




# The chromosome-scale genomes of *Dipterocarpus turbinatus* and *Hopea hainanensis* (Dipterocarpaceae) provide insights into fragrant oleoresin biosynthesis and hardwood formation

Sibo Wang<sup>1,†</sup>, Hongping Liang<sup>1,2,†</sup>, Hongli Wang<sup>1,2</sup>, Linzhou Li<sup>1,3</sup>, Yan Xu<sup>1,2</sup>, Yang Liu<sup>1</sup>, Min Liu<sup>1</sup>, Jinpu Wei<sup>1</sup>, Tao Ma<sup>4</sup> , Cheng Le<sup>5</sup>, Jinlong Yang<sup>5,6</sup>, Chengzhong He<sup>7</sup>, Jie Liu<sup>8</sup>, Jianming Zhao<sup>8</sup>, Yuxian Zhao<sup>9</sup>, Michael Lisby<sup>10</sup>, Sunil Kumar Sahu<sup>1,\*</sup>  and Huan Liu<sup>1,2,\*</sup> 

<sup>1</sup>State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen, China

<sup>2</sup>College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China

<sup>3</sup>Department of Biotechnology and Biomedicine, Technical University of Denmark, Lyngby, Denmark

<sup>4</sup>Key Laboratory of Bio-resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, China

<sup>5</sup>BGI-Yunnan, BGI-Shenzhen, Yunnan, China

<sup>6</sup>College of Forensic Science, Xi'an Jiaotong University, Xi'an, China

<sup>7</sup>Southwest Forestry University, Kunming, Yunnan, China

<sup>8</sup>Forestry Bureau of Ruili, Yunnan Dehong, Ruili, China

<sup>9</sup>Chinese Academy of Forestry, Beijing, China

<sup>10</sup>Department of Biology, University of Copenhagen, Copenhagen, Denmark

Received 13 August 2021;

revised 1 October 2021;

accepted 12 October 2021.

\*Correspondence (Tel +86 136 128 80402;

fax 0755 33945530;

email liuhuan@genomics.cn) (H. L.) and

(Tel +86 135 351 24942; fax 0755

33945530;

email sunilkumarsahu@genomics.cn

(S. K. S.))

<sup>†</sup>Equal contribution.

## Summary

Dipterocarpaceae are typical tropical plants (dipterocarp forests) that are famous for their high economic value because of their production of fragrant oleoresins, top-quality timber and usage in traditional Chinese medicine. Currently, the lack of Dipterocarpaceae genomes has been a limiting factor to decipher the fragrant oleoresin biosynthesis and gain evolutionary insights into high-quality wood formation in Dipterocarpaceae. We generated chromosome-level genome assemblies for two representative Dipterocarpaceae species *viz.* *Dipterocarpus turbinatus* Gaertn. f. and *Hopea hainanensis* Merr. et Chun. Our whole-genome duplication (WGD) analysis revealed that Dipterocarpaceae underwent a shared WGD event, which showed significant impacts on increased copy numbers of genes related to the biosynthesis of terpene, *BAHD* acyltransferases, fatty acid and benzenoid/phenylpropanoid, which probably confer to the formation of their characteristic fragrant oleoresin. Additionally, compared with common soft wood plants, the expansion of gene families was also found to be associated with wood formation, such as in *CESA* (cellulose synthase), *CSLE* (cellulose synthase-like protein E), laccase and peroxidase in Dipterocarpaceae genomes, which might also contribute to the formation of harder, stronger and high-density timbers. Finally, an integrative analysis on a combination of genomic, transcriptomic and metabolic data from different tissues provided further insights into the molecular basis of fragrant oleoresins biosynthesis and high-quality wood formation of Dipterocarpaceae. Our study contributes the first two representative genomes for Dipterocarpaceae, which are valuable genetic resources for further researches on the fragrant oleoresins and superior-quality timber, genome-assisted breeding and improvement, and conservation biology of this family.

**Keywords:** genome, long reads, Dipterocarpaceae, whole-genome duplication, Evolution, fragrance, oleoresin, wood formation.

## Introduction

Dipterocarpaceae plants play an important ecological role in studying the succession of forest communities because they are representative tropical tree species of pantropical rainforests (Sasaki, 2006). The Dipterocarpaceae family belongs to the order of the Malvales and consists of 16 genera and about 695 known species around the world (Appanah and Turnbull, 1998; Christenhusz and Byng, 2016), which are mainly distributed in tropical lowland wet rainforest area, from South America to Africa (Seychelles), and India,

as well as Southeast Asia areas, *that is* China, Indonesia, Malaysia and Philippines (Appanah and Turnbull, 1998).

Dipterocarpaceae plants are of high economic value due to the presence of unique fragrant oleoresins, high-quality timber and bioactive components for the preparation of traditional Chinese medicines (Dyrmosse *et al.*, 2017; Rana *et al.*, 2010; Shen *et al.*, 2017). The *Dipterocarpus* species serves as the principal source of fragrant oleoresins, which are present primarily in their tree trunk, and is a mixture of amorphous polymer compounds (Appanah and Turnbull, 1998). Other Dipterocarpaceae genera

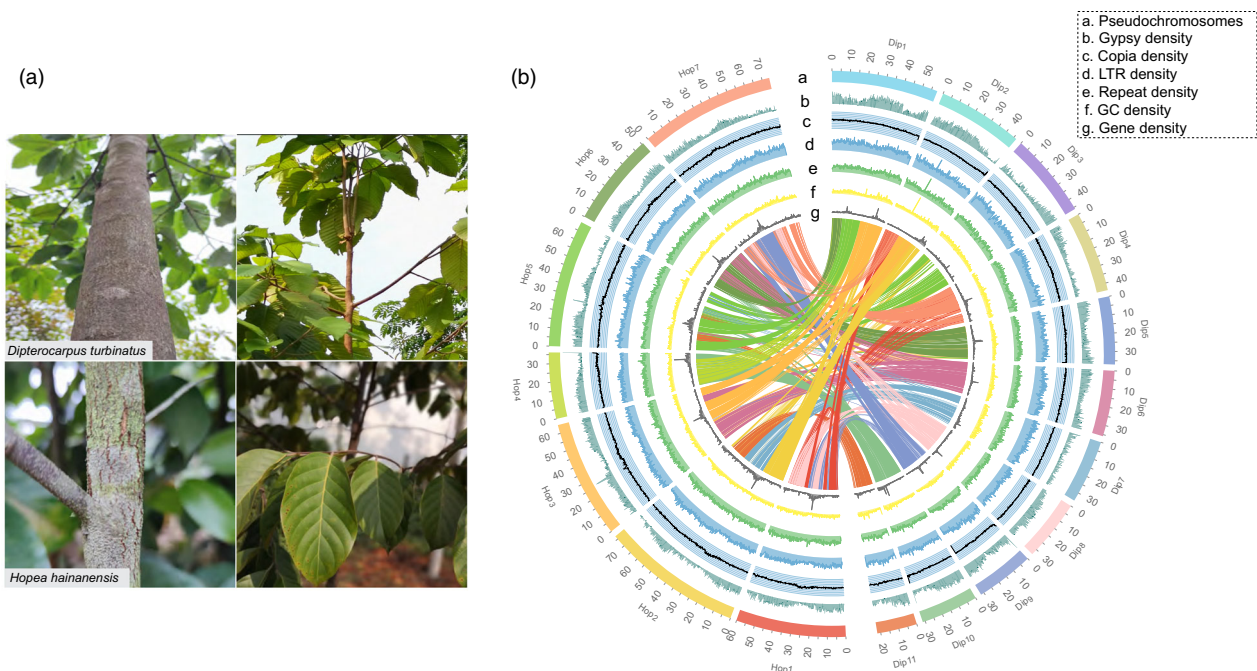
also produce fragrant oleoresins but are not that important, such as *Shorea*, *Vatica*, *Hopea*, *Dryobalanops* and *Parashorea* (Anarkarjard and Kegl, 1998; Appanah and Turnbull, 1998). A famous oleoresin is procured from *D. turbinatus* wood (Figure 1a), which is the principal source to produce 'keruing' through the distillation and purification of fragrant oleoresins and also a famous perfumery over the world (Appanah and Turnbull, 1998; Aslam and Ahmad, 2015). The fragrance of oleoresins is good for health, as documented in ancient Chinese medical literature, such as inducing resuscitation, clearing away heat and detoxifying, relief of swelling, pain and tumescence (Peter and Babu, 2012; Zhou and Yang, 2017). The aromatic resins secreted from *D. turbinatus* are treated as 'panacea' for treatments of rheumatism, stubborn eczema and other skin diseases in Xishuangbanna of China (Zhou and Ren, 2007). Modern pharmacological and chemical studies also support the potential medicinal value of fragrant resins of Dipterocarpaceae since it is featured with antioxidant, antibacterial, anti-fungi and anti-inflammatory activities (Aslam *et al.*, 2015; Yongram *et al.*, 2019). The composition of oleoresins extracted from Dipterocarpaceae is exceedingly complex, including dozens of chemical compounds (Kamariyah *et al.*, 2012); however, the main active compounds present in aromatic oleoresins of Dipterocarpaceae are various terpenoids (Kamariyah *et al.*, 2012; Messer *et al.*, 1990). Notably, it consists of distinct sesquiterpenes, gurjunene, which may be one of the main components that contribute to the unique fragrance of Dipterocarpaceae (Appanah and Turnbull, 1998).

Additionally, Dipterocarpaceae are highly demanded in the plywood industry in tropical Asia. The wood of Dipterocarpaceae is hard and tight with fine texture and strong moisture resistance (Rana *et al.*, 2009). Especially, *Hopea hainanensis* (Figure 1a) is the wood, which grows much slower and needs hundreds of years to attain full maturity (Appanah and Turnbull, 1998;

Oldfield and Lusty, 1998). However, its wood is extremely precious and famous for high strength, corrosion and insect resistance, durable and long-lasting (Appanah and Turnbull, 1998). The wood of Dipterocarpaceae is widely used as bridge, ship and furniture construction materials (Appanah and Turnbull, 1998).

Many species of Dipterocarpaceae such as *Dipterocarpus* and *Hopea* are now endangered (Oldfield *et al.*, 1998), mainly because of their low genetic diversity and overconsumption (Meng and Xu, 2005). There are a large number of inbreeds in the population of Dipterocarpaceae, resulting in a large loss of genetic variation of this species (Appanah and Turnbull, 1998; Meng and Xu, 2005; Zhou and Yang, 2017). Additionally, the overexploitation of natural populations and destruction of wild habitat further resulted in a dramatic decline of the population of Dipterocarpaceae in the past decades (Meng and Xu, 2005). Considering the increasing demands in the market of fragrant resins and timber of Dipterocarpaceae and decreasing population in many Dipterocarpaceae plants, it is important to interrogate the genomic background to explore the genome feature of Dipterocarpaceae and to accelerate genome-assisted improvement in breeding systems.

Although several woody and economically important plant genomes have been reported earlier (Baek *et al.*, 2018; Butkhuip *et al.*, 2011; Chang *et al.*, 2018; Fan *et al.*, 2020, 2021; Hofmeister *et al.*, 2020; Sahu *et al.*, 2019; Shang *et al.*, 2020b; Zhang *et al.*, 2019), the lack of Dipterocarpaceae genome has been a limiting factor to decipher the fragrant oleoresin biosynthesis and gain evolutionary insights into high-quality wood formation in Dipterocarpaceae. In this study, high-quality chromosome-level genome assemblies of *D. turbinatus* and *H. hainanensis* (Figure 1b) were obtained by combining the Oxford Nanopore long-read sequencing and Hi-C scaffolding data. Comparative analyses of the *D. turbinatus* and *H. hainanensis*



**Figure 1** Morphology and genome features of *D. turbinatus* and *H. hainanensis*. (a) Leaf and stem of *D. turbinatus* and *H. hainanensis*, respectively. (b) The genomic landscape of *D. turbinatus* and *H. hainanensis*.

genomes with the other representative fragrant plants and timber trees provided an impetus to explore the evolution and differentiation mechanisms of fragrance biosynthesis and wood formation in Dipterocarpaceae. Through gene mining, exploration of transcriptome and metabolic data generated from different tissues, we present new insights into the molecular basis of fragrant resins biosynthesis and wood formation of Dipterocarpaceae. Our study shall serve as a foundation for future research on the evolution, phytochemistry and ecology of Dipterocarpaceae.

## Results

### Genome assembly and annotation

The genomes sizes of *D. turbinatus* and *H. hainanensis* were estimated with K-mer analysis to be 426.6 and 442.2 Mb, respectively (Additional file 1: Figure S1). The two genomes were sequenced and assembled using a combination of Oxford Nanopore long-read, MGI-SEQ short read and Hi-C paired-end read data. After primary assembly, correction, polishing and scaffolding, final assemblies of 421.2 and 434.3 Mb with contig N50 of 29 Mb and 9 Mb were obtained for *D. turbinatus* and *H. hainanensis*, respectively (Table 1). To refine both assemblies of *D. turbinatus* and *H. hainanensis*, their Hi-C data were mapped to their draft assemblies; as a result, 99.9% and 99.6% of the entire genome sequences were effectively anchored, structured and oriented into 11 and 7 pseudo-chromosomes, respectively (Figure 1b, Figure 2, Additional file 1: Figure S2, Additional file 2: Table S1), which is in accordance with the reported number of chromosomes in somatic cells of *D. turbinatus* ( $2n = 22$ ) and *H. hainanensis* ( $2n = 14$ ) (Appanah and Turnbull, 1998). To assess the quality of the two genome assemblies, we performed BUSCO analysis and found that 90.7% and 91.4% complete eukaryotic conserved genes exist in the *D. turbinatus* and *H. hainanensis* genomes, respectively (Table 1 and Additional file 2: Table S2). Moreover, the LTR Assembly Index (LAI) value was 14.27 and 17.01 for *D. turbinatus* and *H. hainanensis*, respectively. Taken together, the above results indicated high degree of contiguity and completeness of the two Dipterocarpaceae genomes according to the current standards (Ellinghaus and Kurtz, 2008; Ou and Chen, 2018; Xu and Wang, 2007).

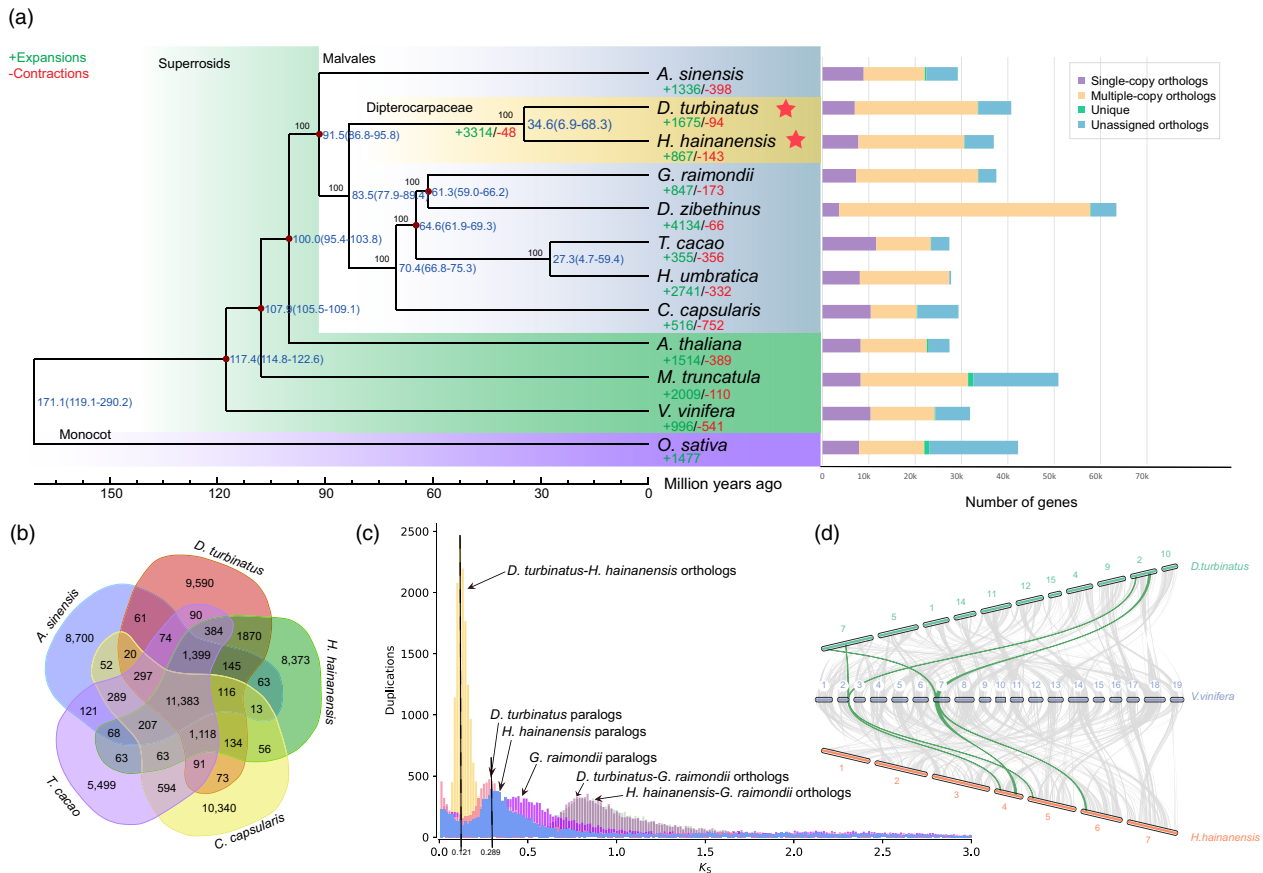
Based on the high-quality genomes of *D. turbinatus* and *H. hainanensis*, we found that *D. turbinatus* and *H. hainanensis* genomes contain 46.4% (195.5 Mb) and 52.4% (227.6 Mb) transposable elements, respectively (Additional file 2: Table S3). Long terminal repeats retrotransposons (LTR-RTs) are the predominant components and comprise 26.5% and 31.7% of the genomes of *D. turbinatus* and *H. hainanensis*, respectively. Among the LTRs, the *Gypsy* elements were the most abundant in both genomes, followed by the *Copia* elements (Additional file 2: Table S3). The majority of intact LTRs of *Ty3/gypsy* and *Ty1/copia* in both Dipterocarpaceae genomes showed a similar pattern of insertion time of LTR-RTs (Additional file 1: Figure S3), indicating LTR elements in Dipterocarpaceae underwent recent bursts and suggesting the major proportion of LTR-RT elements in Dipterocarpaceae became active recently (Liu *et al.*, 2021). By combining *ab initio*, homologue-based and transcriptome-based approaches, a total of 40 707 and 36 967 protein-coding genes were predicted from *D. turbinatus* and *H. hainanensis* genomes, respectively, of which 88.6% and 89.7% shared homologs with annotated genes in public protein databases (Table 1 and Additional file 2: Table S4).

**Table 1** Assembly and annotation features of the *D. turbinatus* and *H. hainanensis* genome

Species	<i>D. turbinatus</i>	<i>H. hainanensis</i>
Assembly feature		
Estimated genome size	426.64 Mb	442.20 Mb
Assembled genome size	421 171 415	434 306 036
GC content	32.71%	32.91%
N50 of contigs (bp)	245 261	6 614 031
N50 of scaffold (bp)	29 439 453	9 083 733
Total length of contig	421 013 775	434 305 978
Longest scaffold	49 593 048	30 910 234
Complete BUSCOs	90.70%	91.40%
Genome annotation		
Repeat region	46.40%	50.70%
Number of protein-coding genes	40 707	36 967
Average length of transcripts (bp)	2459.20	2509.02
Average exon length (bp)	205.48	206.30
Average intron length (bp)	346.05	353.34
HiC		
Anchor size	420 837 890	432 493 699
Anchor rate	99.92%	99.58%
Number of pseudo-chromosomes	11	7
N50 of scaffold (bp)	37 048 376	69 706 040
Longest scaffold	58.2 M	76.9 M

### Phylogenetics and genome evolution of Dipterocarpaceae

A phylogenetic tree was constructed for 12 selected plant species, including 8 representative species from Malvales, using genes extracted for 294 orthologous single-copy nuclear genes. Molecular dating analysis suggests that Dipterocarpaceae diverged from the most recent common ancestor with Malvaceae at around 83.5 Mya, followed by the divergence of *Dipterocarpus* and *Hopea* at around 34.6 Mya (Figure 2a). The topology of our phylogenetic tree constructed by nuclear single-copy gene contradicted with the previous study that used several discrete genes (Hernandez-Gutierrez and Magallon, 2019). Their phylogeny showed a sister relationship between Thymelaeaceae and Dipterocarpaceae. However, their result also conflicts with other previous phylogenetic studies based on complete chloroplast genome sequences (Cvetković and Hsingier, 2017; Lee *et al.*, 2019; Yan *et al.*, 2019). To further prove our nuclear phylogeny, we employed concatenation and coalescent methods by using RAXML and ASTRAL among 14 representative species (including 13 Malvales, transcriptome data from 1KP) (Leebens-Mack *et al.*, 2019), and both concatenation and coalescent phylogenetic trees showed consistent results with the phylogeny of Figure 2a with high bootstrap values (Additional file 1: Figure S4). To test whether there is cytonuclear discordance (the discordance between nuclear and organellar phylogenies) for Thymelaeaceae and Dipterocarpaceae, we assembled the chloroplast genomes of *D. turbinatus* and *H. hainanensis* in this study. Additionally, we also downloaded all the plastid genomes of Malvales from NCBI and performed phylogenetic analysis based on the complete chloroplast genome sequences by using RAXML and ASTRAL. Interestingly, both chloroplast genome-based phylogenetic trees showed a similar topology that *D. turbinatus* and *H. hainanensis*



**Figure 2** Evolution of the Dipterocarpaceae genomes and gene families. (a) Phylogenetic tree constructed by maximum likelihood based on the concatenation of single-copy nuclear genes, the distribution of genes in each species is shown in the right panel. (b) Venn diagram of gene families in Dipterocarpaceae and other Malvales species. (c) The distributions frequencies of synonymous substitutions ( $K_s$ ) for orthologs among two Dipterocarpaceae and *G. raimondii* (*D. turbinatus* (Dip), *H. hainanensis* (Hop) and *G. raimondii* (Goss)). (d) Synteny patterns between genomic regions from two Dipterocarpaceae and *V. vinifera*. A collinear relationship is highlighted by one syntenic set shown in green colours.

of this study clustered with other Dipterocarpaceae species, and a sister relationship could be observed between Thymelaeaceae and Dipterocarpaceae (Additional file 1: Figure S4). Based on the phylogeny, we found 3,314 gene families expanded and 48 gene families contracted in the ancestor of Dipterocarpaceae. We also performed the KEGG enrichment for the expanded gene families in the common ancestor of the two newly sequenced species, which showed the expansion of mRNA surveillance pathway, Propanoate metabolism, Terpenoid backbone biosynthesis, and Pantothenate, CoA biosynthesis and Brassinosteroid biosynthesis pathways (Additional file 2: Table S5). From the gene families clustered with the other three representative species of the Malvales family (*Aquilaria sinensis*, *Theobroma cacao* and *Corchorus capsularis*), 19 833 gene families were Dipterocarpaceae-specific while 11 383 gene families were shared among all the selected species of Malvales family (Figure 2c). Functional analysis showed that unique gene families in Dipterocarpaceae were preferentially enriched in the terms monoterpenoid biosynthesis, flavone and flavonol biosynthesis and fatty acid biosynthesis (Additional file 2: Table S6).

To investigate whether whole-genome duplication (WGD) events happened in Dipterocarpaceae, the number of synonymous substitutions per synonymous site ( $K_s$ ) was characterized of *D. turbinatus*, *H. hainanensis*, *C. capsularis*, *A. sinensis* and *G.*

*raimondii*, respectively. *D. turbinatus* and *H. hainanensis* displayed almost the same  $K_s$  distributions of all paralogous gene pairs, and it unveiled evidence for a recent WGD event in the common ancestor of Dipterocarpaceae ( $K_s$  peak at  $\sim 0.29$ ) after their divergence with *Corchorus* (Figure 2a and c). To provide additional evidence of shared WGD event between the *D. turbinatus* and *H. hainanensis*, we extracted paralogous pairs of *D. turbinatus* and *H. hainanensis* genes derived from their respective WGDs and constructed phylogenetic trees to further confirm that the WGD event was shared by *D. turbinatus* and *H. hainanensis* (Additional file 1: Figure S5). In addition, we detected a 2:2:1 syntenic relationship among *D. turbinatus*, *H. hainanensis* and *V. vinifera*, which provided additional evidence for a WGD event in the common ancestor of *Dipterocarpus* and *Hopea* (Figure 2d and Additional file 1: Figure S6). That is, a single *V. vinifera* region could be aligned to two genomic regions in the *D. turbinatus* or *H. hainanensis* genome (Figure 2e).

#### Gene duplication contributes to the aromatic scent in Dipterocarpaceae

Gene duplication has long been regarded as one of the major driving forces in plant evolution, which may endow genes with potential sub-functionalization and neo-functionalization (Liu

*et al.*, 2021; Qiao *et al.*, 2019). We identified a total of 33 230 and 29 176 duplicated genes from *D. turbinatus* and *H. hainanensis* genomes, respectively, which were classified into five categories, that is, the WGD duplicates, tandem duplicates (TD), transposed duplicates (TRD), proximal duplicates (PD) and dispersed duplicates (DD) (Additional file 2: Table S7). Both sequenced species in this study exhibited a similar trend concerning the numbers in each category of the duplicates.

Next, we analysed the duplication-induced expansion in gene families by combining and overlapping each type of duplicated gene and expanded gene families (EGFs) (Figure 3a). WGD events contributed to the highest proportion of expansion of the gene families compared with other duplication types in both *D. turbinatus* and *H. hainanensis*. KEGG enrichment of expanded gene families in *D. turbinatus* and *H. hainanensis* indicated that, in *D. turbinatus*, the expansion of terpenoid backbone biosynthesis genes might be mainly caused by WGD events; the expansion of sesquiterpenoid and triterpenoid biosynthesis gene families was induced by TD; and the expansion of phenylpropanoid biosynthesis gene families derived from DSD events. In comparison, in *H. hainanensis*, TRD but WGD events contributed to the expansion of terpenoid backbone biosynthesis. Additionally, TD-involved gene family in *H. hainanensis* was in response to the expansion of butanoate metabolism that is also related to fragrant volatiles (Butkhop *et al.*, 2011). Similar to *D. turbinatus*, DSD-involved gene family expansion also participated in the expansion of Phenylpropanoid biosynthesis genes in *H. hainanensis*. Taking these together, whole-genome duplication and other types of duplication events (such as tandem duplications, dispersed duplicates) have significant impacts on the increase of copy numbers in genes related to terpene and phenylpropanoid biosynthesis that are involved in the generation of aromatic scent, which may contribute to the characteristic fragrance (aroma component) in Dipterocarpaceae (Additional file 2: Table S8).

### Evolution of terpene biosynthesis and regulation-related genes

The containing of volatile organic compounds is the main reason behind the process of unique fragrance of Dipterocarpaceae (Shang *et al.*, 2020a; Yang *et al.*, 2018). Previous metabolic studies on the oleoresins of *D. turbinatus* and *H. hainanensis* showed that dozens of distinct chemical constituents, including the azulone,  $\alpha$ -gurjunene,  $\alpha$ -copaene,  $\delta$ -elemene and borneol, exist in these two species (Zhou and Ren, 2007), which is consistent with our results of chemical detection analyses (Additional file 2: Table S8). Monoterpenes, sesquiterpenes and iridoids are usually generated via the 2-C-methyl-D-erythritol 4-phosphate (MEP) pathway and the mevalonate (MVA) pathway (Figure 4a). A total of 55/58 associated genes from two pathways were found in the *D. turbinatus* and *H. hainanensis* genomes, respectively (Additional file 2: Table S9). Our results showed that the copy number of some of these genes was expanded in both *D. turbinatus* and *H. hainanensis* genomes. For example, genes of Isopentenyl-diphosphate Delta-isomerase, which is responsible for a reversible reaction between isopentenyl diphosphate and dimethylallyl diphosphate for sesquiterpene and monoterpene biosynthesis, expanded in both *D. turbinatus* and *H. hainanensis* (Additional file 2: Table S9). Notably, gurjunene (sesquiterpene) and borneol (monoterpene) might play key roles in the contribution of unique fragrance in Dipterocarpaceae (Appanah and Turnbull, 1998). The gene expression profile across wood and leaf

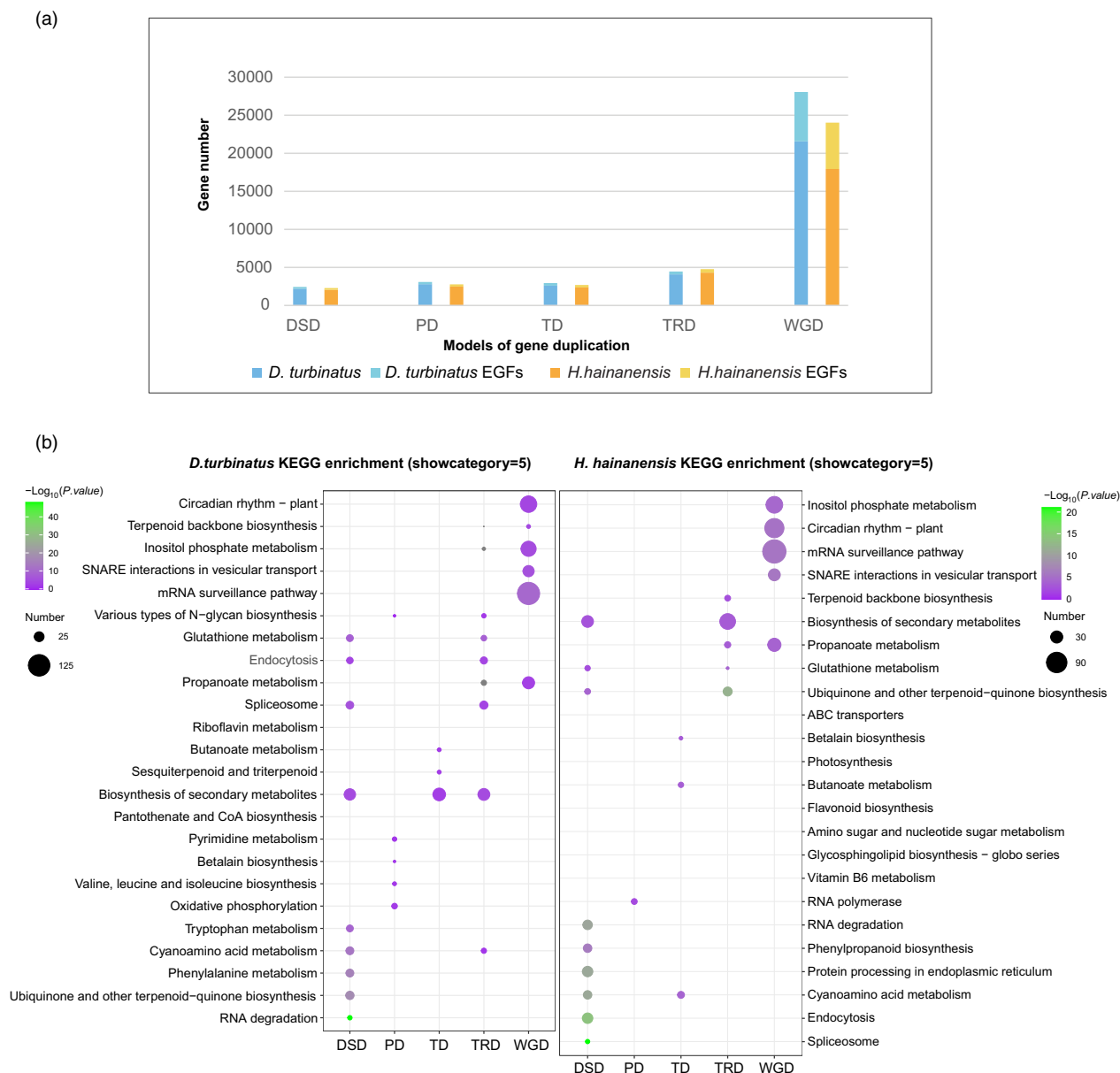
tissues revealed transcripts of numerous MEP/MVA pathway genes, such as HMGR, DXS and GGPPS genes, were most abundant in wood tissue, which coincided with the fact that fragrant oleoresins only accumulated in tree wood but not leaves. (Figure 4a). Interestingly, we identified the number of *LAMT* homologs, which are key enzymes for biosynthesis of Iridoids, showed remarkable expansion in both *D. turbinatus* and *H. hainanensis* compared with other plants (Figure 4b). Terpene synthases (TPSs) are the vital enzymes responsible for the catalytic reaction in the MVA and MEP pathway to generate a basic skeleton of terpenoid compounds. We identified 20/21 TPS genes in *D. turbinatus* and *H. hainanensis*, respectively (Figure 5a and Additional file 1: Figure S7), which did not exhibit a remarkable difference in TPS number compared with other selected plants (Additional file 2: Table S10). Similar to previous results in other plants (Xu *et al.*, 2020; Li, Wang, *et al.*, 2021), many TPS genes in *D. turbinatus* and *H. hainanensis* also exhibited a tandem array, and these genes formed TPS gene clusters on chromosomes 4, 8, 10 and 1, 7 (Figure 5b and Additional file 1: Figure S8), suggesting these genes underwent recent tandem duplication events.

The *TPS-a* and *TPS-b* genes encode angiosperm-specific sesquiterpene and monoterpene synthases predominantly (Li, Wang, *et al.*, 2021; Shang *et al.*, 2020a). Surprisingly, *D. turbinatus* only encodes two *TPS-a* (Figure 5a), which is remarkably less than that of *H. hainanensis* and *Aquilaria sinensis*. Phylogenetic analysis of the TPS gene family from selected plants revealed that the *TPS-a* from *D. turbinatus* and *H. hainanensis* forms an individual subclade (Figure 5a), suggesting lineage-specific of *TPS-a* genes might own unique function contribute to the sesquiterpenes (especially for the distinct sesquiterpenes of Dipterocarpaceae essential oil, *i.e.*, alpha and beta gurjunene) accumulation of fragrant resins in *D. turbinatus* and *H. hainanensis*. The comparison of the expression level of TPS genes of *D. turbinatus* and *H. hainanensis* between wood and leaf tissues revealed that the expression levels of Diptu\_12379.t1 (*TPS-a*) and Diptu\_12960.t1 (*TPS-b*) were significantly higher in wood than leaves (Figure 5d), suggesting they might play a primary role in the biosynthesis of sesquiterpene and monoterpene, the main components of fragrant oleoresin of *D. turbinatus*. Terpenes are usually modified by BAHD acyltransferases to produce esters, which are involved in the synthesis of various flavours and fragrances in plants (Xu *et al.*, 2020; Shang *et al.*, 2020a). A total of 51 and 40 BAHD acyltransferases were identified in the genomes of *D. turbinatus* and *H. hainanensis*, respectively (Figure 5c and Additional file 2: Table S11). Phylogenetic analysis revealed the number of BAHD acyltransferases V type in both *D. turbinatus* and *H. hainanensis* was prominently higher than that of *A. sinensis*, *C. capsularis*, *V. vinifera* and *T. cacao*. By comparing the expression level between wood and leaf tissues in *D. turbinatus* and *H. hainanensis*, we found that many BAHD acyltransferase genes from BADH-Ia, BADH-IIIa and BADH-V groups were highly expressed in wood than in leaf tissue of both *D. turbinatus* and *H. hainanensis*, indicating BAHD acyltransferases might also contribute to the unique aroma of Dipterocarpaceae (Figure 5d).

### Evolution of benzenoid/phenylpropanoid biosynthesis-related genes

In plants, Phenylpropanoids and benzenoids are one of the largest class of compounds responsible for the fragrances (Li, Chen, *et al.*, 2021; Zhao *et al.*, 2017). We therefore investigated the

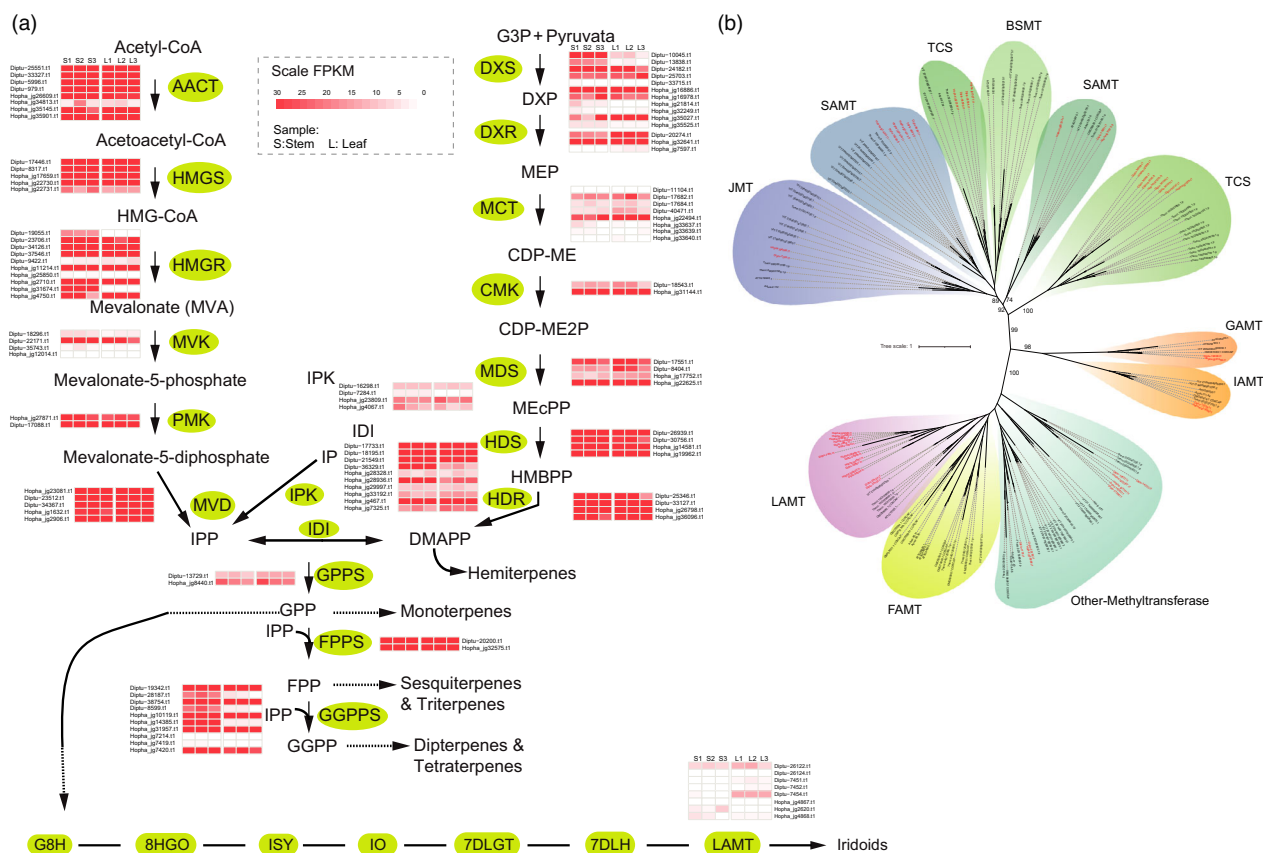




**Figure 3** Gene duplication and evolution. (a) Stacked columns chart shows the number of gene duplication in various duplicated modes and gene duplication-induced expanded gene number. (b) Functional enrichment of genes overlapping between expanded gene families and various modes of gene duplications.

expression of key genes involved in benzenoid/phenylpropanoid biosynthesis in both Dipterocarpaceae genomes (Additional file 2: Table S12). As phenylpropanoid biosynthesis and phenylalanine metabolism are two different metabolic pathways, and they share phenylalanine precursors (at the upstream), we considered both the pathway-related genes to avoid the bias while performing the functional annotation, and found that while the majority of genes involved in phenylpropanoid biosynthesis exhibited higher expression, there were also many genes from phenylalanine synthesis that also displayed higher expression (Additional file 2: Table S13). We found that most of these genes showed remarkably higher level of expression in wood tissue than in leaf tissue in both *D. turbinatus* and *H. hainanensis* (Additional file 2: Table S13). acetyl-CoA:benzyl alcohol acetyltransferase (*BEAT*) plays a key role in yielding benzyl acetate in plants for floral scent,

and we identified 9/6 *BEAT* homologous genes in *D. turbinatus* and *H. hainanensis* (Additional file 2: Table S12). We found two *BEATs* displayed remarkably higher expression levels in stem tissue of *D. turbinatus* than in leaf tissue (Figure 5d), indicating a higher activity of benzyl acetate biosynthesis in the stem tissue for the aroma of *D. turbinatus*. The production of phenylpropanoid/benzenoid compounds in plants is related to the SABATH families (Xu *et al.*, 2020). Phylogenetic analysis showed 29/24 SABATH homologs, including 1/1 *IAMT*, 1/1 *JMT*, 6/5 *SAMT*, 1/1 *GAMT*, 8/8 *LAMT*, 8/5 *TCS* and 6/3 others, existed in genomes of *D. turbinatus* and *H. hainanensis*, respectively (Figure 4b, 6 and Additional file 2: Table S14). Additionally, both the *COMT* and *ICMT*, belonging to the SAM-binding methyltransferase superfamily, are thought to be involved in aromatic compound metabolism. Phylogenetic analysis showed both *D. turbinatus*



**Figure 4** Biosynthetic pathway of MEP/MVA in Dipterocarpaceae. (a) MEP/MVA biosynthesis pathways in Dipterocarpaceae leaf and stem based on transcriptomic analyses. (b) Maximum likelihood phylogenetic tree showing the classification and copy number of SABATH family.

and *H. hainanensis* own similar copy number of *COMT* and *ICMT* compared with other selected plants (Additional file 1: Figure S9 and Additional file 2: Table S14). However, the expression level of these genes in stem tissue from both *D. turbinatus* and *H. hainanensis* was mostly higher than their respective leaf tissue. Thus, transcriptomic data provided evidence for the higher activation of *COMT* and *ICMT* genes in stem tissue of two the Dipterocarpaceae species than in their leaf tissues (Additional file 2: Table S13).

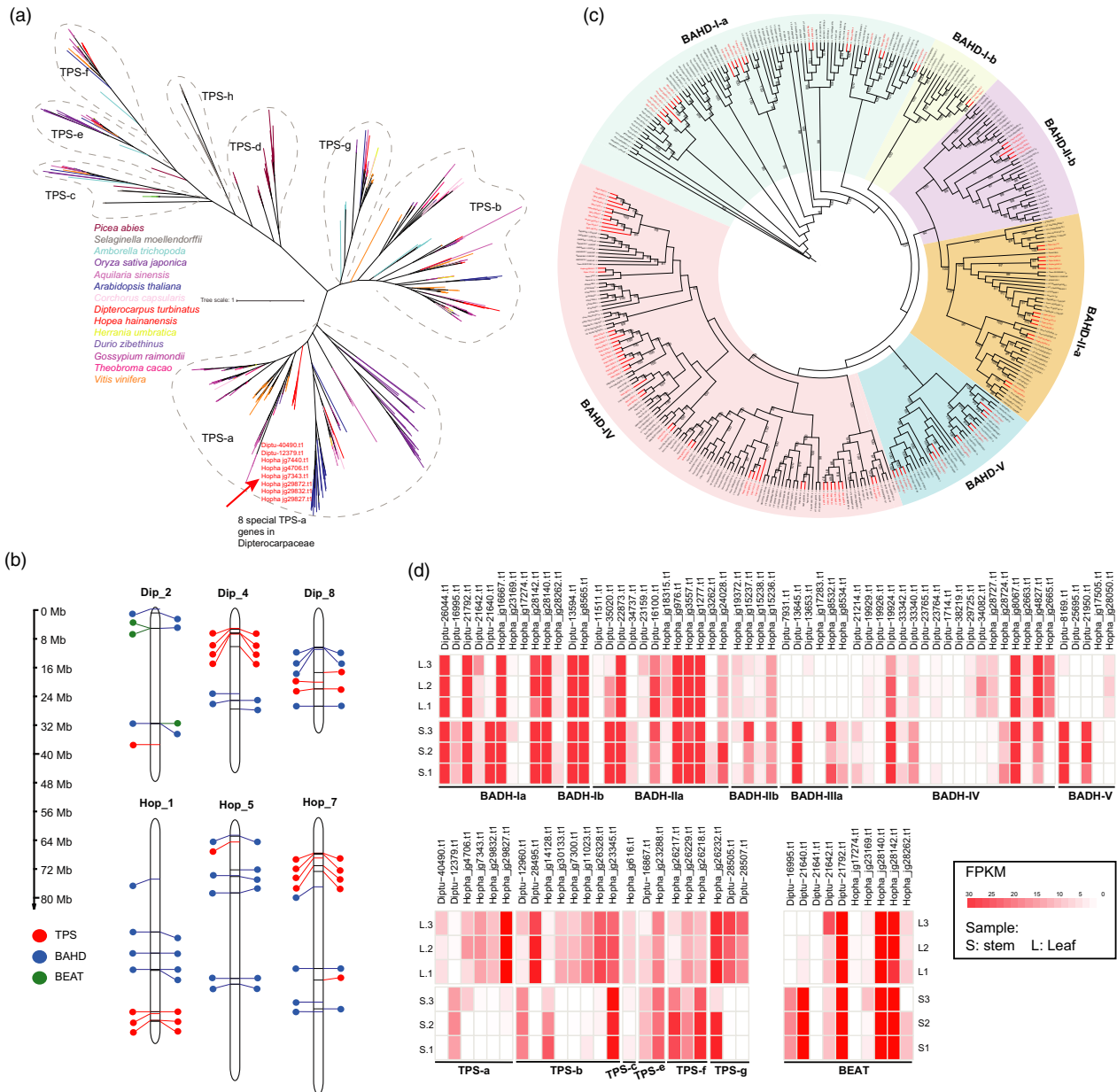
#### Hardwood formation in Dipterocarpaceae

Hardness, high strength and strong moisture resistance are the typical features of wood of Dipterocarpaceae plants (Appanah and Turnbull, 1998). To explore the genetic basis of the wood formation of Dipterocarpaceae, comparative genomics analyses on the genes related to the cell wall and lignin metabolic pathways were performed among the genomes of Dipterocarpaceae and other representative tree species characterized by relatively loose and soft wood texture (i.e. *Populus trichocarpa* and *A. sinensis*). *D. turbinatus* and *H. hainanensis* could encode 64/60 cellulose synthase (*CESA*) (including cellulose synthase-like (*CSL*)), respectively, which is remarkably higher than the number of two representative softwood trees (*Populus trichocarpa* and *A. sinensis*) and other phylogenetic affiliation species (Figure 6a). Phylogenetic analysis indicated that the copy number of *CESA* and *CSLE* exhibited expansion in both *D. turbinatus* and *H. hainanensis* compared with that of *P. trichocarpa* and *A. sinensis* (Figure 6b). *CESA* is involved in the primary cell wall formation

and considered as the most important enzyme involved in the synthesis of cellulose microfibrils in plant cells (Kumar and Turner, 2015). Interestingly, the number of laccase and peroxidase participating in the lignin metabolism, in both *D. turbinatus* and *H. hainanensis*, showed remarkable expansions compared with *P. trichocarpa* and *A. sinensis* featured with the softwood (Figure 6a). Additionally, the expression level of genes related to the lignin metabolism was compared between stem and leaf tissue in *D. turbinatus* and *H. hainanensis*, respectively. Most of the genes associated with the lignin metabolism showed higher expression in stem tissue of both *D. turbinatus* and *H. hainanensis* than their respective leaf tissue, reflecting these expanded lignin-related genes might contribute to high lignin accumulation in stem tissues of both Dipterocarpaceae plants (Figure 6c). Thus, we anticipated that the expanded copy number of the *CESA*, *CLSE*, as well as the laccase and peroxidase in Dipterocarpaceae genomes might contribute to the evolutionary changes of the wood constitution (i.e. the ratio of cellulose microfibrils, hemicellulose and lignin molecule) and further result in the formation of harder, stronger and highly dense wood compared with those softwood plants.

#### Transcription factors and phytohormone contribute to the regulation of fragrant resins biosynthesis and wood formation in Dipterocarpaceae plants

The formation of fragrant resins and hardwood does not rely on a variety of biochemical properties of enzymes. Many transcription factors (TFs) and phytohormones also participate in these



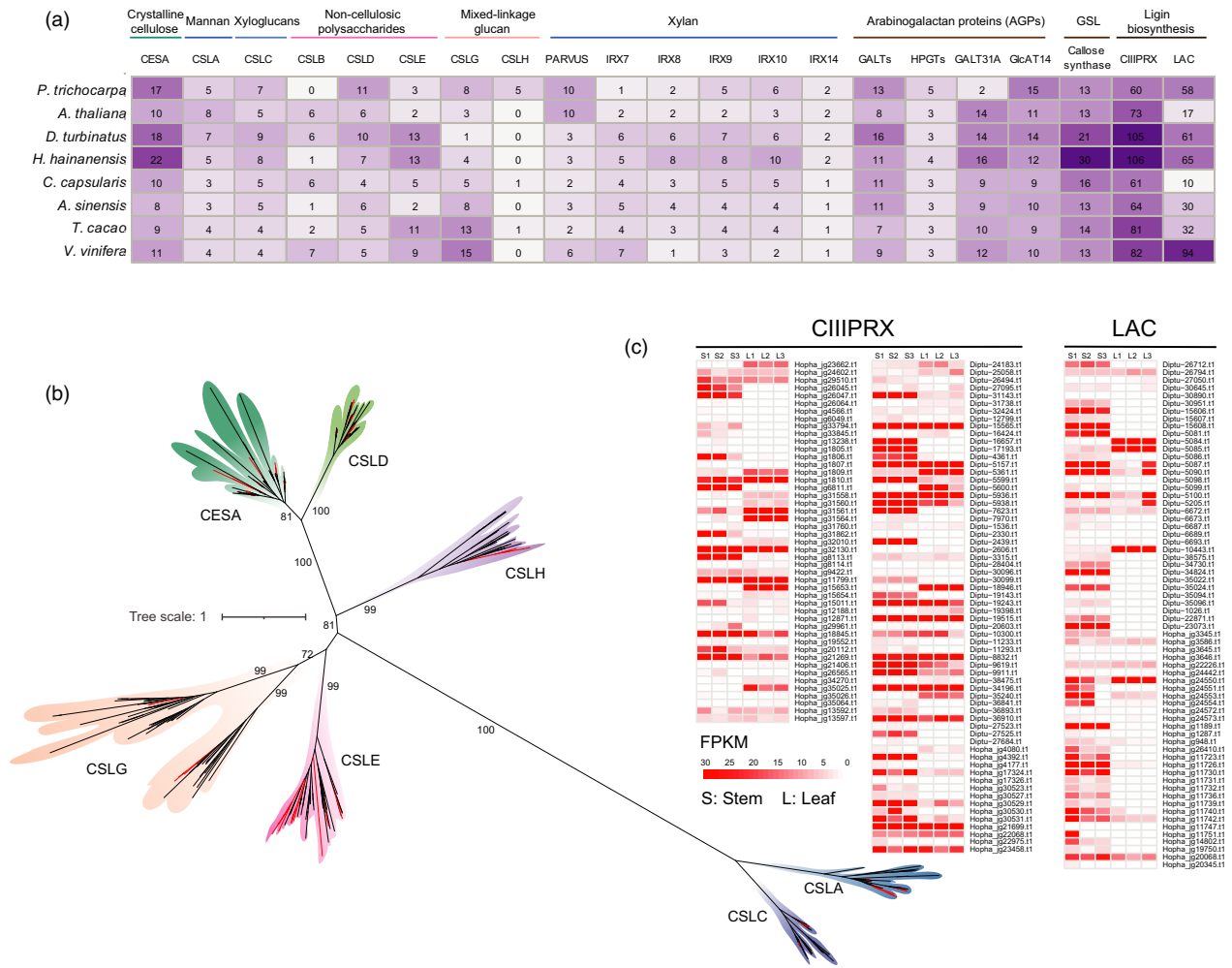
**Figure 5** Phylogeny and expression level of TPS and BAHD families in Dipterocarpaceae. (a) Maximum likelihood phylogenetic tree showing the classification and number of TPS in Dipterocarpaceae. (b) Chromosomal distribution of TPS, BAHD and BEAT genes of two Dipterocarpaceae. (c) Maximum likelihood phylogenetic tree showing the classification and number of BAHD subfamilies in Dipterocarpaceae. (d) Comparison of the relative expression profiles of TPS and BAHD genes between leaf and stem tissues of two Dipterocarpaceae species.

complicated regulatory networks (Plomion and Leprovost, 2001; Shang *et al.*, 2020a). A total of 2,401 and 2,397 TFs were respectively identified in *D. turbinatus* and *H. hainanensis*, which was remarkably higher than plants with relatively close relationships (Additional file 2: Table S15). To study the transcriptional regulatory networks of terpenoids biosynthesis and hardwood formation of Dipterocarpaceae wood, we performed a weighted correlation network analysis of transcript expression between TPSs and relative TFs as well as wood biosynthetic genes and relative TFs. In *D. turbinatus* stems, several *AP2/EREBP*, *bHLH* and *MYB* genes were shown to have a wide association in regulating fragrance-related and wood growth-related genes (Figure 7b). However, in the stem of *H. hainanensis*, the high number of *AP2/*

*EREBP* exhibited a strong co-relationship with both fragrance and wood growth-related genes, which was comparatively fewer by *MYB* and *bHLH* with respect to *D. turbinatus*.

Comparing the various phytohormones content between stem and leaf, we found auxin, ethylene and N6-( $\Delta^2$ -isopentenyl) adenine (a naturally occurring cytokinin) in *D. turbinatus* and *H. hainanensis* both showed significantly higher accumulation in stem tissue than its leaf tissue (Figure 7a and Additional file 2: Table S16). However, other types of phytohormones displayed different patterns between *D. turbinatus* and *H. hainanensis*. Salicylic acid (SA), jasmonic acid (JA) and abscisic acid (ABA) exhibited higher accumulation in the stem tissue of *H. hainanensis* than its leaf tissue, while almost no difference between leaf





**Figure 6** Evolution and expression of key genes involved in wood formation in Dipterocarpaceae. (a) The heat map shows a comparison of the numbers of key genes of related to cell wall formation and lignin metabolism among Dipterocarpaceae and representative plants. (b) Phylogenetic tree of the cellulose synthase (including cellulose synthase-like) genes. (c) Comparison of the relative expression profiles of the Laccase and peroxidase between leaf and stem tissues of two Dipterocarpaceae species.

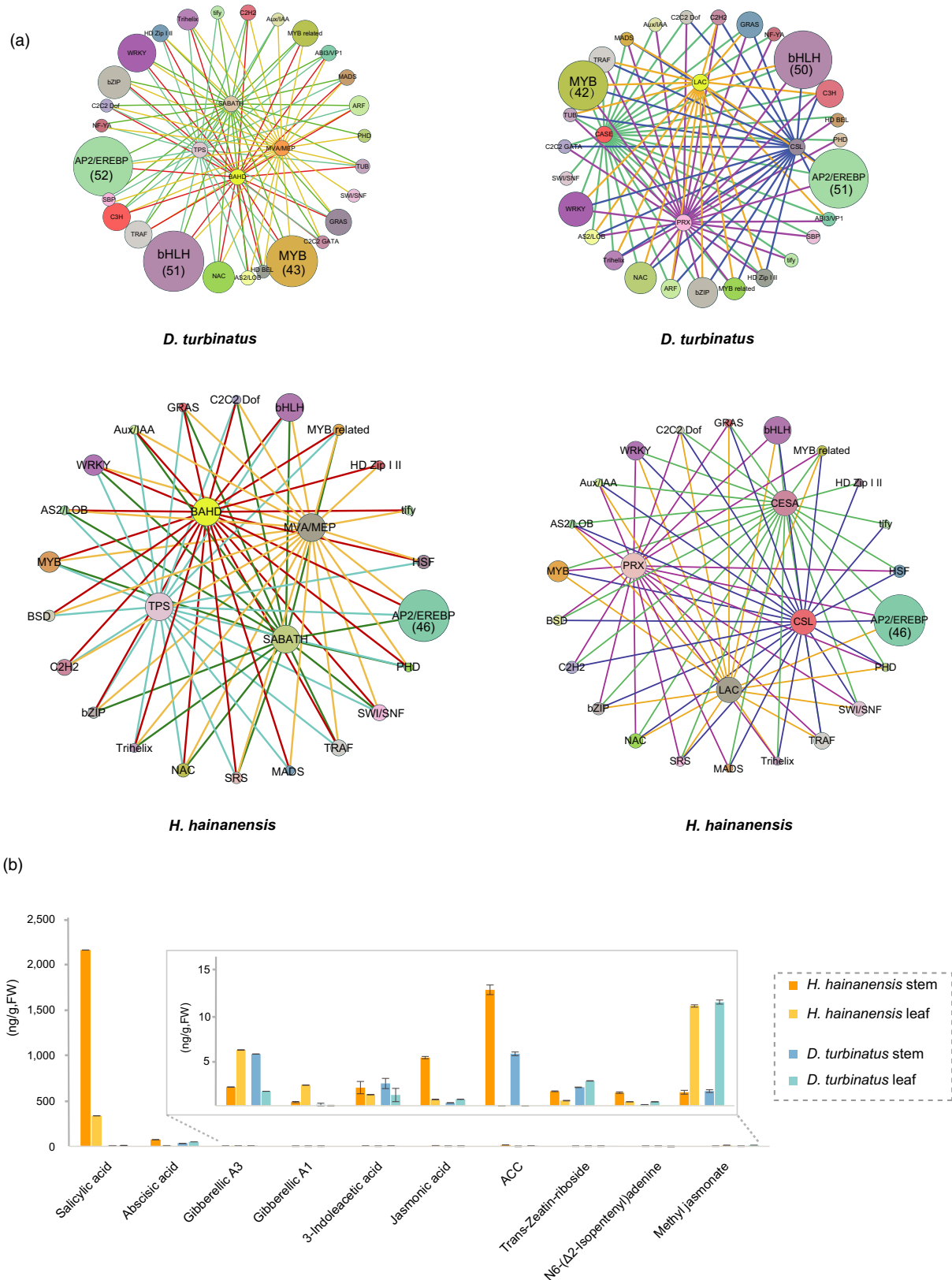
and stem tissues for *D. turbinatus*. Gibberellin accumulation in *D. turbinatus* stem tissue was over twofold higher than its leaf tissue; conversely, Gibberellin was preferred to be produced in *H. hainanensis* leaf rather than its stem (Figure 7a). Thus, these phytohormones might contribute to the regulation of the physiologies of stem tissue between *D. turbinatus* and *H. hainanensis* in different ways.

## Discussion

Dipterocarpaceae are widely known trees (dipterocarp forests) in the tropics and predominate the international tropical timber market, and therefore play an important role in the economy of many Southeast Asian countries (Appanah and Turnbull, 1998). Besides a high value for its hardwood timber, the non-timber species of Dipterocarpaceae are also widely used in the perfume industry and pharmaceutical applications (Appanah and Turnbull, 1998). In this study, we present the nuclear genome assemblies of *D. turbinatus* and *H. hainanensis* at chromosome level by combining the long-read sequences from Nanopore and Hi-C data for super-scaffolding. The first two Dipterocarpaceae

genomes presented here provided a valuable opportunity to determine genome evolutionary signatures and elucidate the genetic basis of these high-value metabolism products.

Phylogenomic analysis of a concatenation of 294 single-copy nuclear genes from 12 representative species (including 8 Malvales) and 226 single-copy nuclear genes from 14 representative species (including 13 Malvales) indicated that the two Dipterocarpaceae species *D. turbinatus* and *H. hainanensis* form a clade and are sister to Malvaceae, and a result is also supported by the coalescent analysis of nuclear genes. This relationship is consistent with previous studies from chloroplast genomes (Lee *et al.*, 2019), but is in conflict with a previous result (Hernandez-Gutierrez and Magallon, 2019), which recovered a sister relationship of Dipterocarpaceae and Thymelaeaceae on the basis of several discrete genes. Based on whole-genome sequences of the two representative plants of Dipterocarpaceae, our analyses revealed that the common ancestor of Dipterocarpaceae experienced a shared WGD event after the divergence from jute species. Polyploidization event is one of important evolutionary formations that have the ability to contribute to species divergence and adaptation (Madlung,



**Figure 7** The regulation of fragrance and timber formation in Dipterocarpaceae plants. (a) Quantitation of eight plant phytohormones amounts in leaf and stem of two Dipterocarpaceae trees. (b) Co-expression networks of fragrance formation-related genes and transcription factors in two Dipterocarpaceae species (left panel). Co-expression networks of cell wall metabolism and lignin-related genes and transcription factors in two Dipterocarpaceae species (right panel), the gene number of highlighted TF were shown in the in brackets.

2013; Scarpino and Levin, 2014); therefore, the occurrence of WGD event in the common ancestor of Dipterocarpaceae might have driven the species richness in Dipterocarpaceae. Dipterocarpus genus serves as the principal source to produce fragrant oleoresins from its wood, but other Dipterocarpaceae genera are with lesser production of fragrant oleoresins thus are less economically important (Appanah and Turnbull, 1998). The most famous oleoresin is procured from *D. turbinatus* wood, which is the principal source to produce perfumery over the world (Appanah and Turnbull, 1998). In consistent with previous studies (Aslam *et al.*, 2015; Wang *et al.*, 1992; Zhou and Ren, 2007), we also found several types of terpenoids in the wood of both *D. turbinatus* and *H. hainanensis* (Additional file 2: Table S8), which were the main fragrances contributor in Dipterocarpaceae. Gurjunene is a unique sesquiterpenes of Dipterocarpaceae which is responsible for the distinct fragrances of Dipterocarpaceae essential oils (Appanah and Turnbull, 1998). Interestingly, we found a few Dipterocarpaceae-specific *TPS-a* genes participating in sesquiterpene accumulation of fragrant resins in Dipterocarpaceae, especially for the *D. turbinatus*. However, only two Dipterocarpaceae-specific *TPS-a* genes were detected in its genome, indicating a diverse catalytic ability of sesquiterpene of *TPS-a* in *D. turbinatus*. The production of terpenes is mainly regulated by the transcription level of *TPS* genes as well as its upstream pathway (MVA/MEP) genes (Li, Wang, *et al.*, 2021). The results of the gene expression analyses revealed a dynamic expression of the *TPS* genes and MVA/MEP pathway-related genes, which may be another explanation for the terpene diversification in the two Dipterocarpaceae plants. In addition to the monoterpenes, sesquiterpenes, the remarkable expansion of LAMTs in both *D. turbinatus* and *H. hainanensis* suggesting iridoids might play an essential role in the formation of Dipterocarpaceae's peculiar scent (Consortium, 2018). Additionally, other fragrance contributors, such as benzenoid/phenylpropanoid biosynthesis-related genes, BAHD acyltransferases and SAM-binding methyltransferase (*COMT* and *ICMT*), most displayed higher activity in stem tissue of *D. turbinatus* and *H. hainanensis* than their respective leaf tissues, further suggesting complicated fragrance composition in Dipterocarpaceae wood. Using genomic data, we classified various types of duplication-induced gene family expansion. Functional enrichment of each type of duplication-induced gene family expansion revealed that whole-genome duplication and other types of duplication events have significant impacts on copy numbers of genes related to terpene, phenylpropanoid biosynthesis and lipid biosynthetic process (Additional file 2: Table S17) that are involved in aromatic oleoresin, which may contribute to the characteristic fragrance oleoresin in Dipterocarpaceae.

Wood is a mixture of polymers, partially composed of crystalline cellulose microfibrils and large amorphous hemicellulose and lignin molecules (Dhugga, 2012; Kumar and Turner, 2015). In wood, cellulose is one of the strongest polymers and hence is mainly responsible for the strength of the wood fibre (Rowell, 2012). Hemicelluloses are amorphous and thus easily hydrolysed into monomer sugars (Mota *et al.*, 2018). However, hemicelluloses are embedded and interact with cellulose and lignin, which significantly increase the strength and toughness of the plant cell wall (Berglund *et al.*, 2020). Generally, the packing density of the cell wall hemicelluloses is in a relatively greater proportion in hardwoods than in softwoods (Khatib, 2016). For tree plants, lignin could be distinguished into softwood lignin and hardwood lignin with different chemical

compositions (Huang *et al.*, 2012). Compared with representative softwood tree (*P. trichocarpa* and *A. sinensis*), Dipterocarpaceae own expanded CESA, CSLE, Laccase and peroxidase. Moreover, most of these genes exhibited higher expression in wood tissue than their expression status in leaves. CESA is involved in the primary cell wall formation and is thought to be the most important enzyme involved in the synthesis of cellulose microfibrils in plant cells (Kumar and Turner, 2015). Additionally, CSLE is proposed to be a Golgi-localized beta-glycan synthase that polymerizes the backbones of non-cellulosic polysaccharides (hemicelluloses) of the plant cell wall (Dhugga, 2012). Increased copy numbers of CESA and CSLE in both Dipterocarpaceae genomes may boost the efficacy of catalytic reactions via dosage effects, resulting in increased metabolic activity towards cellulose microfibrils and hemicelluloses. Laccase is necessary and non-redundant with peroxidase for lignin polymerization during vascular development in *Arabidopsis*. They are responsible for the catalysation of monolignols from corresponding *p*-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) lignin unit (Dixon and Barros, 2019; Wang *et al.*, 2015). The expansion of cell wall-related genes in Dipterocarpaceae genomes might contribute to the evolutionary changes of wood constitution, further inducing the formation of harder, stronger and denser wood.

Additionally, unique physiologies of Dipterocarpaceae wood may also necessitate complicated regulation by transcription factors and phytohormones. Ethylene, 3-Indoleacetic acid and N6-( $\Delta^2$ -isopentenyl) adenine showed higher accumulation in wood of both *D. turbinatus* and *H. hainanensis* compared with their leaves. All three phytohormones were reported to affect the emission of terpenoids and wood formation/growth in other plants (Boncan *et al.*, 2020; Savidge, 2003). Not all phytohormones showed a similar situation in both *D. turbinatus* and *H. hainanensis*. SA and JA displayed >sevenfold higher accumulation in *H. hainanensis* woods than its leaves, but they are not differentiated between woods and leaves of *D. turbinatus*. JA and SA both play important roles in transducing the activation of plant defence systems against pathogen attacks and responses to various abiotic stresses (McDowell and Dangl, 2000); therefore, the higher accumulation of SA and JA in the wood tissue of *H. hainanensis* might suggest its inherent defence ability against the pathogens. Notably, the remarkably expanded gene copy number of 'Protein ENHANCED PSEUDOMONAS SUSCEPTIBILITY (EPS)' in *H. hainanensis* genome compared with *D. turbinatus* (Additional file 2: Table S18) might also provide a clue to explain the higher SA level in *H. hainanensis*. Moreover, EPS encodes BAHD acyl transferase-like protein has been shown to trigger SA accumulation and SA-mediated resistance to virulent and avirulent pathogens in *A. thaliana* (Zheng *et al.*, 2009). Unexpectedly, we found that both Dipterocarpaceae genomes underwent massive loss of resistance gene families related to immunity, such as leucine-rich repeat receptor-like protein kinases (LRR-RLKs) and nucleotide-binding leucine-rich repeat (NBS-LRR) gene families which were markedly lower than that in all the other species investigated (Additional file 2: Table S19). These results might suggest Dipterocarpaceae mainly rely on its complex oleoresin terpene defences against herbivores and pathogens mediated by phytohormones such as JA and SA to compensate for the huge loss of immunity-related genes in their genomes (Celedon and Bohlmann, 2019; Nong *et al.*, 2020). Moreover, the key components of Dipterocarpaceae oleoresin, such as gurjunene and borneol,

have a slightly spicy smell, exhibiting higher activity in deterring aggressive insects (Li, Wang, *et al.*, 2021).

*D. turbinatus* woods displayed ~fourfold higher GA3 accumulation than its leaves. GA3-treated Cannabis exhibited an augmented HMGR activity of MVA pathway that is primarily essential for the synthesis of sesquiterpene (Mansouri and Asrar, 2009). GA3 could potentially have opposite effects on MEPs and MVA pathways, with stimulatory and inhibitory impacts on the key terpenoids produced via MVA and MEP pathways (Mansouri *et al.*, 2009). Consequently, we hypothesized that the high concentration of GA3 found in *D. turbinatus* woods would activate its MVA activity and inhibit the MEP pathway, resulting in increased sesquiterpene production but not increased monoterpene production. Overall, based on our findings, we hypothesize that gene expression dynamics, duplication-induced gene family expansion on aroma metabolism and cell wall associated genes may all contribute to the abundant feature in Dipterocarpaceae woods.

In summary, we presented high-quality assemblies of *D. turbinatus* and *H. hainanensis*, the first two reference genomes from the Dipterocarpaceae family. The integration of multi-omics data advanced our understanding of fragrant oleoresin biosynthesis and hardwood formation in Dipterocarpaceae plants. The two available complete Dipterocarpaceae genomes provided a fundamental resource for comparative genomic studies on the evolutionary mechanisms of secretion traits (fragrant oleoresin) and wood formation in these timber species at the genomic level, which will also be a valuable genetic resource for further research on the genome-assisted breeding and improvement, and conservation biology of Dipterocarpaceae.

## Methods

### Genome and transcriptome sequencing

All plant materials of *Dipterocarpus turbinatus* Gaertn. f. (HCNGB\_00001637) and *Hopea hainanensis* Merr. et Chun. (HCNGB\_00001636) used in this study were collected from Ruili Botanical Garden (Yunnan, China). High-quality DNA was extracted from fresh leaves by using QIAGEN® Genomic kits, and the DNA quantification was checked by Nanodrop and Qubit. PromethION Nanopore sequencer with the long-read DNA sequencing type was used for genome sequencing. The SQK\_LSK109 Ligation Sequencing Kit was used to prepare the sequencing libraries. Finally, a total of 27 and 26 Gb pass reads were generated for *D. turbinatus* and *H. hainanensis*, respectively.

For short read Illumina sequencing, the genomic DNA was isolated from fresh leaves using a modified CTAB protocol (Sahu and Thangaraj, 2012). The extracted DNA was used to create four paired-end libraries (170, 350, 500, and 800 bp) and four mate-pair libraries (2, 6, 10 and 20 Kb) using the Illumina standard methods (San Diego, CA). Following that, the sequencing was performed by employing a whole-genome shotgun sequencing approach on an Illumina HiSeq 2000 platform (San Diego).

### Estimation of the genome size

The genome size of *D. turbinatus* and *H. hainanensis* was estimated with Illumina sequencing short reads through kmer method by using kmerfreq (version 5.0) (Marçais and Kingsford, 2011). From the kmer frequency distribution, the kmer depth was 28 and 35, and the total kmer number was 11 792 799 620 and 15 200 711 260, respectively. The genome size was estimated by the formula: genome size =  $K_{num} / k\text{-mer}_{depth}$ .

### Genome assembly and annotation

NextDenovo (<https://github.com/Nextomics/NextDenovo/>) used Oxford Nanopore long reads to assemble the genome, including reads error correction with parameter 'task=all (correct and assemble)'. We set read\_cutoff = 1k, seed\_cutfiles = 10K while default parameters were used for other settings. At the genome polishing stage, NextPolish was used to correct the genome with three rounds of nanopore reads and thrice with Illumina sequencing reads.

There are three methods to evaluate the quality of the genome assembly. Firstly, the assembly N50 was more than 29Mb and 9Mb, respectively. The completeness of genome assembly was evaluated by BUSCO (version 2) with 'eukaryota\_odb9' database (Additional file 2: Table S4) (Waterhouse *et al.*, 2018). The accuracy used genome mapping rate to Illumina short reads by STAR (version 2.40) (Dobin *et al.*, 2013).

Repeat elements were annotated using a combined strategy. We used both de novo and homolog-based methods to find DNA transposon elements, retrotransposon elements and tandem repeats. For *ab initio* prediction, we used Piler-DF, RepeatScout, MITE-hunter, LTR\_FINDER and RepeatModeler (version 1.0.8; <http://www.repeatmasker.org/RepeatModeler/>). Among them, Piler detected repeat elements such as satellites and transposons, RepeatScout identified all repeat classes, MITE-hunter discovered miniature inverted-repeat transposable elements (MITEs) from the genomic sequence, while LTR-FINDER predicted the location and structure of full-length LTR retrotransposons. All results from *ab initio* prediction were merged as a homolog database to identify repetitive sequences by RepeatMasker (<http://www.repeatmasker.org>). We used LAI (LTR Assembly Index) to evaluate the assembly continuity by evaluating the assembly of repeat sequences. LTR-RT candidates were obtained using LTRharvest with parameters '-minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1 -similar 85 -vic 10 -seed 20 -seqids yes' and LTR\_Finder (version 1.0.7) with '-D 20000 -d 1000 -L 700 -l 100 -p 20 -C -M 0.9' (Ellinghaus *et al.*, 2008). LTR\_retriever was used to filter, unique and then obtain high-confidence LTR retrotransposons with default parameters. Then, the genome LAI score was carried out by LAU program in the LTR\_retriever with default parameter.

Gene models come from homology-based prediction, de novo prediction and RNA-seq-based prediction. We used automated BRAKER2 to obtain accurate gene models which combined de novo and homology-based predictions with GeneMark-ES/ET and AUGUSTUS (Brüna *et al.*, 2021). For training GeneMark-TP and AUGUSTUS, we selected all Malvales proteins from the NR database (non-redundant protein database). All protein-coding genes were against several databases, including NR (plant database), SwissProt, KEGG (plant database), COG, InterProScan (using data from Pfam, PRINTS, SMART, ProDom and PROSITE) and GO by blastp (E-value < 1e-5).

### Pseudochromosome assembly based on Hi-C data

Pseudochromosome validation involved three steps. First, Hi-C-Pro was used to process the Hi-C data from paired-end raw reads to normalized contact maps with a resolution of 100 kb (Servant *et al.*, 2015). The raw data with low quality, unmapped and invalid mapped paired reads were filtered out, and then, the assembly genome was integrated into a pseudochromosome-scale assembly using the 3D de novo assembly (3D DNA) pipeline (Dudchenko *et al.*, 2017). Juicebox Assembly Tools were used to

view '.hic' files from 3D DNA and further improve assembly by hand according to the contact maps (Durand *et al.*, 2016).

### Gene family, phylogenomic analysis and estimation of divergence times

The genomes of species that were used for comparative genomics analysis were downloaded from public databases. OrthoFinder (v.1.1.8) was used to infer a homolog matrix of orthogroups (gene families) among these selected organisms (Emms and Kelly, 2019). Single copy gene families were used to construct phylogenetic trees based on maximum likelihood. In brief, multiple sequence alignment by MAFFT (v.7.310) for each single-copy gene orthogroup, followed by gap position removal (only positions where 50% or more of the sequences have a gap are treated as a gap position). A maximum likelihood phylogenetic tree was constructed for each single-copy gene family. The ASTRAL program was used to combine all single-copy gene trees to a species tree with the multispecies coalescent model (Zhang *et al.*, 2018). The Count software was implemented (with wagner parsimony algorithm) to analyse the orthogroups changes (such as gains, loss, expansions and contractions) of each lineage at every evolutionary node of the phylogenetic tree (Csűös, 2010). Divergence times between species were calculated using the MCMC tree program (<http://abacus.gene.ucl.ac.uk/software/paml.html>) implemented in Phylogenetic Analysis by Maximum Likelihood (PAML) (Yang and evolution, 2007). Expansion and contraction of the orthologous gene families were determined using CAFÉ software (De Bie *et al.*, 2006).

### Analysis of genome synteny and whole-genome duplication

We use the MCscan pipeline ([https://github.com/tanghaibao/jcvi/wiki/MCscan-\(Python-version\)](https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version))) and Circos (Krzywinski *et al.*, 2009) for genome synteny. Ancient whole-genome duplications were generated by command-line tool WGD (Zwaenepoel and Van de Peer, 2019). Then, common evidence for ancient WGDs and synonymous substitutions per synonymous site (Ks) distributions were computed including whole-paranome and one-vs-one ortholog in *D. turbinatus* and *H. hainanensis* and other related genomes (*G. raimondii*, *A. sinensis*, *C. capsularis*, *T. cacao* and *A. thaliana*).

To provide additional evidence of shared WGD event between the *D. turbinatus* and *H. hainanensis*, we extracted paralogous pairs of *D. turbinatus* and *H. hainanensis* genes derived from their respective WGDs and constructed phylogenetic trees. Firstly, we performed gene family cluster by using proteomes of *A. sinensis*, *A. thaliana*, *C. capsularis*, *D. turbinatus*, *H. hainanensis* and *T. cacao* to obtain orthogroups. We identified 4,539 and 3,098 gene pairs from *D. turbinatus* and *H. hainanensis* from Ks peak ( $Ks=0.22 \sim 0.35$ ), respectively. 4,539 gene pairs of *D. turbinatus* were distributed in 3,578 orthogroups, and 3,098 gene pairs of *H. hainanensis* were distributed in 2,550 orthogroups. Then, we identified a total of 1,631 orthogroups containing both gene pairs of *D. turbinatus* and *H. hainanensis*. Next, 125 out of 1,631 orthogroups were randomly selected for phylogenetic study by using the IQtree (version 1.6.12) with parameter of '-bb 10000 -alrt 5000 -nt AUTO'. Next, we reconciled ancestral gene events (duplications, losses and transfers) using a phylogenomic function in NOTUNG (version 2.9) by comparing the gene trees with the species tree. These trees were visualized with NOTUNG (Chen and Durand, 2000; Stolzer *et al.*, 2012). Finally, we also checked the synteny of gene-derived scaffold region between *D. turbinatus*

and *H. hainanensis* in every phylogenetic tree. In summary, 108, 4, 5 out of 125 trees supported type I, II and III topologies, respectively, and a good synteny of gene pair-derived scaffold region could be observed between *D. turbinatus* and *H. hainanensis* in each tree, thereby providing strong evidence of shared WGD event between both the species (Additional file 1: Figure S5).

To investigate *D. turbinatus* and *H. hainanensis* genome evolution, we do further genome-wide duplications identification and classification by DupGen\_finder with default parameters. We identified different modes of gene-duplicated gene pairs and divided them into five types duplications: whole-genome duplicates (WGD), tandem duplicates (TD), proximal duplicates (less than 10 gene distance on the same chromosome: PD), transposed duplicates (transposed gene duplications: TRD) and dispersed duplicates (other duplicates than WGD, TD, PD and TRD: DSD). The target species was *A. thaliana*, and the final gene number came from unique genes.

### Estimation of the divergence time

Divergence times in the phylogeny tree between each species were calculated using the MCMC tree program with -sampfreq 5000 -burnin 5000000 parameter. The sequential PHYLIP format nucleotide sequences and rooted phylogeny tree is derived from Figure 2. The divergence time was searched from TIMETREE (<http://www.timetree.org/>), *G. raimondii*-*D. zibethinus* divergence time (60-77 MYA), *G. raimondii*-*T. cacao* divergence time (62-85 MYA).

### Detection of key candidate functional genes

Based on the following criteria, all candidate genes were screened: firstly, candidate gene sequences were identical to collect query gene sequences gathered from previous studies or public databases, by BLAST ( $< 1e-5$ ); and (2) The candidate genes feature should be similar with the online functional annotation or Swissprot functional annotation query genes.

Regarding the identification of transcription factors (TFs), we used the HMMER search method for transcription factors. The Pfam website (<https://pfam.xfam.org/>) was used to download HMMER domain structure models for each transcription factor when as per the role of TAPscan v.2 database for TFs (<https://plantcode.online.uni-marburg.de/tapscan/>). Preliminary TF candidate genes were collected for each species ( $< 1e-5$ ) by searching the HMM profile. Then, parts of genes were filtered if they are not the homologs according to their functional annotation of SwissProt ( $< 1e-5$ ). In the end, we filtered genes that contained a wrong domain under the TAPscan v.2 transcription factor database domain rules.

To identify genes involved in the terpenoid backbone biosynthesis pathway (Figure 4a), we collected the genes from *A. thaliana* that were documented in this pathway. Using these genes as a query sequence in BlastP, we predicted TPS genes with queries from *Atha*, *Vvin*, *Ptri* and rice, and the two Pfam domains, PF01397 and PF03936, were used to search by using HMMER. For BAHD identification, *Atha* members were used as queries to predict *A. sinensis*, *C. capsularis*, *D. turbinatus*, *H. hainanensis*, *T. cacao* and *V. vinifera* BAHD genes using BLASTP ( $1e-5$ ). The CYP450 genes were searched by both domains, PF00067 and BLAST, with the queries from rice and *Atha*.

Finally, all searched candidate genes were used for phylogenetic analysis to distinguish the orthologs of corresponding functionally characterized genes. For each gene family, mafft-



7.310 used to align and Gblocks used to trim ambiguously aligned positions (Castresana, 2000). All the phylogenetic trees of functional genes were constructed by the maximum likelihood method with RAxML-8.2.4 (Stamatakis, 2014).

### Transcriptome analysis of different tissues

The raw paired-end RNA-seq reads were filtered into clean data by FASTP (Chen *et al.*, 2018). RSEM (<https://deweylab.github.io/RSEM/>) packages were used to estimate gene expression levels from clean reads. The transcriptome reads were mapped to the assembled genome by bowtie2 with default settings. We continue to identify the differentially expressed genes by DESeq2 (Love and Huber, 2014).

### Tissue-specific Co-expression Modules

To explore the dynamic changes of the genes and programs expressed, we performed weighted correlation network analysis (WGCNA) of gene expression in the leaf and stem (FPKM>1) (Langfelder and Horvath, 2008). When module displayed gene highly expressed in all 3 stem tissues and low expressed in all 3 leaf tissues, gene from this kind of module was selected for further co-expression networks.

### Detection of metabolites by LC-MS

The stem and leaf tissues were collected and stored in liquid nitrogen, then transferred to a freezer at  $-80^{\circ}\text{C}$ . For the terpenoids detection, stem and leaf samples were preliminarily disposed of by using 2-chlorophenylalanine (4 ppm) methanol. Next, samples with glass beads were put into the tissue grinder to grind for 90 s at 55 Hz. Following centrifugation at 12000 rpm at  $4^{\circ}\text{C}$  for 10 min, take the supernatant, filter it through  $0.22\ \mu\text{m}$  membrane and transfer the filtrate into the detection bottle before LC-MS analysis. Then, the sample extracts were analysed using an Ultra Performance Liquid Chromatography (UPLC) Vanquish (Thermo) and Q Exactive HF-X system (Thermo). For the quantitative detection of phytohormones, stem and leaf tissue samples were used. The self-construction database which is constructed by reference standards was used to perform qualitative analysis. Additionally, different concentrations of the standards were used to perform quantitative analysis.

### Acknowledgements

This work was supported by National Key R&D Program of China (No. 2019YFC1711000), Major Science and Technology Projects of Yunnan Province (Digitalization, development and application of biotic resource, 202002AA100007), and the Shenzhen Municipal Government of China (No. JCYJ20151015162041454). This work is part of the 10KP project (<https://db.cngb.org/10kp/>) (Cheng *et al.*, 2018). This work is also supported by China National GeneBank (CNGB; <https://www.cngb.org/>).

### Conflicts of interest

The authors declare that they have no competing interests.

### Authors' contributions

H.L. and S.K.S. conceived, designed and supervised the project. J.W., L.C., J.Y., C.H., J.L. and Y.Z. provided resources and materials. S.W., H. Liang., S.K.S. and H.W. analysed the data.

S.K.S., S.W. and H. Liang. wrote the paper. H.L., Y.L., T.M., M.L., L.L., Y.X. and M.L. revised the manuscript. All the authors read, revised and approved the final version of the manuscript.

### Data availability statement

The data sets generated and analysed during the current study are available in the CNGB Nucleotide Sequence Archive (CNSA: <https://db.cngb.org/cnsa/>) under accession number CNP0002018.

### References

- Ankarfjård, R. and Kegl, M. (1998) Tapping oleoresin from *Dipterocarpus alatus* (Dipterocarpaceae) in a Lao village. *Econ. Bot.* **52**, 7–14.
- Appanah, S. and Turnbull, J.M. eds. (1998) A review of dipterocarps: taxonomy, ecology, and silviculture. In *Center for International Forestry Research (CIFOR)*. <https://doi.org/10.17528/cifor/000463>
- Aslam, M.S., Ahmad, M.S. and Mamat, A. (2015) A phytochemical, ethnomedicinal and pharmacological review of genus *Dipterocarpus*. *Int. J. Pharma. Pharma. Sci.* **7**, 27–38.
- Baek, S., Choi, K., Kim, G.-B., Yu, H.J., Cho, A., Jang, H., Kim, C. *et al.* (2018) Draft genome sequence of wild *Prunus yedoensis* reveals massive inter-specific hybridization between sympatric flowering cherries. *Genome Biol.* **19** (1), 1–17.
- Berglund, J., Mikkelsen, D., Flanagan, B.M., Dhital, S., Gaunitz, S., Henriksson, G., Lindström, M.E. *et al.* (2020) Wood hemicelluloses exert distinct biomechanical contributions to cellulose fibrillar networks. *Nat. Commun.* **11**, 1–16.
- Boncan, D.A.T., Tsang, S.S., Li, C., Lee, I.H., Lam, H.-M., Chan, T.-F. and Hui, J.H. (2020) Terpenes and terpenoids in plants: interactions with environment and insects. *Int. J. Mol. Sci.* **21**, 7382.
- Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M. and Borodovsky, M. (2021) BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinform.* **3**, lqaa108.
- Butkhup, L., Jeenphakdee, M., Jorjong, S., Samappito, S., Samappito, W. and Chotivannakul, S. (2011) HS-SPME-GC-MS analysis of volatile aromatic compounds in alcohol related beverages made with mulberry fruits. *Food Sci. Biotechnol.* **20**, 1021–1032.
- Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552.
- Celedon, J.M. and Bohlmann, J. (2019) Oleoresin defenses in conifers: chemical diversity, terpene synthases and limitations of oleoresin defense under climate change. *New Phytol.* **224**, 1444–1463.
- Chang, Y., Liu, H., Liu, M., Liao, X., Sahu, S.K., Fu, Y., Song, B. *et al.* (2018) The draft genomes of five agriculturally important African orphan crops. *GigaScience*, **8**(3), gij152.
- Chen, K., Durand, D. and Farach-Colton, M. (2000) NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.* **7**, 429–447.
- Chen, S., Zhou, Y., Chen, Y. and Gu, J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
- Cheng, S., Melkonian, M., Smith, S.A., Brockington, S., Archibald, J.M., Delaux, P.-M., Li, F.-W. *et al.* (2018) 10KP: A phylodiverse genome sequencing plan. *GigaScience*, **7**(3), gij013.
- Christenhusz, M.J. and Byng, J.W. (2016) The number of known plants species in the world and its annual increase. *Phytotaxa*, **261**, 201–217.
- Consortium and Boachon, B., Buell, C.R., Crisovan, E., Dudareva, N., Garcia, N., Godden, G., Henry, L. *et al.* (2018) Phylogenomic mining of the mints reveals multiple mechanisms contributing to the evolution of chemical diversity in Lamiaceae. *Mol. Plant*, **11**, 1084–1096.
- Csűös, M. (2010) Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, **26**, 1910–1912.
- Cvetković, T., Hinsinger, D.D. and Strijk, J.S. (2017) The first complete chloroplast sequence of a major tropical timber tree in the Meranti family: *Vatica odorata* (Dipterocarpaceae). *Mitochondrial DNA Part B*, **2**, 52–53.

- 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.
- Dhugga, K.S. (2012) Biosynthesis of non-cellulosic polysaccharides of plant cell walls. *Phytochemistry*, **74**, 8–19.
- Dixon, R.A. and Barros, J. (2019) Lignin biosynthesis: old roads revisited and new roads explored. *Open Biol.* **9**, 190215.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P. et al. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
- Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S. et al. (2017) De novo assembly of the aedes aegypti genome using Hi-C Yields chromosome-length Scaffolds. *Science*, **356**, 92–95.
- Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S. and Aiden, E.L. (2016) Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Sys.* **3**, 99–101.
- Dyrmoose, A.-M.-H., Turreira-García, N., Theilade, I. and Meilby, H. (2017) Economic importance of oleoresin (*Dipterocarpus alatus*) to forest-adjacent households in Cambodia. *Nat. His. Bull. Siam Soc.* **62**(1), 67–84.
- Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.* **9**, 1–14.
- Emms, D.M. and Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14.
- Fan, Y., Sahu, S.K., Yang, T., Mu, W., Wei, J., Cheng, L., Yang, J. et al. (2020) Dissecting the genome of star fruit (*Averrhoa carambola* L.). *Hortic Res.* **7**, 1–10.
- Fan, Y., Sahu, S.K., Yang, T., Mu, W., Wei, J., Cheng, L.E., Yang, J. et al. (2021) The *Clausena lansium* (Wampee) genome reveal new insights into the carbazole alkaloids biosynthesis pathway. *Genomics*, **113**, 3696–3704.
- Hernandez-Gutierrez, R. and Magallon, S. (2019) The timing of Malvales evolution: Incorporating its extensive fossil record to inform about lineage diversification. *Mol. Phylogenet. Evol.* **140**, 106606.
- Hofmeister, B.T., Denkena, J., Colome-Tatche, M., Shahryar, Y., Hazarika, R., Grimwood, J., Mamidi, S. et al. (2020) A genome assembly and the somatic genetic and epigenetic mutation rate in a wild long-lived perennial *Populus trichocarpa*. *Genome Biol.* **21**, 259.
- Huang, Y., Wang, L., Chao, Y., Nawawi, D.S., Akiyama, T., Yokoyama, T. and Matsumoto, Y. (2012) Analysis of lignin aromatic structure in wood based on the IR spectrum. *J. Wood Chem. Technol.* **32**, 294–303.
- Kamariyah, A., Ozek, T., Demirci, B. and Baser, K. (2012) Chemical composition of leaf and seed oils of *Dryobalanops aromatica* Gaertn. (Dipterocarpaceae). *ASEAN J. Sci. Technol. Dev.* **29**, 105–114.
- Khatib, J. (2016) *Sustainability of Construction Materials*. Cambridge: Woodhead Publishing.
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome research*, **19**, 1639–1645.
- Kumar, M. and Turner, S. (2015) Plant cellulose synthesis: CESA proteins crossing kingdoms. *Phytochemistry*, **112**, 91–99.
- Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 1–13.
- Lee, S.Y., Ng, W.L., Hishamuddin, M.S. and Mohamed, R. (2019) The complete chloroplast genome sequence of Chengal (*Neobalanocarpus heimii*, Dipterocarpaceae), a durable tropical hardwood. *Mitochondrial DNA Part B*, **4**, 19–20.
- Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W., Grosse, I., Li, Z., Melkonian, M. & Mirarab, S. (2019) One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685.
- Li, J., Wang, Y., Dong, Y., Zhang, W., Wang, D., Bai, H., Li, K. et al. (2021) The chromosome-based lavender genome provides new insights into Lamiaceae evolution and terpenoid biosynthesis. *Horticulture Res.* **8**, 1–14.
- Li, W., Chen, H.-Q., Wang, H., Mei, W.-L. and Dai, H.-F. (2021) Natural products in agarwood and *Aquilaria* plants: Chemistry, biological activities and biosynthesis. *Nat. Prod. Rep.* **38**, 528–565.
- Liu, Y., Zhang, X., Han, K., Li, R., Xu, G., Han, Y., Cui, F. et al. (2021) Insights into amphicarpary from the compact genome of the legume *Amphicarpaea edgeworthii*. *Plant Biotechnol. J.* **19**, 952–965.
- 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**, 1–21.
- Madlung, A. (2013) Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity*, **110**, 99–104.
- Mansouri, H., Asrar, Z. and Mehrabani, M. (2009) Effects of gibberellic acid on primary terpenoids and  $\Delta^9$ -tetrahydrocannabinol in *Cannabis sativa* at flowering stage. *J. Integr. Plant Biol.* **51**, 553–561.
- Marçais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.
- McDowell, J.M. and Dangl, J.L. (2000) Signal transduction in the plant immune response. *Trends Biochem. Sci.* **25**, 79–82.
- Meng, L.-Z. and Xu, Z.-F. (2005) The threatened status and ex situ community conservation approach on Dipterocarpaceae in China. *Guihaia*, **1**(1), 10–15.
- Messer, A., McCormick, K., Hagedorn, H.H., Tumbel, F. and Meinwald, J. (1990) Defensive role of tropical tree resins: antitermitic sesquiterpenes from Southeast Asian Dipterocarpaceae. *J. Chem. Ecol.* **16**, 3333–3352.
- Mota, T.R., Oliveira, D., Marchiosi, R., Ferrarese-Filho, O. and Santos, W. (2018) Plant cell wall composition and enzymatic deconstruction. *AIMS Bioengineering*, **5**, 63–77.
- Nong, W., Law, S.T., Wong, A.Y., Baril, T., Swale, T., Chu, L.M., Hayward, A. et al. (2020) Chromosomal-level reference genome of the incense tree *Aquilaria sinensis*. *Mol. Ecol. Resour.* **20**, 971–979.
- Oldfield, S., Lusty, C. and MacKinven, A. (1998) *The World List of Threatened Trees*. Marseille: World Conservation Press. <https://www.unep.org/resources/report/world-list-threatened-trees>
- Ou, S., Chen, J. and Jiang, N. (2018) Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126.
- Peter, K. and Babu, K.N. (2012) Introduction to herbs and spices: medicinal uses and sustainable production. In *Handbook of Herbs and Spices*, (Vol. **2**, pp. 1–16). Cambridge: Elsevier. <https://doi.org/10.1533/9780857095688.1>
- Plomion, C., Leprovost, G. and Stokes, A. (2001) Wood formation in trees. *Plant Physiol.* **127**, 1513–1523.
- 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**, 1–23.
- Rana, R., Langenfeld-Heyser, R., Finkeldey, R. and Polle, A. (2009) Functional anatomy of five endangered tropical timber wood species of the family Dipterocarpaceae. *Trees*, **23**, 521–529.
- Rana, R., Langenfeld-Heyser, R., Finkeldey, R. and Polle, A. (2010) FTIR spectroscopy, chemical and histochemical characterisation of wood and lignin of five tropical timber wood species of the family of Dipterocarpaceae. *Wood Sci. Technol.* **44**, 225–242.
- Rowell, R.M. (2012) *Handbook of Wood Chemistry and Wood Composites*. Florida: CRC press.
- Sahu, S.K., Liu, M., Yssel, A., Kariba, R., Muthemba, S., Jiang, S., Song, B. et al. (2019) Draft Genomes of Two Artocarpus Plants, Jackfruit (*A. heterophyllus*) and Breadfruit (*A. altiss.*). *Genes*, **11**, 1–17.
- Sahu, S.K., Thangaraj, M. and Kathiresan, K. (2012) DNA extraction protocol for plants with high levels of secondary metabolites and polysaccharides without using liquid nitrogen and phenol. *ISRN Mol. Biol.* **2012**, 205049.
- Sasaki, S. (2006) Ecology and physiology of Dipterocarpaceae. In *Plantation Technology in Tropical Forest Science*, (Suzuki, K., Ishii, K., Sakurai, S. and Sasaki, S., eds), pp. 3–22. Tokyo: Springer. [https://doi.org/10.1007/4-431-28054-5\\_1](https://doi.org/10.1007/4-431-28054-5_1)
- Savidge, R.A. (2003) Tree growth and wood quality. *Wood Quality Its Biol. Basis* **1**(1), 1–29.
- Scarpino, S.V., Levin, D.A. and Meyers, L.A. (2014) Polyploid formation shapes flowering plant diversity. *Am. Nat.* **184**, 456–465.
- 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**, 1–11.
- Shang, J., Tian, J., Cheng, H., Yan, Q., Li, L., Jamal, A., Xu, Z. 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.
- Shang, J., Tian, J., Cheng, H., Yan, Q., Li, L., Jamal, A., Xu, Z. *et al.* (2020b) The chromosome-level wintersweet (*Chimonanthus praecox*) genome provides insights into floral scent biosynthesis and flowering in winter. *Genome Biol.* **21**, 200.
- Shen, J., Zhou, Q., Li, P., Wang, Z., Liu, S., He, C., Zhang, C. *et al.* (2017) Update on phytochemistry and pharmacology of naturally occurring resveratrol oligomers. *Molecules*, **22**, 2050.
- Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
- Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B. and Durand, D. (2012) Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, **28**, i409–i415.
- Wang, J.-L., Ding, J.-K., Chen, Z.-Y. and Yang, C.-R. (1992) The sesquiterpenes and their seasonal variations in the oleoresin essential oils from two species of *Dipterocarpus* in Yunnan. *Acta Botanica Yunnanica* **14**(3), 1–3.
- Wang, J., Feng, J., Jia, W., Chang, S., Li, S. and Li, Y. (2015) Lignin engineering through laccase modification: a promising field for energy plant improvement. *Biotechnol. Biofuels*, **8**, 1–11.
- Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V. *et al.* (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548.
- Xu S., Ding Y., Sun J., Zhang Z., Wu Z., Yang T., Shen F. *et al.* (2021) A high-quality genome assembly of *Jasminum sambac* provides insight into floral trait formation and Oleaceae genome evolution. *Molecular Ecology Resources*. <http://dx.doi.org/10.1111/1755-0998.13497>
- 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.
- Yan, F., Tao, X., Wang, Q.-L., Juan, Z.-Y., Zhang, C.-M. and Yu, H.-L. (2019) The complete chloroplast genome sequence of the medicinal shrub *Daphne giraldii* Nitsche. (Thymelaeaceae). *Mitochondrial DNA Part B*, **4**, 2685–2686.
- Yang, X., Yue, Y., Li, H., Ding, W., Chen, G., Shi, T., Chen, J. *et al.* (2018) The chromosome-level quality genome provides insights into the evolution of the biosynthesis genes for aroma compounds of *Osmanthus fragrans*. *Horticulture Res.* **5**, 1–13.
- Yang Z.J.M.b. and evolution (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.
- Yongram, C., Sungthong, B., Puthongking, P. and Weerapreeyakul, N. (2019) Chemical composition, antioxidant and cytotoxicity activities of leaves, bark, twigs and oleo-resin of *Dipterocarpus alatus*. *Molecules*, **24**, 3083.
- Zhang, C., Rabiee, M., Sayyari, E. and Mirarab, S. (2018) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinform.* **19**, 15–30.
- Zhang, L., Liu, M., Long, H., Dong, W., Pasha, A., Esteban, E., Li, W. *et al.* (2019) Tung tree (*Vernicia fordii*) genome provides a resource for understanding genome evolution and improved oil production. *Genom. Proteomics Bioinform.* **17**, 558–575.
- Zhao, K., Yang, W., Zhou, Y., Zhang, J., Li, Y., Ahmad, S. and Zhang, Q. (2017) Comparative transcriptome reveals benzenoid biosynthesis regulation as inducer of floral scent in the woody plant *Prunus mume*. *Front. Plant Sci.* **8**, 319.
- Zheng, Z., Qualley, A., Fan, B., Dudareva, N. and Chen, Z. (2009) An important role of a BAHD acyl transferase-like protein in plant innate immunity. *Plant J.* **57**, 1040–1053.
- Zhou, B. and Ren, H. (2007) Study on the essential oil of *Dipterocarpus tubinatus* wood. *Flavour Fragrance Cosmetics* **5**(1), 9–11.
- Zhou, L. and Yang, M. (2017) Research and utilization of Dipterocarpaceae plants in China. *World Forestry Res.* **30**, 46–51.
- 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.

## Supporting information

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

**Figure S1** Estimation of genome complexity of two Dipterocarpaceae trees.

**Figure S2** Hi-C contact matrix visualization for chromosomes of two Dipterocarpaceae reference genome assemblies.

**Figure S3** Comparison of insertion dates of LTR-RTs among *D. turbinatus*, *H. hainanensis* and *A. thaliana*.

**Figure S4** The phylogenetic analyses of Malvales based on nuclear and chloroplast genes.

**Figure S5** Phylogenetic evidence of shared WGD event between the *D. turbinatus* and *H. hainanensis*.

**Figure S6** Syntenic blocks between genomes.

**Figure S7** Phylogenetic tree of the TPS proteins.

**Figure S8** Chromosomal distribution of the important fragrance related genes in *D. turbinatus* and *H. hainanensis* genomes, respectively.

**Figure S9** Phylogenetic tree of the SABATH family, COMT and ICMT in the selected plants.

**Table S1** Statistics of contig and chromosome level genome assemblies of *D. turbinatus* and *H. hainanensis*.

**Table S2** Benchmarking Universal Single-Copy Orthologs (BUSCO) evaluation of *D. turbinatus* and *H. hainanensis* genomes.

**Table S3** Annotation of repetitive sequences in *D. turbinatus* and *H. hainanensis* genomes.

**Table S4** Functional annotation of the *D. turbinatus* and *H. hainanensis* protein-coding genes.

**Table S5** KEGG enrichment of expanded gene families of selected evolutionary nodes.

**Table S6** KEGG enrichment of unique gene families in Dipterocarpaceae.

**Table S7** Statistics of expanded genes number induced by different types of gene duplication way.

**Table S8** Aroma components in Dipterocarpaceae.

**Table S9** Genes involved in the 2-C-methyl-D-erythritol 4-phosphate (MVA) and mevalonate (MEP) metabolism in selected plants.

**Table S10** Comparison of TPS gene number in selected plants.

**Table S11** Comparison of BAHD gene number in selected plants.

**Table S12** The candidate genes involved in the benzenoid/phenylpropanoid synthesis pathways in *D. turbinatus* and *H. hainanensis*.

**Table S13** The expression level of candidate genes involved in the terpenoid volatile organic compounds (VOCs) synthesis in wood and leaf tissue of *D.*

**Table S14** Comparison of gene number of SABATH family and SAM-binding methyltransferase superfamily in selected plants.

**Table S15** Comparison of transcription factors of selected plants.

**Table S16** Comparison of various phytohormones content between wood and leaf.

**Table S17** Genes involved in the fatty acid metabolism.

**Table S18** Protein ENHANCED PSEUDOMONAS SUSCEPTIBILITY in selected plants.

**Table S19** Statistics of resistance (R) genes.