, including China, India, Pakistan and West Africa.

Using the genomes of *A. indica*, *M. azedarach* and other Meliaceae species, we reconstructed the inter- and intrafamily phylogenetic relationships. The three phylogenetic trees using different methods were consistent, suggesting that Meliaceae should be divided into two subfamilies: Cedreloideae and Melioideae. The divergence times showed that these two subfamilies diverged during the transition from the Palaeogene to Neogene, accompanied by the formation of modern mountain

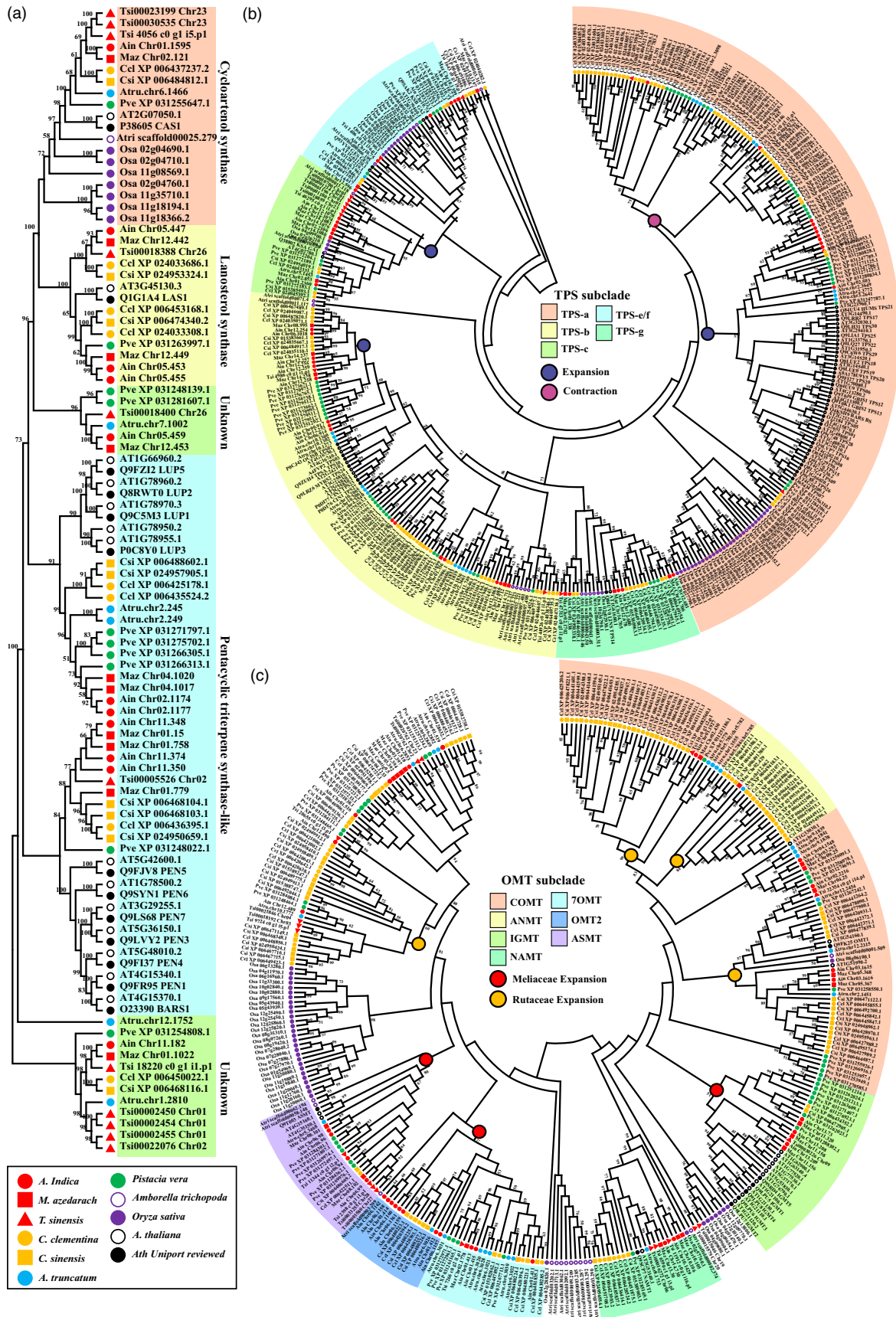


Figure 5 Phylogenetic analysis of OSC (a), TPS (b) and OMT (c) from related Sapindales species. 7OMT, (R, S)-reticuline 7-O-methyltransferase; NAMT, Nicotinate N-methyltransferase 1.

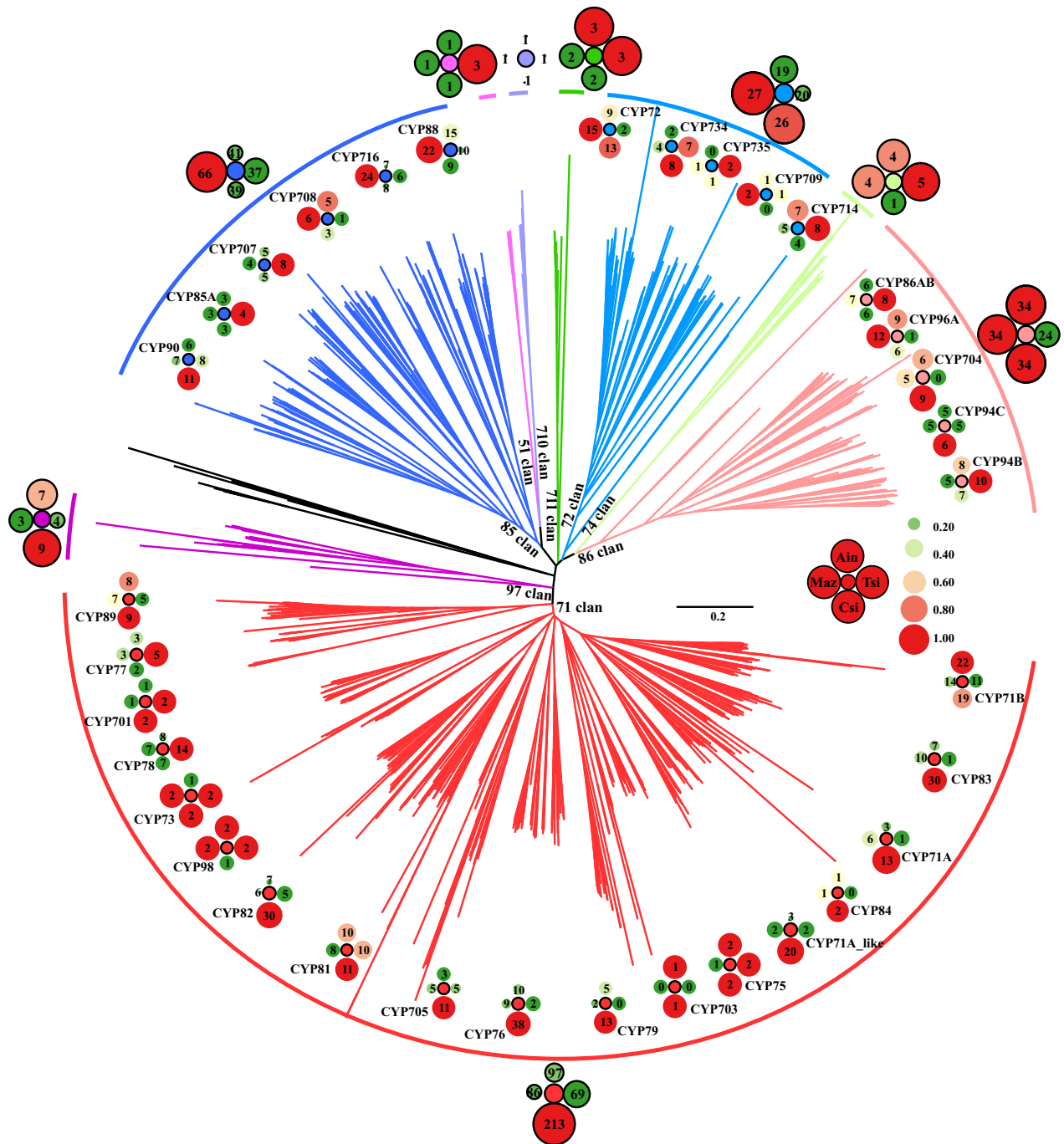


Figure 6 Phylogenetic analysis of CYP from *A. indica* (Ain), *M. azedarach* (Maz), *T. sinensis* (Tsi) and *C. sinensis* (Csi).

and climatic zonation. Moreover, many living genera and species appeared as the arrival of high prosperity of angiosperms and flowering plants. Our molecular clock produced a similar result on the time estimation of Meliaceae appearance, accompanied by adequate species and information. We also checked genes related to wood formation in all species, but there were few differences between the Cedreloideae and Melioideae subfamilies. Therefore, we only used the assembled genomes to investigate limonoid biosynthesis. However, the data and transcripts were used to compensate for the deficiency of *T. sinensis* genome assembly in later investigations.

Wood formation is controlled by transcriptional regulatory networks that consist of transcription factors and secondary cell wall genes. Many genes related to cellulose, hemicellulose and lignin biosynthesis play vital roles in woody plants (Nakano *et al.*, 2015; Taylor-Teeple *et al.*, 2015). For example, five of the 17 identified CESA genes in *Populus trichocarpa*, *CESA4*, *CESA7A*, *CESA7B*, *CESA8A* and *CESA8B5*, exhibited high expression levels in mature stems, especially in secondary vascular tissue development (Kumar *et al.*, 2009; Xu *et al.*, 2021). Moreover, *CSLA1* is highly expressed in the developing xylem of poplar and pine trees. Although monolignol transport and lignification

mechanisms remain unclear, many enzymes critical in lignin monomer synthesis have been studied. For example, 4CL1 inhibition or downregulation of cinnamoyl-CoA-reductase in poplar decreases lignin content in the secondary wall (Wang *et al.*, 2009). In this study, there were more genes involved in both cellulose and hemicellulose biosynthesis in *T. sinensis* from Cedreloideae, while the number of lignin biosynthesis pathways did not increase. This finding may contribute to the different compound contents in the secondary wall between subfamilies, and further analysis of timber constitution would help confirm this hypothesis. Transcription factors also play a role in the secondary cell wall. Inhibiting PtSND2 thinned xylem fibres and reduced lignin and cellulose contents in poplar, reducing secondary cell wall thickness (Wang *et al.*, 2013). NAC1 regulates phenylalanine biosynthesis by activating MYB4 in pines (Pascual *et al.*, 2018). In poplar, PtrMYB3 and PtrMYB20 are highly expressed in ducts and fibres to regulate lignin, cellulose and xylose biosynthesis (McCarthy *et al.*, 2010). Recent research has revealed that PtrMYB074 and PtrWRKY19 transcriptionally activate *PtrbHLH186* for secondary xylem development in *P. trichocarpa* (Liu *et al.*, 2022). However, we found only two related SNDs in the *T. sinensis* genome, which might result from imperfect genome assembly or species-specific features, but we found significant expansion of the downstream transcription factors MYB, WRKY and bHLH. These collective results will contribute to future investigation of secondary wall biosynthesis and wood development. However, more focus should be placed on identifying the expression patterns of critical transcription factors between additional Cedreloideae and Melloioideae species and how they influence biosynthetic pathways.

Terpenes of Sapindales or limonoids from Meliaceae and Rutaceae have drawn attention because of their anticancer, antimicrobial, antioxidant and insecticidal properties (Fernandes Da Silva *et al.*, 2021; Saleem *et al.*, 2018; Zhang and Xu, 2017). Almost 50 limonoid aglycones have been reported in Rutaceae primarily with seco-A,D-ring structures (Bao *et al.*, 2016; Zhang and Xu, 2017). Meliaceae limonoids are known to contain approximately 1500 structurally diverse compounds, of which seco-C-ring limonoids are the most interesting because of their insecticidal activity (Bao *et al.*, 2016; Fernandes Da Silva *et al.*, 2021; Saleem *et al.*, 2018; Tan and Luo, 2011). Many researchers have attempted to identify the key enzymes involved in limonoid biosynthesis, which would improve and increase the application of these compounds. For example, transcriptome analysis of neem fruits, flowers and leaves revealed that GGPPS, farnesyl diphosphate synthase and squalene synthase initiate isoprenoid biosynthesis (Pandreka *et al.*, 2015). Another analysis based on transcriptome datasets from neem leaf and fruit suggested that the CYP450 members CYP16671, CYP16365 and CYP18835 catalyse secondary modifications of bioactive triterpenoids (Bhambhani *et al.*, 2017). Moreover, transcriptomic analysis of *C. grandis* revealed that CYP450s and the transcription factor MYB exhibited high correlation coefficients with limonoid biosynthesis (Wang *et al.*, 2017). Comparative analysis of the terpenoid biosynthesis pathways in *A. indica* and *M. azedarach* also revealed that only six of these genes were upregulated in *A. indica* (Wang *et al.*, 2016). The conserved enzymes, OSC1, CYP71CD and CYP71BQ, were identified in *M. azedarach* and *C. sinensis* are responsible for protolimonoid melianol biosynthesis, contributing to limonoid metabolic engineering and diversification (Hodgson *et al.*, 2019). However, the lack of a high-quality reference genome has been identified as an urgent problem

despite whole genomes of *T. sinensis* and *A. indica* being reported. Regarding terpene synthesis, these two reports concluded: (1) tandem duplication or a recent WGD event may be responsible for most TPS gene expansion in *T. sinensis* (Ji *et al.*, 2021); (2) most *A. indica*-specific TPS and CYP genes were located on chromosome 13 (Du *et al.*, 2022). In our study, all the related genes were compared using the terpene biosynthesis differences between Meliaceae and Rutaceae, between Cedreloideae and Melloioideae and between *A. indica* and *M. azedarach*. Although there were few gene differences in the triterpene backbone of the isoprenoid biosynthesis pathway, the gene numbers varied for isoprenoid chain elongation, cyclization and modification. For example, two expanded, unknown, and Sapindales-specific OSC catalogues may contribute to the unique skeletons of Sapindales compounds, and some expanded TPS subfamilies may enrich for terpene species. Moreover, the majority of these genes were located as clusters to facilitate their roles; however, there were differences in cluster locations between *A. indica* and *M. azedarach*, which may produce different limonoid contents. For example, a cluster consisting of two OSC and four CYP at Ain_Chr11 was not found in the corresponding Maz_Chr01. There were two large TPS clusters at Ain_Chr12 and Ain_Chr02, while there was only one large cluster at Maz_Chr14. Moreover, the OMT of *A. indica* was significantly higher than that of *M. azedarach*, resulting in two large OMT clusters at Ain_Chr01 and Ain_Chr06. The CYP71 clan of Rutaceae *C. sinensis* showed significant expansion, whereas all three Meliaceae CYP74 genes were much more abundant than in Rutaceae. Thus, expansions of different terpenoid oxidases are responsible for different limonoid biosynthesis.

In conclusion, we report improved genomes for *A. indica* and *M. azedarach*. Chloroplast genomes, single-copy gene families and SNP of 11 Meliaceae species revealed their phylogenetic relationships, which should be classified into the Cedreloideae and Melloioideae subfamilies. Moreover, segmental duplication, rather than WGD, in Meliaceae species improved their characteristics. Many downstream transcription factors and cellulose/hemicellulose biosynthesis-related genes in *T. sinensis* expanded significantly compared to those in *A. indica* and *M. azedarach*, and these genes may be involved in wood formation. Moreover, expanded special OSCs catalogues may contribute to the skeleton diversification of Sapindales compounds. Additionally, further research on the clustered genes focusing on terpene chain elongation, cyclization and modification can contribute to improved limonoid biosynthesis. The different expanded clans of TPS, OMT and CYPs are responsible for the different classes of limonoids in these species. These results would benefit further investigations into wood development and limonoid biosynthesis.

Materials and methods

Plant sampling and DNA extraction

Leaves of *Azadirachta indica*, *Aglaiia duperreana*, *Chukrasia tabularis*, *C. tabularis velutina*, *Khaya senegalensis*, *Swietenia macrophylla* and *Trichilia connaroides* were collected from the Insecticidal Herbarium Garden of the South China Agricultural University (Guangzhou, China). The leaves of *Aphanamixis grandifolia* and *Toona sinensis* were collected from the Arboretum of South China Agricultural University (Guangzhou, China). Leaves of *Melia azedarach* were collected from the Kunming Institute of Botany, Chinese Academy of Sciences (Kunming, China) with institutional permission. Leaves of *M.*

toosendan were collected from Southwest University (Chongqing, China).

The following procedures of DNA extraction, library preparation and sequencing, de novo assembly, de novo annotation, Hi-C library construction and sequencing were performed by Nextomics Biosciences Co., Ltd. (Wuhan, China). Genomic DNA was prepared using the SDS method, followed by purification with the QIAGEN Genomic kit (Cat#13343, QIAGEN, USA) according to the provided instructions. DNA degradation and contamination were monitored using 1% agarose gel electrophoresis. DNA purity was detected using a NanoDrop One UV–Vis spectrophotometer (Thermo Fisher Scientific, USA), and DNA concentration was measured using a Qubit 3.0 Fluorometer (Invitrogen, USA).

Library preparation and sequencing

A total of 2 µg of DNA was used for ONT library preparation. After the sample was qualified, size selection of long DNA fragments was performed using the BluePippin system (Sage Science, USA). Next, the ends of the DNA fragments were repaired, and the A-ligation reaction was conducted using the NEBNext Ultra II End Repair/dA-tailing Kit (Cat#E7546, NEB, USA). The adapter in the SQK-LSK109 kit (Oxford Nanopore Technologies, UK) was used for further ligation reactions, and a Qubit 3.0 Fluorometer was used to quantify library fragment size. Sequencing was performed on a Nanopore PromethION sequencer (Oxford Nanopore Technologies, UK) instrument by NextOmics (Wuhan China). Output base-calling was first performed to convert the FAST5 files to FASTQ format using Guppy (Version 3.2.2 + 9fe0a78). Raw reads with mean_qscore_template <7 were then filtered, resulting in pass reads. A further filter with fastp (version 0.19.4) obtained clean data for genome assembly.

De novo assembly and assessment

The assembly was constructed using the overlap layout-consensus/string graph method with NextDenovo (v2.0-beta.1). First, the original clean reads were self-corrected using NextCorrect to obtain consistent sequences. Then, correlations were performed using NextGraph, and the preliminary genome was assembled (Senol et al., 2019). To improve the accuracy of the assembly, the contigs were refined with Racon (v1.3.1) using ONT long reads, and NextPolish (v1.0.5) using Illumina short reads with default parameters (Hu et al., 2020). To discard possibly redundant contigs and generate a final assembly, similarity searches were performed with the parameters: identity = 0.8, overlap = 0.8.

The completeness of genome assembly was assessed using BUSCO (v4.0.5) and CEGMA (v2). To evaluate the accuracy of the assembly, all Illumina paired-end reads were mapped to the assembled genome using the BWA (Burrows-Wheeler Aligner, Version 0.7.12-r1039), and the mapping rate and genome coverage of sequencing reads were assessed using Samtools (v1.4). Base accuracy of the assembly was calculated using BCFtools (v1.8.0). The coverage of the expressed genes was examined using HISAT2 (v2.1.0) with default parameters. The draft genome assembly was submitted to the Nucleotide Sequence Database library (NT), and aligned sequences were eliminated to avoid including the mitochondrial sequences.

Hi-C library construction and sequencing

Genomic DNA was extracted, and sequencing data were obtained using the Illumina NovaSeq platform. In brief, freshly harvested leaves were cut into 2 cm pieces and vacuum-filtrated

in nuclei isolation buffer supplemented with 2% formaldehyde. Crosslinking was halted with glycine, and fixed tissue was then ground to powder before resuspending in nuclei isolation buffer. Purified nuclei were digested with 100 units of DpnII and marked by incubation with biotin-14-dCTP. The ligated DNA was sheared into 300–600 bp fragments, followed by blunt-end repair, A-tailing and purification through biotin-streptavidin-mediated pull-down. Finally, the Hi-C libraries were quantified and sequenced using the Illumina NovaSeq platform (PE150, Illumina, USA).

Paired-end reads were generated, and quality control of the Hi-C raw data was performed using HiC-Pro (Servant et al., 2015). First, low-quality sequences (quality scores <20), adaptor sequences and sequences shorter than 30 bp were filtered out using fastp (Version 0.12.6), and then the clean paired-end reads were mapped to the draft assembled sequence using Bowtie2 (v2.3.2) to obtain unique mapped paired-end reads. Valid interaction paired reads were identified and retained using HiC-Pro (v2.8.1) from unique mapped paired-end reads for further analysis (Burton et al., 2013). The scaffolds were further clustered, ordered and oriented to chromosomes by LACHESIS (<https://github.com/shendurelab/LACHESIS>), with parameters CLUSTER_MIN_RE_SITES = 100, CLUSTER_MAX_LINK_DENSITY = 2.5, CLUSTER NONINFORMATIVE RATIO = 1.4, ORDER MIN N RES IN TRUNK = 60, and ORDER MIN N RES IN SHREDS = 60. Finally, placement and orientation errors exhibiting discrete chromatin interaction patterns were manually adjusted.

Repeat and ncRNA annotation

Tandem repeats were annotated using GMATA (v2.2) and Tandem Repeats Finder (TRF, Version 4.07b). GMATA identifies simple repeat sequences, and TRF recognizes all tandem repeat elements. TEs were identified using a combination of *ab initio*- and homology-based methods. Briefly, an *ab initio* repeat library was first predicted using MITE-Hunter (Han and Wessler, 2010) and RepeatModeler (version open-1.0.11) with default parameters. This library was aligned to Repbase (<http://www.girinst.org/repbase>) with TEclass (Abrusán et al., 2009). RepeatMasker (version 1.331) was used to search for known and novel TEs. Overlapping TEs belonging to the same repeat class were collated and combined.

To obtain ncRNA, we searched the database using a prediction model. Transfer RNAs (tRNAs) were predicted using tRNAscan-SE (v2.0) and eukaryotic parameters. MicroRNA, rRNA, small nuclear RNA, and small nucleolar RNA were detected using Infernal (v1.1.2) cmscan to search the Rfam database (<http://rfam.xfam.org/>). rRNAs and their subunits were predicted using RNAmmer (v1.2).

Gene prediction and functional annotation

Three independent approaches—*ab initio* prediction, homology search and reference-guided transcriptome assembly—were used for gene prediction. GeMoMa (v1.6.1) was used to align the homologous peptides from related species to the assembly and obtain the gene structure information. For RNA-seq-based gene prediction, the filtered mRNA-seq reads were aligned to the reference genome using STAR (2.7.3a, default). The transcripts were then assembled using StringTie (v1.3.4d) and open reading frames (ORFs) were predicted using PASA (v2.3.3). For de novo prediction, AUGUSTUS (v3.3.1) with default parameters was utilized for *ab initio* gene prediction using the RNA-seq assembled training set. EVidenceModeler (v1.1.1) was used to produce an integrated gene set in which genes with TEs were removed using

the TransposonPSI package (<http://transposonpsi.sourceforge.net/>). Untranslated regions (UTRs) and alternative splicing regions were determined using PASA based on RNA-seq assemblies. The longest transcripts for each locus were retained, and regions outside the ORFs were designated as UTRs.

Gene function information, motifs and domains were assigned by comparisons with public databases, including SwissProt (<http://www.gpmaw.com/html/swiss-prot.html>), NR, KEGG (<http://www.genome.jp/kegg/>), KOG/COG (<http://www.ncbi.nlm.nih.gov/COG/>) and GO (<http://www.geneontology.org/>). Putative domains and gene GO terms were identified using the InterProScan program (version 5.32–71.0) with default parameters. For the other four databases, BLASTp (v2.7.1) was used to compare the protein sequences against the four well-known public protein databases with an E-value cut-off of 1×10^{-5} , and the results with the lowest E-values were retained.

Gene cluster and gene family analyses

Gene clustering was conducted using OrthoFinder (Emms and Kelly, 2019), and the input gene sets were collected from sequenced plant species. The extracted protein sequences were aligned pairwise to identify conserved orthologs using BLASTp set to an E-value threshold of $\leq 1 \times 10^{-5}$. Orthologous intergenomic gene pairs, paralogous intra-genome gene pairs and single-copy gene pairs were further identified from the OrthoFinder results. Genes with no homologues in other plant genomes were identified as species-specific. The GO and KEGG enrichment analyses were conducted using the AllEnricher software (Zhang *et al.*, 2020), and a *P*-value < 0.05 was used as the significance threshold.

The coding sequences were extracted from single-copy families, and each ortholog group was aligned multiple times using MAFFT (v7.313). Poorly aligned sequences were eliminated using Gblocks (version 0.91b), and the GTRGAMMA substitution model of RAXML was used for phylogenetic tree construction with 1000 bootstrap replicates. The generated tree files were displayed using MEGA (version 10.1.8). Based on the phylogenetic tree, RelTime of MCMCTree (Puttick, 2019) was used to compute the mean substitution rates along each branch and estimate the species divergence time. Three fossil calibration times were obtained from the TimeTree database (<http://www.timetree.org/>), including the divergence times of Toona and Citrus (50.1–69.2 Mya), the divergence time of Citrus and Acer (68.0–82.8 Mya) and the divergence time of monocot and eudicot (lower boundary of 130.0 Mya).

Gene family expansion and genes under positive selection

According to the results of OrthoMCL (<http://orthomcl.org/orthomcl/>), expansions and contractions of orthologous gene families were detected using CAFE (v4.2.1), which uses a birth and death process to model gene gain and loss over phylogeny.

The average nonsynonymous (K_a)/synonymous (K_s) substitution rate values were calculated, and the branch-site likelihood ratio test was conducted using Codeml implemented in the PAML package (version 4.8) to identify positively selected genes. Genes with a value < 0.05 under the branch-site model were considered positively selected genes.

Screening for whole-genome duplication events

4DTv and K_s were used to detect WGD events. First, protein sequences were extracted, and all-vs-all paralog analysis was

performed using the best hits from primary protein sequences by self-BLASTp in these plants. After filtering by identity and coverage, the BLASTp results were subjected to MCScanX (Wang *et al.*, 2012), and the collinear blocks were identified. Finally, 4DTv was calculated for the syntenic block gene pairs using KaKs_Calculator (Version 2.0), and potential WGD events in each genome were evaluated based on their K_s and 4DTv distribution. WGD (Zwaenepoel *et al.*, 2019) was used to conduct the K_s estimation under the default parameters.

Duplication source detection

To investigate the resources of the duplication genes during gene family evolution, we identified genome-wide duplications for *A. indica* and *M. azedarach* using *T. sinensis* as an outgroup. We identified different modes of gene duplication, namely WGD, tandem duplicates (TD), proximal duplicates (less than 10 gene distance on the same chromosome: PD), transposed duplicates (transposed gene duplications: TRD) or dispersed duplicates (other duplicates than WGD, TD, PD and TRD: DSD) using DupGen_finder (Qiao *et al.*, 2019) with default parameters.

Gene family analysis

The genomes of all species were obtained from the NCBI database (Supplementary Table S1). A simple HMM search and BLASTn (2.2.30+) were used for sequence searching within the TBtools software (Chen *et al.*, 2020). The original genes and selected Pfam IDs (Table S12 and Table S13) of the relevant genes were selected from the reviewed *A. thaliana* sequences from the UniProt website (<https://www.uniprot.org/>). The representative genes were collected using TBtools, and redundant sequences were screened manually. All the final sequences were aligned with the MUSCLE algorithm of MEGA 11 (MEGA, USA) and then screened by length ($> 30\%$) and conservative amino acid sites of the main Pfam sequence HMM logo. Adobe Illustrator CS6 (Adobe, USA) was used to edit and produce the figures.

Acknowledgements

The authors sincerely appreciated all the material supplied institutions and individuals, including Kunming Institute of Botany, Chinese Academy of Sciences (Kunming, China) and Prof. He Lin from Southwest University (Chongqing, China). This research was funded by the National Natural Science Foundation of China (No. 32000342), the Chinese Postdoctoral Science Foundation (2020 M672650) and the National Key Research and Development Program of China (Grant No. 2018YFD0200300). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Conflict of interest

The authors declare no conflicts of interest.

Authors' contributions

G.Z. and L.G. carried out conceptualization, supervision and project administration. G.C., Y.L. and X.Y. were involved in formal analysis. J.W. and Q.Z. carried out software. G.C. also contributed to data curation and writing—original draft. C.L. and P.L. were involved in validation. G.C. and Y.L. carried out investigation. G.C. and X.Y. carried out resources. Q.Z. and J.W. were involved in writing—reviewing and editing. G.C. and G.Z. carried out

funding acquisition. All authors discussed the results and commented on the manuscript.

Data availability statement

The data supporting the findings of this work are available within the paper and its Supplementary Information files. The datasets generated and analysed during this study are available from the corresponding author upon request. The genome assemblies of *A. indica* (CNS0590388) and *M. azedarach* (CNS0590389) are available at CNA0051891 and CNA0051892 of project CNP0003339 on China National GeneBank DataBase (CNGBdb).

References

- Aarthy, T., Mulani, F.A., Pandreka, A., Kumar, A., Nandikol, S.S., Haldar, S. and Thulasiram, H.V. (2018) Tracing the biosynthetic origin of limonoids and their functional groups through stable isotope labeling and inhibition in neem tree (*Azadirachta indica*) cell suspension. *BMC Plant Biol.* **18**, 230.
- Abrusán, G., Grundmann, N., DeMester, L. and Makalowski, W. (2009) TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329–1330.
- Baez, L.A., Ticha, T. and Hamann, T. (2022) Cell wall integrity regulation across plant species. *Plant Mol. Biol.* **109**, 483–504.
- Bao, H., Zhang, Q., Ye, Y. and Lin, L. (2016) Naturally occurring furanoditerpenoids: distribution, chemistry and their pharmacological activities. *Phytochem. Rev.* **16**, 235–270.
- Bhambhani, S., Lakhwani, D., Gupta, P., Pandey, A., Dhar, Y.V., Kumar, B.S., Asif, M.H. et al. (2017) Transcriptome and metabolite analyses in *Azadirachta indica*: identification of genes involved in biosynthesis of bioactive triterpenoids. *Sci. Rep.* **7**, 5043.
- 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.
- Chen, C., Chen, H., Zhang, Y., Thomas, H.R., Frank, M.H., He, Y. and Xia, R. (2020) TBtools: An integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* **13**, 1194–1202.
- Du, Y., Song, W., Yin, Z., Wu, S., Liu, J., Wang, N., Jin, H. et al. (2022) Genomic analysis based on chromosome-level genome assembly reveals an expansion of terpene biosynthesis of *Azadirachta indica*. *Front. Plant Sci.* **13**, 853861.
- Emms, D.M. and Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238.
- Endler, A. and Persson, S. (2011) Cellulose synthases and synthesis in *Arabidopsis*. *Mol. Plant* **4**, 199–211.
- Fernandes Da Silva, M.F.D.G., Da Silva, P.L., Amaral, J.C., Fernandes Da Silva, D., Rossi Forim, M. and Fernandes, J.B. (2021) Nortriterpenes, chromones, anthraquinones, and their chemosystematics significance in Meliaceae, Rutaceae, and Simaroubaceae (Sapindales). *Rev. Bras. Bot.* **45**, 15–40.
- Grima-Pettenati, J., Soler, M., Camargo, E.L.O. and Wang, H. (2012) Transcriptional regulation of the lignin biosynthetic pathway revisited: new players and insights. *Adv. Bot. Res.* **61**, 173–218.
- Han, Y. and Wessler, S.R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199.
- Hao, Z. and Mohnen, D. (2014) A review of xylan and lignin biosynthesis: foundation for studying *Arabidopsis irregular xylem* mutants with pleiotropic phenotypes. *Crit. Rev. Biochem. Mol. Biol.* **49**, 212–241.
- He, Z., Feng, X., Chen, Q., Li, L., Li, S., Han, K., Guo, Z. et al. (2022) Evolution of coastal forests based on a full set of mangrove genomes. *Nat. Ecol. Evol.* **6**, 738–749.
- Hodgson, H., De La Pena, R., Stephenson, M.J., Thimmappa, R., Vincent, J.L., Sattely, E.S. and Osbourn, A. (2019) Identification of key enzymes responsible for protolimonoid biosynthesis in plants: Opening the door to azadirachtin production. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 17096–17104.
- 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.
- Ji, Y.T., Xiu, Z., Chen, C.H., Wang, Y., Yang, J.X., Sui, J.J., Jiang, S.J. et al. (2021) Long read sequencing of *Toona sinensis* (A. Juss) Roem: A chromosome-level reference genome for the family Meliaceae. *Mol. Ecol. Resour.* **21**, 1243–1255.
- Julian, J.D. and Zobotina, O.A. (2022) Xyloglucan biosynthesis: from genes to proteins and their functions. *Front. Plant Sci.* **13**, 920494.
- Krishnan, N.M., Jain, P., Gupta, S., Hariharan, A.K. and Panda, B. (2016) An improved genome assembly of *Azadirachta indica* A. Juss. *G3 (Bethesda)* **6**, 1835–1840.
- Krishnan, N.M., Pattnaik, S., Jain, P., Gaur, P., Choudhary, R., Vaidyanathan, S., Deepak, S. et al. (2012) A draft of the genome and four transcriptomes of a medicinal and pesticidal angiosperm *Azadirachta indica*. *BMC Genomics* **13**, 464.
- Kubitzki, K. (ed) (2011) *Flowering Plants Eudicots-Sapindales, Cucurbitales, Myrtaceae*. Berlin: Springer.
- Kumar, M., Thammanagowda, S., Bulone, V., Chiang, V., Han, K.H., Joshi, C.P., Mansfield, S.D. et al. (2009) An update on the nomenclature for the cellulose synthase genes in *Populus*. *Trends Plant Sci.* **14**, 248–254.
- Kuravadi, N.A., Yenagi, V., Rangiah, K., Mahesh, H.B., Rajamani, A., Shirke, M.D., Russiachand, H. et al. (2015) Comprehensive analyses of genomes, transcriptomes and metabolites of neem tree. *PeerJ* **3**, e1066.
- Li, R.S., Zhu, J.H., Guo, D., Li, H.L., Wang, Y., Ding, X.P., Mei, W.L. et al. (2021) Genome-wide identification and expression analysis of terpene synthase gene family in *Aquilaria sinensis*. *Plant Physiol. Biochem.* **164**, 185–194.
- Lian, X., Zhang, X., Wang, F., Wang, X., Xue, Z. and Qi, X. (2020) Characterization of a 2,3-oxidosqualene cyclase in the toosendanin biosynthetic pathway of *Melia toosendan*. *Physiol. Plant.* **170**, 528–536.
- Lin, Y.J., Chen, H., Li, Q., Li, W., Wang, J.P., Shi, R., Tunlaya-Anukit, S. et al. (2017) Reciprocal cross-regulation of VND and SND multigene TF families for wood formation in *Populus trichocarpa*. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E9722–E9729.
- Liu, H., Gao, J., Sun, J., Li, S., Zhang, B., Wang, Z., Zhou, C. et al. (2022) Dimerization of PtrMYB074 and PtrWRKY19 mediates transcriptional activation of *PtrbHLH186* for secondary xylem development in *Populus trichocarpa*. *New Phytol.* **234**, 918–933.
- Mabberley, D.J., Pannell, C.M. and Sing, A.M. (1995) Meliaceae. In *Series I, Spermatophyta: Flowering Plants Vol 12, Part 1 (Malesiana, F., ed)*, pp. 1–407. Leiden: Leiden University.
- McCarthy, R.L., Zhong, R., Fowler, S., Lyskowski, D., Piyasena, H., Carleton, K., Spicer, C. et al. (2010) The Poplar MYB transcription factors, PtrMYB3 and PtrMYB20, are involved in the regulation of secondary wall biosynthesis. *Plant Cell Physiol.* **51**, 1084–1090.
- Mei, Y., Jing, D., Tang, S., Chen, X., Chen, H., Duanmu, H., Cong, Y. et al. (2022) InsectBase 2.0: a comprehensive gene resource for insects. *Nucleic Acids Res.* **50**, D1040–D1045.
- Muellner, A.N., Samuel, R., Chase, M.W., Coleman, A. and Stuessy, T.F. (2008) An evaluation of tribes and generic relationships in Melioideae (Meliaceae) based on nuclear ITS ribosomal DNA. *Taxon* **57**, 98–108.
- Muellner, A.N., Schaefer, H. and Lahaye, R. (2011) Evaluation of candidate DNA barcoding loci for economically important timber species of the mahogany family (Meliaceae). *Mol. Ecol. Resour.* **11**, 450–460.
- Muellner-Riehl, A.N., Weeks, A., Clayton, J.W., Buerki, S., Nauheimer, L., Chiang, Y., Cody, S. et al. (2016) Molecular phylogenetics and molecular clock dating of Sapindales based on plastid *rbcl*, *atpB* and *trnL-trnF* DNA sequences. *Taxon* **65**, 1019–1036.
- Nagel, J., Culley, L.K., Lu, Y., Liu, E., Matthews, P.D., Stevens, J.F. and Page, J.E. (2008) EST analysis of hop glandular trichomes identifies an O-methyltransferase that catalyzes the biosynthesis of xanthohumol. *Plant Cell* **20**, 186–200.
- Nakano, Y., Yamaguchi, M., Endo, H., Rejab, N.A. and Ohtani, M. (2015) NAC-MYB-based transcriptional regulation of secondary cell wall biosynthesis in land plants. *Front. Plant Sci.* **6**, 288.
- Ohtani, M., Nishikubo, N., Xu, B., Yamaguchi, M., Mitsuda, N., Goué, N., Shi, F. et al. (2011) A NAC domain protein family contributing to the regulation of wood formation in poplar. *Plant J.* **67**, 499–512.

- Oyediji Amusa, M.O., Van Wyk, B. and Oskolski, A. (2020) Wood anatomy of South African Meliaceae: evolutionary and ecological implications. *Bot. J. Linn. Soc.* **193**, 165–179.
- Pandreka, A., Dandekar, D.S., Haldar, S., Uttara, V., Vijayshree, S.G., Mulani, F.A., Aarthy, T. et al. (2015) Triterpenoid profiling and functional characterization of the initial genes involved in isoprenoid biosynthesis in neem (*Azadirachta indica*). *BMC Plant Biol.* **15**, 214.
- Pascual, M.B., Llebres, M.T., Craven-Bartle, B., Canas, R.A., Canovas, F.M. and Avila, C. (2018) *PpNAC1*, a main regulator of phenylalanine biosynthesis and utilization in maritime pine. *Plant Biotechnol. J.* **16**, 1094–1104.
- Pennington, T. and Styles, B. (1975) A generic monograph of the Meliaceae. *Blumea* **22**, 419–540.
- Puttick, M.N. (2019) MCMCtreeR: functions to prepare MCMCtree analyses and visualize posterior ages on trees. *Bioinformatics* **35**, 5321–5322.
- Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S. et al. (2019) Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biol.* **20**, 38.
- Riesco Muñoz, G., Imaña Encinas, J. and Elias De Paula, J. (2019) Wood density as an auxiliary classification criterion for botanical identification of 241 tree species in the order Sapindales. *Eur. J. For. Res.* **138**, 583–594.
- Saleem, S., Muhammad, G., Hussain, M.A. and Bukhari, S. (2018) A comprehensive review of phytochemical profile, bioactives for pharmaceuticals, and pharmacological attributes of *Azadirachta indica*. *Phytother. Res.* **32**, 1241–1272.
- Senol, C.D., Kim, J.S., Ghose, S., Alkan, C. and Mutlu, O. (2019) Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. *Brief. Bioinform.* **20**, 1542–1559.
- Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E. et al. (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259.
- Sun, Y., Shang, L., Zhu, Q.H., Fan, L. and Guo, L. (2022) Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci.* **27**, 391–401.
- Tan, Q.G. and Luo, X.D. (2011) Meliaceous limonoids: chemistry and biological activities. *Chem. Rev.* **111**, 7437–7522.
- Taylor-Teeple, M., Lin, L., de Lucas, M., Turco, G., Toal, T.W., Gaudinier, A., Young, N.F. et al. (2015) An *Arabidopsis* gene regulatory network for secondary cell wall synthesis. *Nature* **517**, 571–575.
- Tundis, R., Loizzo, M.R. and Menichini, F. (2014) An overview on chemical aspects and potential health benefits of limonoids and their derivatives. *Crit. Rev. Food Sci. Nutr.* **54**, 225–250.
- Vanholme, R., Demedts, B., Morreel, K., Ralph, J., Boerjan, W. and Great, L.B.R.C. (2010) Lignin biosynthesis and structure. *Plant Physiol.* **153**, 895–905.
- Wang, D.B.F.U., Bai, H.B.F.U., Chen, W.B.F.U., Lu, H.B.F.U. and Jiang, X.B.F.U. (2009) Identifying a cinnamoyl coenzyme A reductase (CCR) activity with 4-coumaric acid: coenzyme A ligase (4CL) reaction products in *Populus tomentosa*. *J. Plant Biol.* **52**, 482–491.
- Wang, F., Wang, M., Liu, X., Xu, Y., Zhu, S., Shen, W. and Zhao, X. (2017) Identification of putative genes involved in limonoids biosynthesis in Citrus by comparative transcriptomic Analysis. *Front. Plant Sci.* **8**, 782.
- Wang, H.H., Tang, R.J., Liu, H., Chen, H.Y., Liu, J.Y., Jiang, X.N. and Zhang, H.X. (2013) Chimeric repressor of PtSND2 severely affects wood formation in transgenic *Populus*. *Tree Physiol.* **33**, 878–886.
- Wang, J., Guo, Y., Yin, X., Wang, X., Qi, X. and Xue, Z. (2022) Diverse triterpene skeletons are derived from the expansion and divergent evolution of 2,3-oxidosqualene cyclases in plants. *Crit. Rev. Biochem. Mol. Biol.* **57**, 113–132.
- Wang, X., Cnops, G., Vanderhaeghen, R., De Block, S., Van Montagu, M. and Van Lijsebettens, M. (2001) *AtCSLD3*, a cellulose synthase-like gene important for root hair growth in *Arabidopsis*. *Plant Physiol.* **126**, 575–586.
- Wang, Y., Chen, X., Wang, J., Xun, H., Sun, J. and Tang, F. (2016) Comparative analysis of the terpenoid biosynthesis pathway in *Azadirachta indica* and *Melia azedarach* by RNA-seq. *Springerplus* **5**, 819.
- Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H. et al. (2012) MCS-X: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49.
- Xie, J., Li, M., Zeng, J., Li, X. and Zhang, D. (2022) Single-cell RNA sequencing profiles of stem-differentiating xylem in poplar. *Plant Biotechnol. J.* **20**, 417–419.
- Xu, H.H., Lai, D. and Zhang, Z.H. (2017) Research and application of botanical pesticide azadirachtin. *J. South China Agric. Univ.* **38**, 1–11.
- Xu, W., Cheng, H., Zhu, S., Cheng, J., Ji, H., Zhang, B., Cao, S. et al. (2021) Functional understanding of secondary cell wall cellulose synthases in *Populus trichocarpa* via the Cas9/gRNA-induced gene knockouts. *New Phytol.* **231**, 1478–1495.
- Xue, Z., Duan, L., Liu, D., Guo, J., Ge, S., Dicks, J., ÓMáille, P. et al. (2012) Divergent evolution of oxidosqualene cyclases in plants. *New Phytol.* **193**, 1022–1038.
- Zhang, D., Hu, Q., Liu, X., Zou, K., Sarkodie, E.K., Liu, X. and Gao, F. (2020) AllEnricher: a comprehensive gene set function enrichment tool for both model and non-model species. *BMC Bioinform.* **21**, 106.
- Zhang, Y. and Xu, H. (2017) Recent progress in the chemistry and biology of limonoids. *RSC Adv.* **7**, 33522–35191.
- Zheng, X., Li, P. and Lu, X. (2019) Research advances in cytochrome P450-catalysed pharmaceutical terpenoid biosynthesis in plants. *J. Exp. Bot.* **70**, 4619–4630.
- Zhou, F. and Pichersky, E. (2020) The complete functional characterisation of the terpene synthase family in tomato. *New Phytol.* **226**, 1341–1360.
- Zwaenepoel, A., Van de Peer, Y. and Hancock, J. (2019) wgd-simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* **35**, 2153–2215.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Basic data figures for genome assembly and annotation.

Figure S2 Phylogenetic analysis of Meliaceae species.

Figure S3 Screening for WGD events of Phylogenetic analysis of Meliaceae species.

Figure S4 GO annotation of expansion proteins of *A. indica* (A, B) and *M. azedarach* (C, D).

Figure S5 Gene family analysis of cellulose and hemicellulose biosynthesis genes.

Figure S6 Gene family analysis of lignin biosynthesis genes.

Figure S7 The whole-transcriptional regulation network in *A. thaliana* with STRING.

Figure S8 Gene family analysis of NAC transcription factors.

Figure S9 Gene family analysis of MYB transcription factors.

Figure S10 Gene family analysis of WRKY transcription factors.

Figure S11 Gene family analysis of bHLH transcription factors.

Figure S12 Gene family analysis of isoprenoid biosynthesis-related genes.

Figure S13 Gene family analysis of carbon chain extension and cyclization-related genes.

Table S1 Statistics for assembly basic data of the *A. indica* and *M. azedarach* genome

Table S2 Statistics for assembly completeness of the *A. indica* and *M. azedarach* genome

Table S3 Statistics for final assembly of the *A. indica* and *M. azedarach* genome

Table S4 Statistics for HiC assembly of the *A. indica* and *M. azedarach* genome

Table S5 Statistics for chromosomes of the *A. indica* and *M. azedarach* genomes

Table S6 Statistics of noncoding RNA of the *A. indica* and *M. azedarach* genome

Table S7 Statistics for repeat annotations of the *A. indica* and *M. azedarach* genome

Table S8 Statistics of completeness and annotated number of *A. indica* and *M. azedarach* genome

Table S9 Statistics of functional annotated genes of the *A. indica* and *M. azedarach* genome

Table S10 Genome size estimation of nine melican species with GenomeScope (k = 21)

Table S11 published sapindales genomes (March 30th, 2022)

Table S12 Statistical table of secondary wall biosynthesis-related genes

Table S13 Statistical table of limonoids biosynthesis-related genes.