Plant Communications
Research Article

CellPress
Partner Journal

Thirteen Dipterocarpoideae genomes provide insights into their evolution and borneol biosynthesis

Zunzhe Tian[1,6], Peng Zeng[2,6], Xiaoyun Lu[1,3,4,6], Tinggan Zhou[1], Yuwei Han[1], Yingmei Peng[1], Yunxue Xiao[5], Botong Zhou[1], Xue Liu[1], Yongting Zhang[1], Yang Yu[1], Qiong Li[1], Hang Zong[1], Feining Zhang[1], Huifeng Jiang[3,4,*], Juan He[1,*] and Jing Cai[1,*]

[1]School of Ecology and Environment, Northwestern Polytechnical University, Xi'an 710072, China

[2]State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macau, China

[3]Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China

[4]National Technology Innovation Center of Synthetic Biology, Tianjin 300308, China

[5]Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Kuming 650223, China

[6]These authors contributed equally to this article.

***Correspondence: Huifeng Jiang (jiang_hf@tib.cas.cn), Juan He (hejuan@nwpu.edu.cn), Jing Cai (jingcai@nwpu.edu.cn)**

https://doi.org/10.1016/j.xplc.2022.100464

ABSTRACT

Dipterocarpoideae, the largest subfamily of the Dipterocarpaceae, is a dominant component of Southeast Asian rainforests and is widely used as a source of wood, damar resin, medicine, and essential oil. However, many Dipterocarpoideae species are currently on the IUCN Red List owing to severe degradation of their habitats under global climate change and human disturbance. Genetic information regarding these taxa has only recently been reported with the sequencing of four Dipterocarp genomes, providing clues to the function and evolution of these species. Here, we report on 13 high-quality Dipterocarpoideae genome assemblies, ranging in size from 302.6 to 494.8 Mb and representing the five most species-rich genera in Dipterocarpoideae. Molecular dating analyses support the Western Gondwanaland origin of Dipterocarpaceae. Based on evolutionary analysis, we propose a three-step chromosome evolution scenario to describe the karyotypic evolution from an ancestor with six chromosomes to present-day species with 11 and 7 chromosomes. We discovered an expansion of genes encoding cellulose synthase (*CesA*), which is essential for cellulose biosynthesis and secondary cell-wall formation. We functionally identified five bornyl diphosphate synthase (BPPS) genes, which specifically catalyze the biosynthesis of borneol, a natural medicinal compound extracted from damar resin and oils, thus providing a basis for large-scale production of natural borneol *in vitro.*

Key words: Dipterocarpoideae, genome, chromosome evolution, cellulose synthase, borneol

Tian Z., Zeng P., Lu X., Zhou T., Han Y., Peng Y., Xiao Y., Zhou B., Liu X., Zhang Y., Yu Y., Li Q., Zong H., Zhang F., Jiang H., He J., and Cai J. (2022). Thirteen Dipterocarpoideae genomes provide insights into their evolution and borneol biosynthesis. Plant Comm. **3**, 100464.

INTRODUCTION

Southeast Asia contains nearly 15% of the world's tropical rainforests, including 4 of 25 globally important biodiversity hotspots (Sodhi et al., 2010). Dipterocarpaceae is an ecologically important family that occupies a dominant position in the tropical rainforests of Southeast Asia (Ashton, 1988). Dipterocarpaceae plants are also economically valuable owing to their unique fragrant oleoresins, high-quality timber, and bioactive components for traditional Chinese medicine (Rana et al., 2010; Dyrmose et al., 2017). However, many Dipterocarpaceae species are currently listed on the IUCN Red List (https://www.iucnredlist.org/) because of severe habitat deterioration under global climate change and intense human activities (Lewis, 2006; Bellard et al., 2014; O'Brien et al., 2014).

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

As the largest subfamily within the Dipterocarpaceae, Dipterocarpoideae contains 470–650 species within 13 genera (Appanah and Turnbull, 1998; Symington et al., 2004). However, the evolutionary history of dipterocarps is a long-disputed question. The main hypothesis suggests that Dipterocarpoideae originated in Western Gondwanaland in the Cretaceous period (~120 million years ago [mya]), predating the separation of Africa and South America (Ghazoul, 2016), based on fossil distribution in separated parts of Gondwanaland (Appanah and Turnbull, 1998) and very low dipterocarp dispersal ability (i.e., limited seed dispersal, poor seed dormancy, and seed salt intolerance) (Stone, 1983; Ashton and Gunatilleke, 1987). More recent phylogenetic analysis using a limited number of gene sequences and new fossil data supports the hypothesis that Dipterocarpaceae originated in Western Gondwanaland, although at a later date in the mid-Cretaceous (~102.9 mya) (Bansal et al., 2022). However, two recent molecular dating studies based on whole-genome sequences indicate that Dipterocarpaceae diverged from its most recent common ancestor with Malvaceae ~83.5 (Wang et al., 2022) to 86–98 mya (Ng et al., 2021), much later than the split of Gondwanaland. Thus, it is necessary to re-estimate divergence times based on additional whole-genome sequences and newly discovered fossils. Although the Dipterocarpoideae genomes are small, they can vary up to 2.64-fold, and the reason for this remains controversial (Grover and Wendel, 2010; Ng et al., 2016). Furthermore, although the basic chromosome numbers of the Dipterocarpeae (*Anisoptera*, *Cotylelobium*, *Dipterocarpus*, *Upuna*, and *Vatica*) and Shoreae (*Dryobalanops*, *Hopea*, *Neobalanocarpus*, *Parashorea*, and *Shorea*) tribes have been ascertained (11 and 7, respectively) (Ashton and Arboretum, 1979; Bawa, 1998) and all Dipterocarpoideae species share a recent whole-genome duplication (WGD) event (Ng et al., 2021; Wang et al., 2022), the chromosomal evolutionary history across Dipterocarpoideae remains unclear. Thus, clarifying the genomic background is important for exploring the genomic features of the Dipterocarpaceae family.

Dipterocarpoideae timber is widely used because of its hard and fine texture and strong moisture resistance (Rana et al., 2009). Wood tissue is mainly composed of secondary cell walls (SCWs) containing cellulose, lignin, and hemicellulose (Meents et al., 2018). A recent study reported on the expansion of several gene families associated with wood formation (e.g., cellulose synthase, *CesA*; cellulose synthase-like E, *CSLE*; laccase; and peroxidase) in two Dipterocarpaceae genomes compared with *Populus trichocarpa* and *Aquilaria sinensis*, and these expansions may have contributed to harder timber formation (Wang et al., 2022). However, the evolutionary characteristics of other genes involved in SCW synthesis in Dipterocarpoideae remain unknown. In addition, the resins and essential oils of Dipterocarpoideae plants are a natural source of borneol, also known as natural bingpian, which has been used for disease prevention and treatment in China for more than 2000 years (Li and Mu, 2004; Yang et al., 2020; Kulkarni et al., 2021). However, traditional extraction is costly in terms of labor and energy. Furthermore, chemically synthesized borneol is inherently impure and contains unexpected by-products, such as levogyral borneol and optically inactive iso-borneol (Zou et al., 2017). Considering the increasing demand for Dipterocarpaceae-derived borneol and timber, it is necessary to explore the genomic characteristics of Dipterocarpaceae to accelerate genome-assisted reproduction and clarify key enzymes of borneol synthesis to reconstruct the synthesis pathway through metabolic engineering.

Here, we *de novo* assembled 13 high-quality Dipterocarpoideae species genomes from the five most species-rich genera. Comparative genomic analysis revealed the evolution of Dipterocarpoideae species and a three-step chromosome evolution scenario. We also identified genes involved in borneol biosynthesis. Thus, this study provides a foundation for future research on the evolution, wood formation, and phytochemistry of Dipterocarpaceae.

RESULTS AND DISCUSSION

Genome sequencing, assembly, and annotation

We *de novo* sequenced and assembled the genomes of 13 species (four *Dipterocarpus* species, two *Hopea* species, three *Shorea* species, three *Vatica* species, and one *Parashorea* species) in the Dipterocarpoideae subfamily using long-read sequencing with Oxford Nanopore Technologies (ONT) and short-read sequencing with MGISEQ-2000 (BGI). For each species, one ONT library and one MGISEQ library were constructed, generating 24.6–66.8 Gb of long-read data and 50.8–81.4 Gb of short-read data, respectively (Supplemental Table 1). Thirteen scaffold-level genomes were obtained, ranging in size from 302.6 to 494.8 Mb, close to the estimates based on flow cytometry and *K*-mer analysis (Supplemental Table 2), with scaffold N50 lengths of 2.9–41.8 Mb and contig N50 lengths of 1.8–10.6 Mb (Table 1). Using high-throughput chromosome conformation capture (Hi-C) technology, 98.3%–99.7% of the contigs were anchored to 7 or 11 pseudochromosomes in five representative species (*Dipterocarpus gracilis*, *Hopea mollissima*, *Parashorea chinensis*, *Shorea henryana*, and *Vatica rassak*) (Table 1 and Supplemental Figure 1), consistent with previous karyotype studies (Ng et al., 2016). Genome assembly completeness was evaluated by benchmarking universal single-copy orthologs (BUSCOs) and the proportion of properly aligned BGI paired-end reads. More than 96.8% of the 1614 single-copy embryophyte genes were complete in the 13 species (Table 1 and Supplemental Table 3). The mapping ratios of properly paired-end short-read data for the 13 genomes ranged from 94.7% to 99.5%, with coverage ratios of 97.7%–99.7% (Table 1 and Supplemental Table 4). Thus, we obtained 13 high-quality genomes.

To annotate the genes in the genome, we used PacBio and RNA-sequencing technologies to generate 20.2–31.3 Gb of full-length cDNA data and 6.6–9.2 Gb of short-read data, which were then assembled into transcripts. A pipeline combining transcriptome data, *de novo* predictions, and homology-based predictions was used to construct gene models for the Dipterocarpoideae genomes. A total of 33 950–44 078 protein-coding genes were annotated in these genomes, over 83.15% of which were functionally classified and supported in public databases, including InterPro, Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), SwissProt, and TrEMBL (Supplemental Table 5). The protein-coding genes of the 13 species were similar in average gene length (2618.11–2885.76 bp), average coding sequence length (1090–1168 bp), average exon number per

Species	Assembly size (Mb)	Scaffold N50 (Mb)	Contig N50 (Mb)	Mapping ratio (%)	Coverage (%)	Chromosome number	Chromosome loading ratio (%)	BUSCO ratio of genome (%)	Gene number
Dipterocarpus alatus	444.6	2.9	1.8	95.5	98	–	–	98.5	42 423
Dipterocarpus gracilis	385	22.1	2.1	96.4	98.7	11	99.5	98.2	38 375
Dipterocarpus intricatus	357.6	18.2	6.6	96.3	98.8	–	–	98.9	37 824
Dipterocarpus zeylanicus	362.3	27	3.5	94.7	99	–	–	97.9	37 356
Hopea mollissima	361.6	29.3	2.8	96.2	98.6	7	99.7	98	35 939
Hopea odorata	478.5	26.9	4.1	99.5	98	–	–	97.1	44 078
Parashorea chinensis	315.3	39.7	9.6	94.8	99.7	7	98.3	98.3	36 597
Shorea leprosula	323.6	8.1	5.5	95.9	99.2	–	–	98.4	34 469
Shorea roxburghii	306.2	21.5	10.1	97.5	99	–	–	98.8	34 473
Shorea henryana	302.6	41.8	10.6	96.9	99.2	7	98.6	98.3	33 950
Vatica xishuangbannaensis	456.5	12.5	4.1	96	98.2	–	–	98.1	36 861
Vatica odorata	494.8	10.8	3	96.9	97.7	–	–	96.8	39 825
Vatica rassak	476.2	7.5	2.1	97	97.8	11	97.1	97.7	39 603

Table 1. Statistics of 13 Dipterocarpoideae genome assemblies.

gene (4.96–5.34), average exon length (214.81–232.21 bp), and average intron length (357.00–447.29 bp) (Supplemental Table 6), indicating that the gene structures of these species were relatively conserved (Figure 1). BUSCO evaluation also showed that at least 92.2% of the 1614 single-copy embryophyte genes were completely annotated in the 13 genomes (Supplemental Table 7).

Comparative genomic analysis

We compared the 13 Dipterocarpoideae genomes with five other sequenced genomes of angiosperm species (*Amborella trichopoda*, *Arabidopsis thaliana*, *A. sinensis*, *Theobroma cacao*, and *Oryza sativa*). Based on gene family clustering analysis, we identified 429 single-copy orthologous gene families among the 18 species and combined all alignments to produce a superalignment matrix to construct a phylogenetic tree. The phylogenetic tree showed that all 13 Dipterocarpoideae species were clustered into one monophyletic group, but three *Shorea* species (*S. henryana*, *S. roxburghii*, and *S. leprosula*) were separated into two branches with 100% support, with *S. leprosula* and *P. chinensis* clustered on one of the two branches (Figure 2A). This finding is consistent with previous studies using other molecular markers (Cao et al., 2006; Heckenhauer et al., 2017). However, classical taxonomy treats all *Shorea* species as a monophyletic group based on floral traits (e.g., number of stamens, anther structure). This conflict between molecular phylogenetic analysis and classical taxonomy may be due to certain morphological characters (e.g., stamens and anthers) evolving in parallel in Shoreae (Heckenhauer et al., 2018), suggesting that the *Shorea* classification may need to be revised.

We used MCMCTree to estimate the divergence times of the Dipterocarpoideae lineages based on Dipterocarpaceae fossils and secondary calibrations (Vega et al., 2006; Bell et al., 2010; Bansal et al., 2022). According to MCMCTree analysis (Figure 2A), the Dipterocarpoideae lineages diverged from *T. cacao* ~147.3 mya (95% confidence interval: 125.8–159.4 mya), much earlier than the dates reported by Ng et al. (Ng et al., 2021) (86–98 mya) and Wang et al. (Wang et al., 2022) (~83.5 mya). Our findings also suggested that the divergence event between Dipterocarpaceae and *T. cacao* occurred before the Atlantic separation of South America and Africa, resolving previous conflicting estimates of the split of dipterocarps from their closest relative in the Malvaceae (*T. cacao*) and the separation time of South America and Africa. The earliest divergence within the Dipterocarpoideae lineage arose ~105 mya (84.7–123.7 mya), relatively consistent with the recent study of Bansal (~94.6 mya) (Bansal et al., 2022), and suggesting the

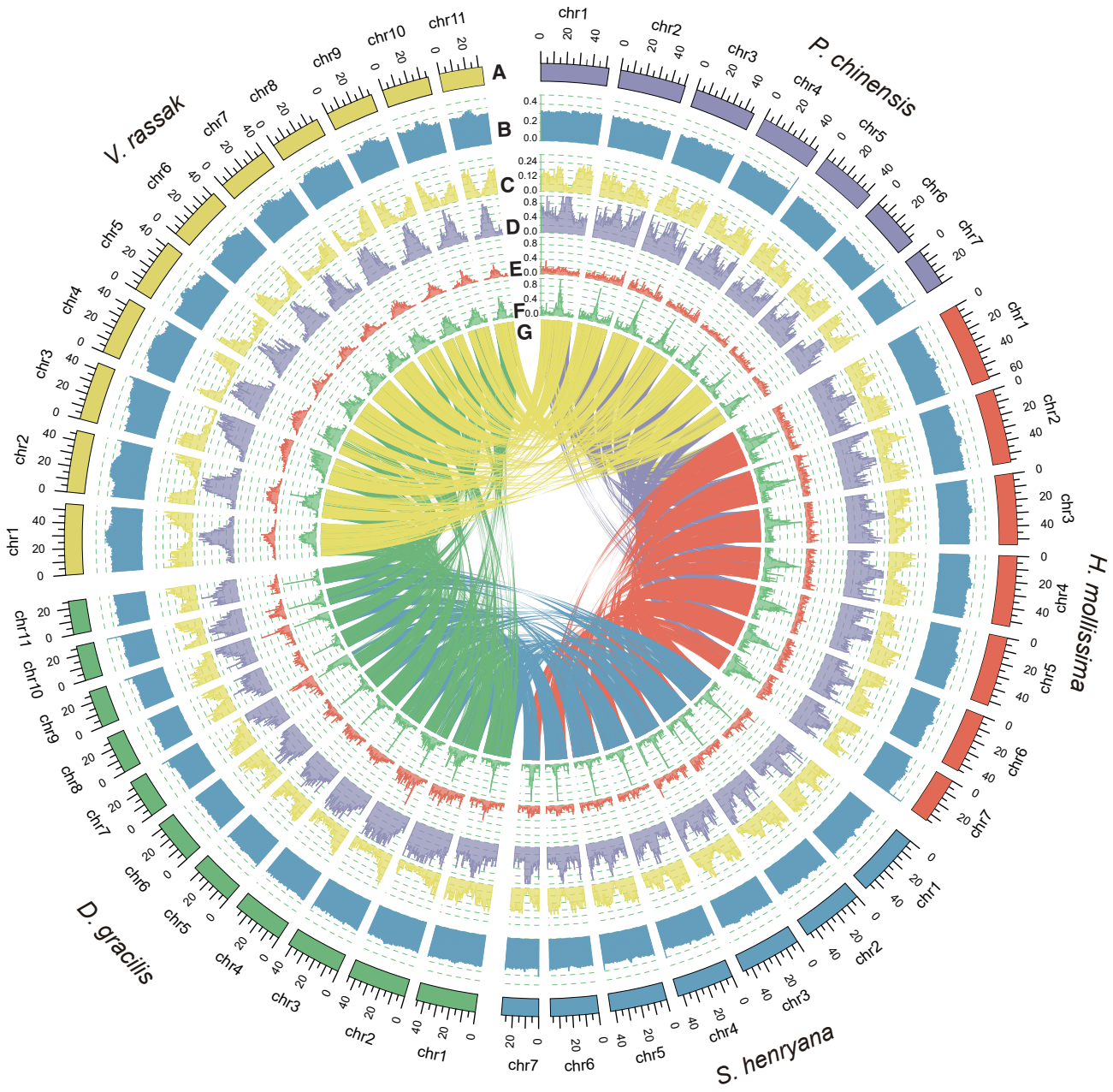

Figure 1. Characteristics of five chromosome-level genomes (*D. gracilis*, *H. mollissima*, *P. chinensis*, *S. henryana*, and *V. rassak*).
(a) Chromosomes, (b) GC density, (c) gene density, (d) transposable element (TE) coverage, (e) Copia coverage, (f) Gypsy coverage, and (g) links between two adjacent genomes connecting syntenic gene blocks, as detected using MCScanX.

evolution of Dipterocarpoideae during the mid-Cretaceous, and supporting the hypothesis of the Western Gondwanaland origin of Dipterocarpaceae. In the future, additional genomic data from the subfamily Monotoideae in the Africa n continent will help to clarify the evolutionary history of dipterocarps.

Significant gene family expansion or contraction is associated with adaptive divergence in related species (Dassanayake et al., 2011). To explore their adaptive divergence, we compared Dipterocarpoideae species with five other angiosperm species and identified 1237 expanded gene families and 1441 contracted gene families in the common ancestor of Dipterocarpoideae species (Supplemental Figure 2). Based on GO analysis, the expanded gene families were enriched in cellulose biosynthesis, plant hormones, and immune response (Supplemental Table 8 and see Figure 4E).

Transposable elements (TEs), which are common features in eukaryotic genomes, can critically affect genome evolution and gene regulation (Chénais et al., 2012; Kim et al., 2017). The mobility of TEs can induce deleterious mutations, gene disruptions, and chromosomal rearrangements (Klein and O'Neill, 2018). To examine differences in TEs among the 13 genomes, we combined *de novo* and homology-based

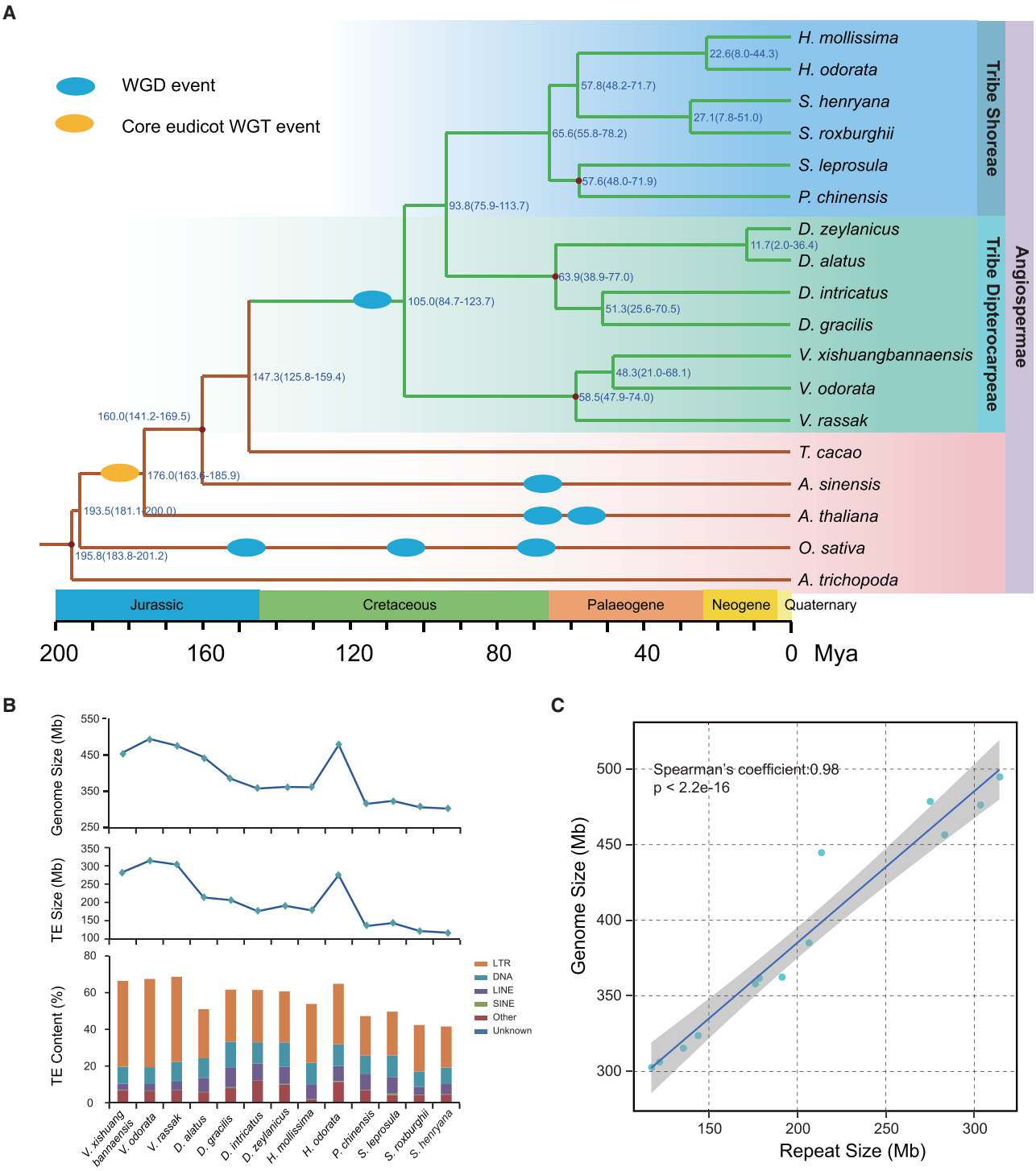

Figure 2. Comparative genomic analysis.

(A) Phylogenetic tree of 18 species, including divergence times, based on 429 single-copy genes. Ovals indicate polyploidization events during angiosperm history. All nodes were supported with bootstrap values of 100.

(B) Genome size, TE size, and different TE contents in 13 species.

(C) Correlation between genome size and TE content.

approaches and identified 38.81%–63.68% of TEs from the 13 assembled genomes. Among the annotated TEs, long terminal repeats (LTRs) were the most abundant, accounting for 21.6%–48.1% of total TEs (Figure 2B and Supplemental Tables 9 and 10). TEs can affect genome size by affecting genome expansion (Novák et al., 2020). We therefore explored the potential correlation between genome size and TE content in the 13 Dipterocarpoideae species. Results showed a

correlation coefficient of 0.98 ($p < 2.2 \times 10^{-16}$), indicating that TEs were the main cause of genome size variation (Figures 2B and 2C). Furthermore, the high LTR content (>45%) in *Vatica* may be responsible for its larger genome size compared with species in other genera (Figure 2B and Supplemental Tables 9 and 10). To further examine genome size variation in Dipterocarpoideae, we estimated historical TE expansion activity via Kimura distance analysis (Chalopin et al., 2015). Results showed that all species in each genus shared a typical genus-specific dynamic history of TEs, but different genera had remarkably different TE histories, as reflected in the different ratios of recent TEs among the five genera (Supplemental Figure 3). Interestingly, we found that almost every pair of LTRs in the full-length LTR retrotransposons diverged by less than 0.1 across the 13 species, indicating recent LTR insertion (Supplemental Figure 4).

Whole-genome duplication and three-step chromosome evolution

Polyploidization or WGD events have widely shaped the genomic evolution of angiosperms (Jiao et al., 2011; Wendel et al., 2016). To investigate WGD events during Dipterocarpoideae evolution, we determined the distribution of synonymous substitutions per synonymous site (*Ks*) using syntenic paralogs within each genome. Results showed that two polyploidization events occurred in all Dipterocarpoideae species (Figure 3A and Supplemental Figure 5). To confirm the polyploidization events in the Dipterocarpoideae species, we conducted an intra-/intergenomic search for syntenic blocks across five species (*H. mollissima*, *S. henryana*, *P. chinensis*, *D. gracilis*, and *V. rassak*) with chromosome-level assemblies. The widespread occurrence of 6:6 syntenic blocks within and between Dipterocarpoideae species also suggested that two polyploidization events (one triplication and one duplication) occurred in these species (Supplemental Figure 6). Combining the syntenic relationships and *Ks* distributions, we found that two of the six syntenic blocks were located at the lower *Ks* peak and four were located at the higher *Ks* peak, indicating that the recent polyploidization event was a WGD, as reported in recent studies (Ng et al., 2021; Wang et al., 2022), whereas the other polyploidization event was a whole-genome triplication. Thus, together with the above phylogeny, these results showed that the Dipterocarpoideae species shared a whole-genome triplication (γ) event, as reported for the common ancestor of core eudicots (Jaillon et al., 2007; Tang et al., 2008a; Jiao et al., 2012), and a WGD event that occurred after the divergence from *T. cacao* but before the earliest divergence time within Dipterocarpoideae species, consistent with previous research (Ng et al., 2021; Tang et al., 2022).

Angiosperm genomes have been shaped by many rounds of paleopolyploidization and various genomic changes, including chromosome fusion, fission, and loss (Jaillon et al., 2007). In addition to the common WGD (α) event found in all 13 Dipterocarpoideae species prior to their divergence, Dipterocarpeae also showed different chromosomal rearrangement histories after the WGD, first giving rise to the tribe Dipterocarpeae (*D. gracilis* and *V. rassak*), with 11 chromosomes, and then to the tribe Shoreae (*S. henryana*, *H. mollissima*, and *P. chinensis*), with 7 chromosomes. To reconstruct the ancestral pre-WGD chromosomes of Dipterocarpoideae species, we carried out intra-/intergenomic alignment of the five chromosome-level genomes. The intragenomic syntenic relationships showed that the species with 11 chromosomes (*D. gracilis* and *V. rassak*) had more intact chromosome-to-chromosome collinearity relationships (chr1 vs. chr9, chr2 vs. chr4, chr3 vs. chr8, chr10 vs. chr11, chr5 + chr6 vs. chr7), whereas the collinearity relationships in species with seven chromosomes (*H. mollissima*, *S. henryana*, and *P. chinensis*) were mostly fragmented among the chromosomes, except for one intact chromosome pair (chr3 vs. chr4) (Supplemental Figure 7). These results indicated that species with 11 chromosomes retained more intact ancestral homologous chromosome pairs generated by the WGD (α) event and experienced less chromosomal rearrangement than species with 7 chromosomes. In addition, most of the syntenic blocks of chr5 and chr6 in the Dipterocarpeae tribe could be mapped to different chromosomes in the genome of the outgroup species *T. cacao* (Figure 3B), suggesting that chr5 and chr6 were two independent chromosomes in the ancestral genome of Dipterocarpoideae before the WGD (α) and that chr7 fused with the other copies of chr5 and chr6 after the WGD (α). Thus, the ancestor of Dipterocarpoideae species probably had 6 protochromosomes (named AS1–6) before the WGD event and 12 protochromosomes after the WGD event (Figure 3D).

Based on the phylogenetic relationships among Dipterocarpoideae species, we found that species of the Shoreae tribe originated from the Dipterocarpeae tribe (Figure 2A). To characterize the evolutionary scenario under which species changed from 11 to 7 chromosomes, we performed intergenomic syntenic analysis and determined the *Ks* distribution of syntenic homologs between *D. gracilis* and *P. chinensis*, representing Dipterocarpeae and Shoreae, respectively. As the WGD event occurred earlier than the *D. gracilis–P. chinensis* divergence, the syntenic blocks with lower and higher *Ks* values were orthologous and paralogous, respectively (Supplemental Figure 8). The orthologous relationships indicated that the chromosomes of *P. chinensis* originated from *D. gracilis* via 16 chromosome fissions and 20 chromosome fusions, with 14 of the 16 fission sites also supported by *V. rassak* and 18 of the 20 fusion sites also supported by *S. henryana* and *H. mollissima*. Only two additional fission sites and two additional fusion sites were supported by one species (Figure 3C), possibly owing to species-specific chromosomal rearrangement events or genome assembly errors.

Thus, we propose that the ancestral Dipterocarpoideae genome underwent a three-step evolutionary process of chromosome structural variation. First, the number of chromosomes in Dipterocarpoideae ancestors evolved from 6 to 12 through a WGD event; second, the 12-chromosome ancestors evolved into 11-chromosome species through two fission and three fusion events between chr5 and chr6 (Figures 3B and 3D); and third, the 11-chromosome species evolved into 7-chromosome species through at least 16 fissions and 20 fusions (Figures 3C and 3D). This three-step chromosome evolutionary history in Dipterocarpoideae may have contributed to its speciation and diversification.

Gene family analysis related to SCW formation

Dipterocarpoideae timber is widely used because of its hardness, high strength, and strong moisture resistance (Appanah and Turnbull, 1998). Wood tissue in trees consists mainly of SCWs,

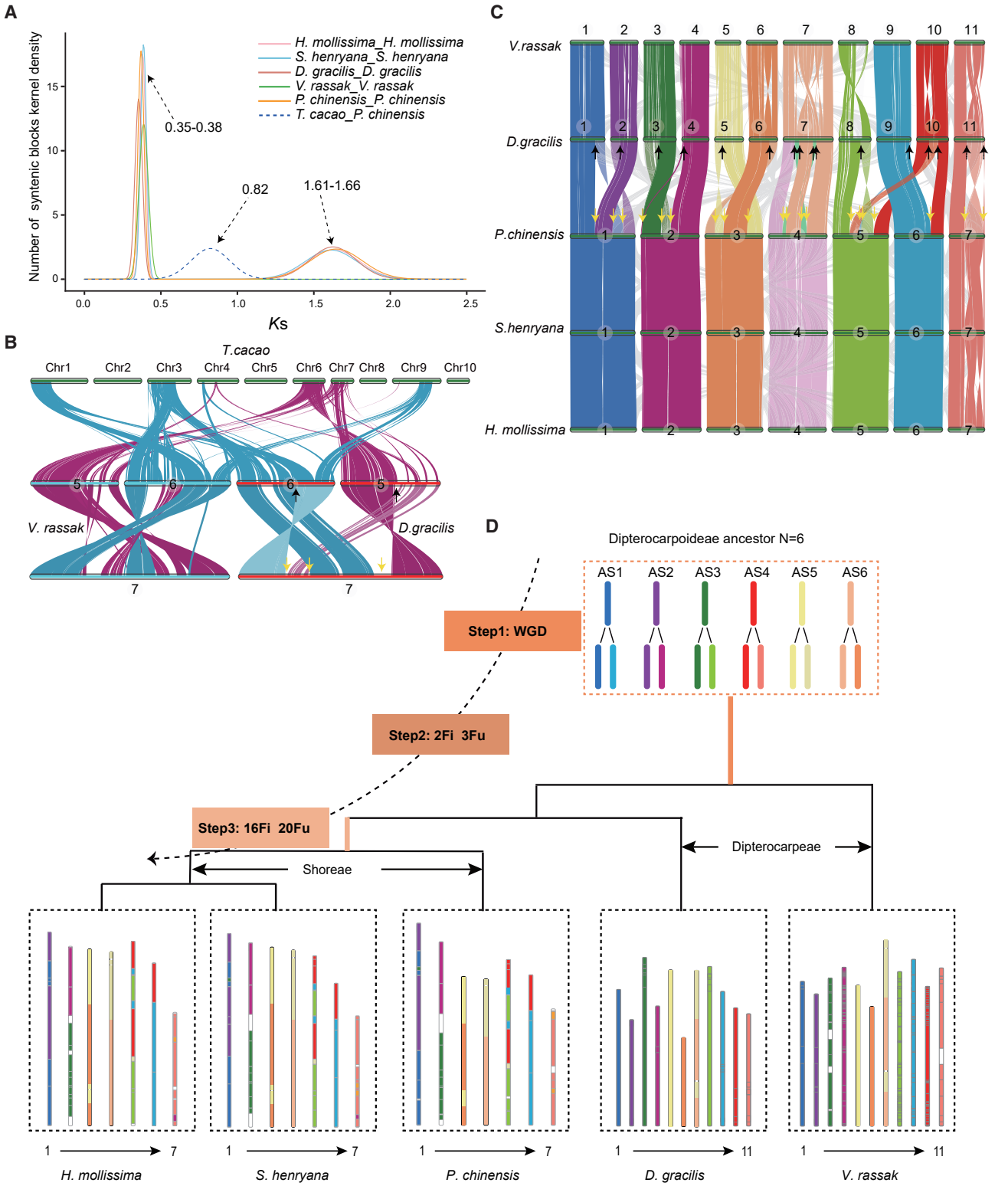

Figure 3. Whole-genome duplication and chromosome evolutionary history in Dipterocarpoideae species.

(A) Synonymous substitution rate distributions of syntenic block pairs in five Dipterocarpoideae species and between *P. chinensis* and *T. cacao*.

(B) Collinearity analysis between Dipterocarpeae tribe (chr5, chr6, chr7) and *T. cacao*; black and yellow arrows represent positions of fission and fusion, respectively.

(C) Collinearity analysis of five chromosome-level genomes. Black and yellow arrows represent positions of fission and fusion, respectively; dotted and solid arrows indicate positions supported by multiple and single species, respectively.

(D) Three-step chromosome evolution scenario of Dipterocarpoideae species from six ancestral chromosomal Dipterocarpoideae karyotypes.

which reinforce tracheary elements and strengthen fibers to permit upright growth and the formation of forest canopies. SCWs are mainly composed of cellulose, lignin, and hemicellulose (Meents et al., 2018). To explore the genetic basis of SCWs in Dipterocarpoideae species, we identified gene families related to cellulose and lignin biosynthesis (Supplemental Tables 11 and 12, Supplemental Figures 9 and 10, and Figure 4A). Cellulose, which accounts for 40%–45% of the wood cell wall, plays a key role in the biomechanical properties of cells and is important for the growth and formation of wood (Higuchi, 2006; Mellerowicz and Sundberg, 2008; Popper et al., 2011). The total number of *CesA* and *CSL* genes in the 13 Dipterocarpoideae genomes ranged from 33 to 54, fewer than previously reported in Dipterocarpoideae genomes (64/60) (Wang et al., 2022). This discrepancy may be related to our stricter filtering of genes that lacked four conserved motifs (i.e., QxxRW, DD, DCD, and TED), as per previous studies (Morgan et al., 2013; Zou et al., 2017). Among the cellulose-related gene families, the *CSL* gene family and all subfamilies in the Dipterocarpoideae genomes did not differ significantly in copy number compared with the closely related species *T. cacao* and *A. sinensis* (Supplemental Table 11 and Supplemental Figure 9), inconsistent with a previous study that reported *CSLE* subfamily expansion in *D. turbinatus* and *H. hainanensis* compared with *A. sinensis* (Wang et al., 2022). However, we identified 13–16 expanded *CesA* genes in Dipterocarpoideae compared with *A. thaliana* (10 *CesA* genes), *T. cacao* (11 *CesA* genes), and *A. sinensis* (six *CesA* genes) (Figure 4A and Supplemental Table 11). An increase in *CesA* gene copies may enhance the biosynthesis of cellulose and promote the formation of hardwood (Popper et al., 2011; Polko and Kieber, 2019). Phylogenetic analysis of *CesA* genes from the Dipterocarpoideae species showed that these genes could be divided into eight subfamilies, similar to *T. cacao*, with two new groups (groups 1 and 2) not present in *A. thaliana* and *A. sinensis* (Figure 4A and Supplemental Table 11). By counting the numbers of *CesA* genes in syntenic blocks from the WGD, we found that the Dipterocarpoideae species retained a higher ratio of *CesA* family genes derived from the WGD than did *A. thaliana*, *T. cacao*, and *A. sinensis*, indicating that the WGD event played a more important role in the expansion of this gene family in Dipterocarpoideae than in the two outgroup species (Supplemental Table 11).

As plant hormones play important roles in the regulatory cascades that guide SCW formation and patterning (Didi et al., 2015), we explored genes related to plant hormones. Results showed that genes encoding cytokinin oxidase/dehydrogenase (*CKX*) were expanded in 12 Dipterocarpoideae species (more than nine *CKX* genes) (Figure 4B and Supplemental Table 13) compared with *T. cacao* (seven *CKX* genes) and *A. thaliana* (seven *CKX* genes). Similar to the pattern found for the *Ces*A family, most *CKX* groups retained two paralogs from the WGD (α), showing that the WGD event also played a major role in the expansion of the *CKX* gene family. We also identified several rapidly evolving genes (REGs) (i.e., *IAA*, *SAUR*, *BRI1*, *BAK1*, *BZR1/2*, *DELLA*, and *TF*) and one positively selected gene (PSG) (i.e., *AHP*) (Figure 4C and Supplemental Tables 14 and 15) involved in plant hormone signaling pathways in the Dipterocarpoideae species. The *AHP* gene showed Dipterocarpoideae-specific mutations in the histidine-containing phototransfer domain (Figure 4D), which may control the SCW transcriptional cascade. Based on GO enrichment analysis of the expanded gene families in the Dipterocarpoideae species, we also found several enriched GO terms related to osmotic regulation, including transmembrane transporter activity (GO: 0022857), potassium ion transmembrane transport (GO: 0071805), and anion transport (GO: 0006820), which may be related to their adaptation to the high-rain environments of tropical rainforests. Therefore, we hypothesize that both expanded gene families (*CesA* and *CKX*) and genes under natural selection jointly contribute to SCW formation in Dipterocarpoideae species (Kato et al., 2015).

Functional identification of bornyl diphosphate synthase in *Dipterocarpus* species

Many *Dipterocarpus* species can produce essential oils (Ashton and Arboretum, 1979) that possess significant antiseptic, antibacterial, antiviral, antioxidant, antiparasitic, antifungal, and insecticidal activities (Burt, 2004; Kaloustian et al., 2008). Essential oils contain a variety of volatile molecules such as terpenes and terpenoids (Bakkali et al., 2008). Terpene synthases (TPSs) are important enzymes in the formation of the basic carbon skeleton of terpenes (Chen et al., 2011). Here, we obtained 99 TPS genes from four *Dipterocarpus* species (28 in *D. alatus*, 22 in *D. gracilis*, 24 in *D. intricatus*, and 25 in *D. zeylanicus*) based on a BLAST search (identity >30, $p < 1 \times 10^{-5}$) and an hmmsearch using HMMER (Supplemental Figure 11). Phylogenetic analysis of the TPS gene family and known TPS genes of *A. thaliana* showed that the TPS gene family in the four *Dipterocarpus* genomes could be divided into five subfamilies: TPS-a, TPS-b, TPS-c, TPS-e/f, and TPS-g (Supplemental Figure 11). Compared with *A. thaliana* and two reported Dipterocarpoideae genomes, there were no significant differences in the number of whole TPS gene families in the four species (Wang et al., 2022). Among the five subfamilies, TPS-e/f was expanded in the Dipterocarpoideae genomes (seven to nine genes) (Supplemental Figure 11) compared with *A. thaliana* (two genes) and *T. cacao* (four genes). The increase in copy number of TPS-e/f genes may have increased terpene production in the Dipterocarpoideae species (Falara et al., 2011).

Borneol, a bicyclic monoterpene found in several species of Dipterocarpaceae, has been widely used for anxiety, pain, and anesthesia in Chinese medicine for more than 2000 years (Appanah and Turnbull, 1998; Zhang et al., 2017; Kulkarni et al., 2021). Bornyl diphosphate synthase (BPPS), a key enzyme for borneol biosynthesis, is found in Lamiaceae and Lauraceae plants. This enzyme is involved not only in borneol biosynthesis but also in the production of multiple compounds (Whittington et al., 2002; Hurd et al., 2017; Wang et al., 2018; Lei et al., 2021). To identify BPPS genes in Dipterocarpaceae, we filtered the functional motif sequences and obtained 14 candidate genes that contained the three motifs conserved in BPPS proteins (i.e., RRX8W, DDXXD, and NSE/DTE) (Whittington et al., 2002; Hurd et al., 2017) (Figure 5B).

Borneol synthesis is a process of complex carbon rearrangement that produces a variety of intermediate products (Figure 5A) (O'Brien et al., 2018). To verify their functions, we expressed the 14 candidate proteins in *Escherichia coli* and performed enzymatic characterization *in vitro* using geranyl diphosphate (GPP) as a substrate. Unlike other monoterpene synthases, BPPS

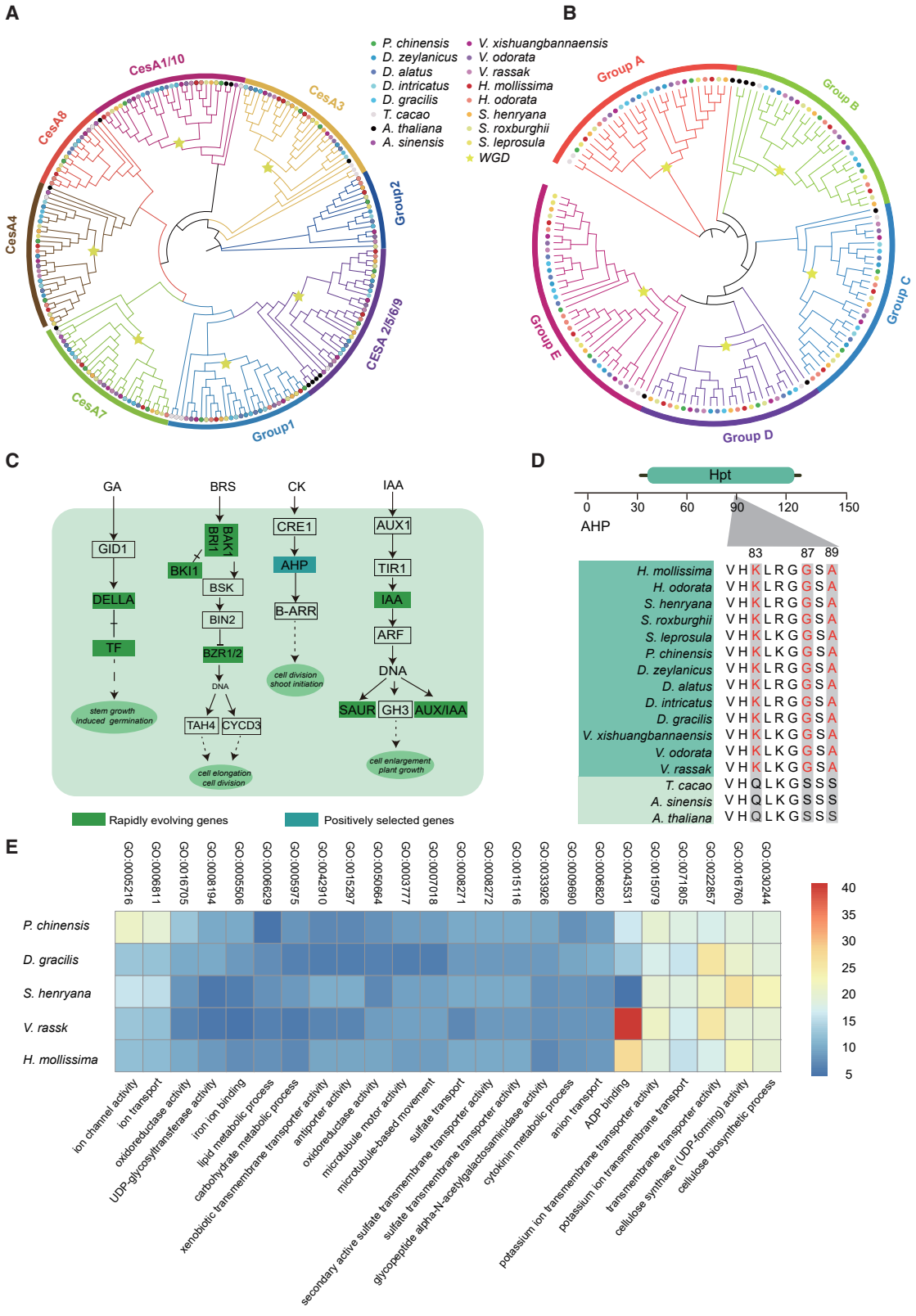

Figure 4. Overview of gene families and genes associated with SCW formation.

(A and B) Maximum-likelihood phylogenetic trees of the *CesA* **(A)** and the *CKX* **(B)** gene families (*CesA* and *CKX* gene families were identified from 13 Dipterocarpoideae species, *T. cacao*, *A. sinensis* and *A. thaliana*). Different colored dots correspond to different species, different colored circles correspond to different *CesA* or *CKX* groups, and yellow stars indicate approximate positions of WGD events in the gene trees.

(legend continued on next page)

catalyzes the conversion of GPP to borneol diphosphate, resulting in the need for additional alkaline phosphatase to obtain borneol. All the mass spectra of mixed standards and samples were analyzed by gas chromatography-mass spectrometry (GC-MS) (Supplemental Figure 12). The characteristic fragment masses of monoterpenes such as borneol are well known (93.0699, 95.0855). Hence, we analyzed the extracted ions of samples and mixed standards that contained six monoterpenes (Figure 5C).

Five of the 14 candidate proteins (*Di*TPS3, *Di*TPS2, *Dz*TPS3, *Da*TPS3, and *Dg*TPS1) were able to catalyze borneol synthesis (Figure 5C and Supplemental Figure 13). Unlike previously reported BPPS proteins that produce borneol with low specificity (Hurd et al., 2017; Wang et al., 2018), the five BPPS proteins of *Dipterocarpus* produced borneol with very high specificity and almost no other intermediate products (Figure 5C). The specific activities of these five BPPS enzymes can contribute to our understanding of monoterpene diversification in *Dipterocarpus* plants and lay a foundation for further pathway reconstruction and metabolic engineering to produce monoterpenoids, especially borneol. To date, the absence of Dipterocarpoideae species genomes has hindered research on their evolutionary history, conservation, and utilization (Cao, 2004; Ng et al., 2016). Here, based on 13 genomes, we systematically elucidated the evolutionary history of Dipterocarpoideae and the mechanism underlying SCW formation. We also identified five functionally specific BPPS proteins. Our research provides an important genetic resource for the conservation and utilization of Dipterocarpoideae species.

METHODS

Plant material, library construction, and sequencing

The 13 Dipterocarpoideae species (Table 1) used in this study were collected from the Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Yunnan, China.

Fresh leaves were harvested and immediately frozen with liquid nitrogen to preserve genomic DNA for isolation. For long-read sequencing, an ONT library was constructed using large (>20 kb) DNA fragments with a Ligation Sequencing Kit (SQK-LSK109; ONT, Oxford, UK) and then sequenced using the ONT PromethION platform. For short-read sequencing, paired-end libraries with an insertion size of 350 bp were constructed according to the manufacturer's protocols and sequenced using the MGISEQ-2000 sequencing platform. For Hi-C sequencing, libraries were created from fresh leaves of five species by BioMarker Technologies (Beijing, China) according to published methods (Xie et al., 2015) and sequenced using the Illumina platform for 150-bp paired-end reads. For RNA sequencing, PacBio Iso-Seq (PB) and RNA libraries were constructed following the manufacturer's recommendations and sequenced using the PacBio RS II and Illumina platforms, respectively.

Detailed sequencing information on the 13 species is provided in Supplemental Table 1. All genomic sequences were generated by BioMarker Technologies (Beijing, China).

Genome size estimation

To estimate the genome sizes of the 13 species, we counted the 17-mer frequency from the MGISEQ-2000 150-bp paired-end reads using GCE v1.0.2 software (Liu et al., 2013) (https://github.com/fanagislab/GCE/tree/master/gce-1.0.2). Genome size was then calculated according to the following formula: genome size = *K*-mer coverage/mean *K*-mer depth. To validate the genome size estimates, we performed flow cytometry analysis compared with *Solanum lycopersicum* to estimate nuclear DNA content (Supplemental Table 2).

Genome assembly and assessment

NextDenovo v.2.1 (https://github.com/Nextomics/NextDenovo) was used for self-correction of long reads sequenced using the ONT platform and assembled based on reads longer than 15 kb. The assembled genomes were polished for two rounds with NextPolish v.1.01 (https://github.com/Nextomics/NextPolish/issues/new) using the ONT long reads and BGI short reads. The Purge Haplotigs pipeline (Roach et al., 2018) was then used to remove heterozygous regions of the genome based on ONT reads. We used the 3D-DNA pipeline to organize scaffolds into pseudochromosomes of the five species (Dudchenko et al., 2017) based on Hi-C reads. After the first round of 3D-DNA, we used the Juicebox Assembly Tools v.1.9.9 to improve the assembled genome by manual adjustment (Durand et al., 2016), including changing the order, correcting mis-joins and orientations of draft scaffolds with translocation and inversion errors, and adjusting chromosome boundaries.

To evaluate the quality of the 13 assembled genomes, genome completeness was assessed using the embryophyta_odb10 dataset of BUSCO (v.4.0.5) (Seppey et al., 2019) with 1614 single-copy genes. In addition, BWA (v.0.7.17) (https://github.com/lh3/bwa) was used to map BGI reads to the assembled genomes, and iTools (v.0.24) was then used to calculate the mapping rate and genome coverage ratio (https://github.com/BGI-shenzhen/Reseqtools).

Repeat annotation

Repetitive sequences and TEs in the genome were identified using complementary homology- and *de novo*-based methods. We used RepeatMasker (v.4.1.0) (Chen, 2004) to search for known repetitive elements using a crossmatch program with a Repbase-derived RepeatMasker library. We constructed an LTR library using RepeatModeler (v.1.0.5), RepeatScout (Price et al., 2005), PILER, and LTR_FINDER (Xu and Wang, 2007) with default parameters. The consensus TE sequences from the repetitive element library were imported into RepeatMasker to annotate tandem repeats within the genomes. Full-length LTR retrotransposons were identified in the 13 genomes using LTR_FINDER and LTRharvest (Ellinghaus et al., 2008) with default parameters.

(C) Plant hormone metabolic pathways in Dipterocarpoideae species.
(D) Dipterocarpoideae-specific mutations located in the histidine-containing phototransfer (Hpt) domain of the *AHP* gene in Dipterocarpoideae species.
(E) Gene Ontology (GO) terms enriched in expanded gene families from five Dipterocarpoideae lineages. Each lineage was represented by one species in each genus.

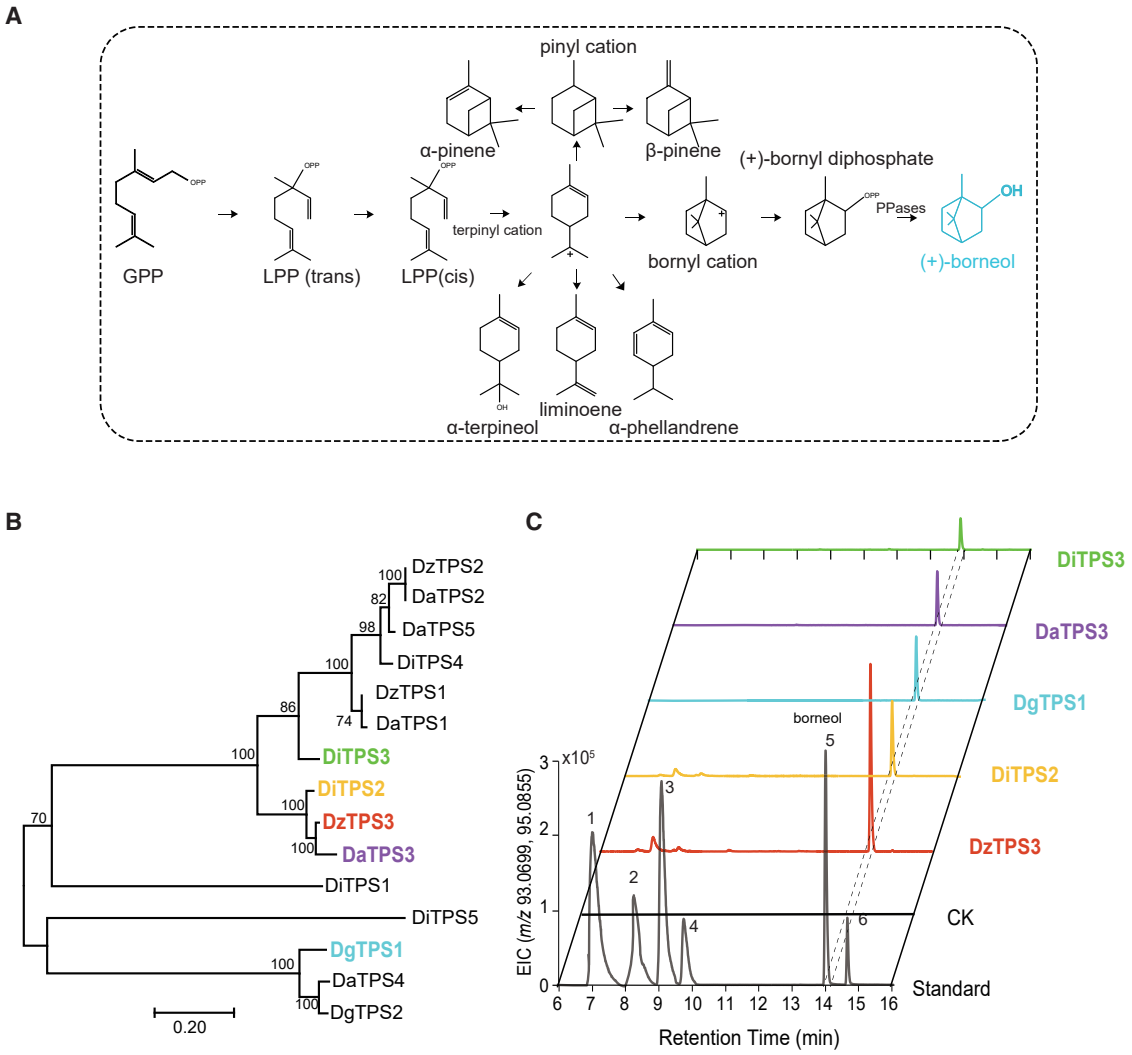

Figure 5. Biosynthetic pathway for borneol and functional identification of BPPS genes.
(A) Primary pathway leading to the formation of borneol and other monoterpenoid products, according to a previous study (Ma et al., 2021).
(B) Phylogenetic analysis of 14 candidate genes from the genus *Dipterocarpus*. Genes in colored fonts function in borneol production.
(C) GC-MS analysis of five BPPS genes (*DiTPS2*, *DiTPS3*, *DzTPS3*, *DaTPS3*, and *DgTPS1*). Peak 1, α-pinene; peak 2, β-pinene; peak 3, α-phellandrene; peak 4, limonene; peak 5, borneol; peak 6, α-terpineol.

Protein-coding gene prediction and functional annotation

Protein-coding genes of the 13 assembled genomes were predicted using a combination of homology-, *de novo*-, and transcriptome-based predictions from the repeat-masked genome. Protein sequences from six species (*A. sinensis*, *A. thaliana*, *T. cacao*, *Corchorus olitorius*, *Durio zibethinus*, and *Gossypium tomentosum*) were downloaded from Phytozome (v.11.0) (Goodstein et al., 2012) and aligned to the assembled genomes using TBLASTn (Altschul et al., 1990) with a filter parameter ($E < 1 \times 10^{-5}$). We used Solar software (Yu et al., 2006) to predict the gene structures of the corresponding genomic regions for each BLAST hit with GeneWise (Birney et al., 2004). For *de novo* prediction, two gene prediction programs, Augustus (Stanke et al., 2006) and GlimmerHMM (Majoros et al., 2004), were used to predict coding regions in the repeat-masked genome, with parameters trained from closely related species. We assembled the RNA-sequencing reads and full-length transcription PacBio reads into transcripts using Trinity (Haas et al., 2013) and the Iso-Seq3 pipeline, respectively. We then mapped the transcripts to the assembled genomes using Blat (Kent, 2002). Finally, we integrated all predictions using EVidenceModeler (EVM v.1.1.1) (Stanke et al., 2008) to generate consensus gene sets. We also used BUSCO to evaluate gene annotation completeness.

For functional annotation, the predicted protein-coding genes were searched against the SwissProt, TrEMBL (Bairoch and Apweiler, 2000), KEGG (Kanehisa and Goto, 2000), and InterPro (Quevillon et al., 2005) databases.

Polyploidization analysis and genome structural comparisons

Ancient WGD or polyploidy is highly prevalent in plants (Jiao et al., 2011; Soltis and Soltis, 2016; Wendel et al., 2016). We performed intra- and intergenomic comparisons between *T. cacao* and

the 13 Dipterocarpoideae species using the WGDI pipeline (Sun et al., 2021). We first identified syntenic blocks of homologous genes within and between all the genomes. We then calculated synonymous substitutions per synonymous site (*Ks*) between homologous genes within syntenic blocks using the Nei–Gojobori algorithm (Nei and Gojobori, 1986). We used the medium *Ks* value to represent each syntenic block. Multipeak fitting of the curve was performed using the Gaussian approximation function (cftool), and the coefficient of determination (R^2) was set to ≥ 0.90.

To track chromosomal evolutionary history among Dipterocarpoideae species, we identified intra-/interspecific syntenic blocks using MCscan (Python version) with default parameters and at least 10 genes required to define a syntenic block (https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)). *Ks* values were calculated with the KaKs_Calculator (Zhang et al., 2006) for each gene pair in the aligned blocks. For genome structure comparisons, we performed intragenome collinearity analysis of the five chromosome-level genomes of the Dipterocarpaceae species and intergenome collinearity analysis of the closely related outgroup species *T. cacao* using MCScanX (Wang et al., 2012). The syntenic blocks were identified based on all-versus-all BLAST alignment in the JCVI package (Tang et al., 2008b) with the parameter "—minspan = 30". We then extracted chromosome arrangements and calculated the numbers of fissions and fusions based on visualizations and phylogeny as follows: chromosomal rearrangement events were identified in the key chromosomes chr5, chr6, and chr7 of the Dipterocarpeae tribe using the *T. cacao* outgroup as the ancestral species, and chromosomal rearrangement events were identified during the evolution of 11-chromosome species into 7-chromosome species using *D. gracilis* as the ancestral species.

Phylogenetic analysis

We analyzed protein-coding genes from 18 species: *A. trichopoda*, *A. thaliana*, *A. sinensis*, *T. cacao*, *O. sativa*, and our assembled 13 genomes. We performed an all-against-all comparison using BLASTP (Altschul and Gish, 1996) with a filter parameter ($E < 1 \times 10^{-5}$). The OrthoMCL (Li et al., 2003) method was used to cluster the BLASTP results into paralogous and orthologous clusters. These analyses resulted in 47 976 gene families containing 656 565 genes from the 18 species. We identified 429 shared single-copy gene families among the 18 species and constructed a phylogenetic tree. First, proteins of the single-copy gene families were aligned using MUSCLE (Edgar, 2004), and a phylogenetic tree was constructed using MrBayes (v.3.1.2) (Huelsenbeck and Ronquist, 2001). Divergence times were estimated using MCMCTree in PAML (v.3.15; options: correlated molecular clock, JC69 model) (http://abacus.gene.ucl.ac.uk/software/paml.html). After a burn-in of 10 000 iterations, the MCMC process was performed 100 000 times with a sample frequency of 100. As fossils play an important role in divergence time estimation, we used first fossil appearances to constrain the minimum age of a clade. A recent study identified eight fossil pollen types belonging to six living genera of Dipterocarpoideae from Sudan and India, with some fossils being the oldest found in their genus (Bansal et al., 2022). We used three fossils from Mahi to constrain our estimation procedure, i.e., minimum ages of 54 mya for *Shorea*, 54 mya for *Vatica*, and 68.5 mya for *Dipterocarpus* (Bansal et al., 2022). We also used the following age constraints in our estimation process: age of crown-group Angiospermae of 167–199 mya based on fossil records and molecular data (Bell et al., 2010) and minimum age of 113 mya for Malvales (Vega et al., 2006).

Gene family expansion and contraction

To investigate gene family evolution during early diversification, we selected five Dipterocarpoideae species (*H. mollissima*, *S. henryana*, *P. chinensis*, *D. gracilis*, and *V. rassak*, representing *Hopea*, *Shorea*, *Parashorea*, *Dipterocarpus*, and *Vatica* genera, respectively) and five closely related species (*T. cacao*, *A. sinensis*, *A. thaliana*, *O. sativa*, and *A. trichopoda*) to identify orthologous gene families. We then measured the expansion and contraction of orthologous gene families using CAFÉ v.4.2 (https://github.com/hahnlab/CAFE) (De Bie et al., 2006) with a probabilistic graphical model. This program uses a birth and death process to model gene gain and loss over a phylogeny. For significantly expanded gene families in the Dipterocarpoideae species, the R package clusterProfiler was used to analyze GO term enrichment (Yu et al., 2012).

Identification of positively selected genes and rapidly evolving genes

To identify PSGs and REGs in the Dipterocarpoideae lineage, we used OrthoMCL (Li et al., 2003) to identify homologous gene clusters among the most closely related species (*T. cacao*) and the 13 Dipterocarpoideae. Orthologous groups with single-copy genes shared by all species were used for further analysis. Multiple sequence alignment was performed for each orthologous group using MUSCLE (Edgar, 2004) with default parameters, and the phylogenetic tree was constructed using MrBayes (v.3.1.2) (Huelsenbeck and Ronquist, 2001). We used the single-copy genes and tree topology for identification of PSGs and REGs using the Codeml program in the PAML package (v.4.8) (Yang, 2007). We also computed *p* values based on chi-square statistics, and genes with $p < 0.05$ were treated as candidate genes under positive selection.

Analysis of gene families related to cellulose synthase (*CesA*), lignin biosynthesis, and cytokinin oxidase/dehydrogenase (*CKX*)

A hidden Markov model (PF00535, PF03552) was used to identify *CesA/CSL* family members in the Dipterocarpoideae species with HMMER v.3.0 (Finn et al., 2011). All obtained proteins without four conserved motifs (QxxRW [Q, R, and W represent glutamine, arginine, and tryptophan, respectively, and x represents any amino acid], DD, DCD, and TED) (Morgan et al., 2013) were removed. In addition, we obtained *A. thaliana CesA/CSL* protein sequences from previous research and classified the Dipterocarpoideae *CesA/CSL* protein sequences based on their phylogenetic relationships with those of *A. thaliana* (Little et al., 2018). For lignin-related genes, we first collected homologous sequences of phenylpropanoid-lignin biosynthesis genes in *Arabidopsis*, rice, and poplar and then aligned the predicted amino acid sequences to the genomes using BLASTP with $E < 1 \times 10^{-10}$. Gene hits that covered more than 200 amino acids with at least 50% protein sequence identity in the alignment were considered to be candidate genes. The hidden Markov model profiles of the cytokine-binding domain (PF09265) and

FAD-binding 4 (PF01565) were used to identify *CKX* sequences with HMMER v.3.0 (Finn et al., 2011). Multiple sequence alignments were generated with MUSCLE, and maximum-likelihood phylogenetic trees were constructed using MEGA (v.7.0) (Kumar et al., 2016). We used MCScanX to identify proteins in the syntenic blocks of each family.

Functional identification of borneol synthase genes

Candidate TPS genes from *Dipterocarpus* species were identified by BLAST using filter parameters (homology identity >30, $E < 1 \times 10^{-5}$) with known TPS genes of *A. thaliana* as query sequences. Two Pfam domains (PF01397 and PF03936) were used to search against protein sequence sets of the four *Dipterocarpus* species using hmmsearch in HMMER. To obtain BPPS candidate genes, we filtered sequences that did not contain three motifs (RRX8W, DDXXD, and NSE/DTE) as described in a previous study (Wang et al., 2018).

Candidate TPS proteins were synthesized and cloned into the pET-28a vector. All TPS proteins were expressed in the *E. coli* BL21(DE3) strain. We purified the proteins using a nickel-affinity chromatography column and determined protein concentrations using a BCA Protein Assay Reagent Kit (Pierce, USA). Monoterpene standards (α-pinene, β-pinene, α-phellandrene, limonene, borneol, and α-terpineol) were purchased from Sigma–Aldrich (UK). The mixed standard was prepared by blending the monoterpene standards at an equimolar ratio.

In vitro enzyme assays were performed following previously described methods (Wang et al., 2018). Each enzyme assay used a 200-μl reaction mixture comprising 50 mM Tris–HCl (pH 7.2), 10 mM $MgCl_2$, 5 mM DTT, 1 mM PMSF, and 3 mM GPP. The catalytic reaction was started by the addition of 1 μg/μl purified protein, and the reaction mixtures were then incubated at 30°C for 15 min. The reactions were terminated at 80°C for 5 min and then quenched on ice. To promote dephosphorylation of the products, 3 μl of calf intestinal alkaline phosphatase (TaKaRa, Japan) was added, followed by 1 h of incubation at 37°C. We added 300 μl of hexane, vortexed the mixture for 15 min to extract the monoterpenoids, and then performed GC-MS analysis.

GC-MS analysis was carried out on an Agilent 7890A gas chromatograph system coupled with a quadrupole time-of-flight mass spectrometer and an inert electron ionization ion source (Agilent, USA). A DB-5MS capillary column coated with 5% diphenyl cross-linked with 95% dimethylpolysiloxane (30 m × 250 μm inner diameter, 0.25-μm film thickness; J&W Scientific, USA) was used. The oven temperature program was as follows: 50°C (2 min), increase to 180°C (5 min) at 5°C min^{-1}, followed by an increase of 10°C min^{-1} to 230°C. The injection, transfer line, and ion source temperatures were 250°C, 290°C, and 230°C, respectively. The sample injection volume was 1 μl in split mode with a 10:1 ratio, and mass spectrometry data were acquired in full-scan mode with an *m/z* range of 35–650 at a rate of 5 spectra/s after a solvent delay of 7.5 min.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Plant Communications Online*.

ACCESSION NUMBERS

The 13 Dipterocarpoideae genome sequences and raw data used in this study were deposited at the China National GeneBank Database (CNGBdb, https://db.cngb.org/) under accession no. CNP0002104.

FUNDING

This work was supported by the "Thousand Talents Plan" (5113190037), the Talents Team Construction Fund of Northwestern Polytechnical University (NWPU), and the Fundamental Research Funds for the Central Universities (3102019JC007) to J.C.

AUTHOR CONTRIBUTIONS

J.C., J.H., and H.J. initiated the genome sequencing project. Z.T. collected materials for the project. Z.T. and X.X. contributed to plant collection. Z.T., P.Z., and T.Z. contributed to genome assembly and annotation. Z.T., P.Z., and Y.H. performed gene family and phylogenomic analyses. Z.T. and P.Z. performed genome structure, whole-genome duplication, and other analyses. X.L. and Q.L. performed functional experiments and GC-MS analyses. J.C., J.H., and Z.T. wrote the manuscript. Y.H., Y.P., Y.Z., Y.Y., Q.L., H.Z., X.L., and F.Z. were involved in improvement of the manuscript.

ACKNOWLEDGMENTS

We thank Profs. Jin Chen and Yaowu Xing from the Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences for collecting materials. We thank Prof. Wen Wang from Northwestern Polytechnical University for suggestions and discussions on the project design. We also thank Haikuan Zhang from Berry Genomics for suggestions on chromosome evolution analysis. No conflict of interest is declared.

Received: March 7, 2022
Revised: September 26, 2022
Accepted: October 21, 2022
Published: October 27, 2022

REFERENCES

Altschul, S.F., and Gish, W. (1996). Local alignment statistics. In Methods in Enzymology (Elsevier), pp. 460–480.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

Appanah, S., and Turnbull, J.M. (1998). A Review of Dipterocarps: Taxonomy, Ecology, and Silviculture.

Ashton, P.S., and Gunatilleke, C.V.S. (1987). New light on the plant geography of Ceylon. I. Historical plant geography. J. Biogeogr. **14**:249–285.

Ashton, P.S. (1988). Dipterocarp biology as a window to the understanding of tropical forest structure. Annu. Rev. Ecol. Syst. **19**:347–370.

Ashton, P.S., and Arboretum, A. (1979). Dipterocarpaceae. Flora Malesiana-Series 1. Spermatophyta **9**:237–552.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. **28**:45–48.

Bakkali, F., Averbeck, S., Averbeck, D., and Idaomar, M. (2008). Biological effects of essential oils–a review. Food Chem. Toxicol. **46**:446–475.

Bansal, M., Morley, R.J., Nagaraju, S.K., Dutta, S., Mishra, A.K., Selveraj, J., Kumar, S., Niyolia, D., Harish, S.M., Abdelrahim, O.B., et al. (2022). Southeast Asian Dipterocarp origin and diversification driven by Africa-India floristic interchange. Science **375**:455–460.

Bawa, K. (1998). In Conservation of Genetic Resources in the Dipterocarpaceae. A Review of Dipterocarps: Taxonomy, Ecology

and Silviculture, S. Appanah and J.M. Turnbull, eds. (Bogor, Indonesia: CIFOR), pp. 45–55.

Bell, C.D., Soltis, D.E., and Soltis, P.S. (2010). The age and diversification of the angiosperms re-revisited. Am. J. Bot. **97**:1296–1303.

Bellard, C., Leclerc, C., Leroy, B., Bakkenes, M., Veloz, S., Thuiller, W., and Courchamp, F. (2014). Vulnerability of biodiversity hotspots to global change. Global Ecol. Biogeogr. **23**:1376–1386.

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

Burt, S. (2004). Essential oils: their antibacterial properties and potential applications in foods—a review. Int. J. Food Microbiol. **94**:223–253.

Cao, C.-P., Gailing, O., Siregar, I., Indrioko, S., and Finkeldey, R. (2006). Genetic variation at AFLPs for the Dipterocarpaceae and its relation to molecular phylogenies and taxonomic subdivisions. J. Plant Res. **119**:553–558.

Cao, Z. (2004). Chinese Obstetrics and Gynecology (Beijing: People's Medical Publishing House), pp. 2838–2862.

Chalopin, D., Naville, M., Plard, F., Galiana, D., and Volff, J.-N. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biol. Evol. **7**:567–580.

Chen, F., Tholl, D., Bohlmann, J., and Pichersky, E. (2011). The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. Plant J. **66**:212–229.

Chen, N. (2004). Using Repeat Masker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinf. **5**:4.10.11–14.10.14.

Chénais, B., Caruso, A., Hiard, S., and Casse, N. (2012). The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. Gene **509**:7–15.

Dassanayake, M., Oh, D.-H., Haas, J.S., Hernandez, A., Hong, H., Ali, S., Yun, D.-J., Bressan, R.A., Zhu, J.-K., Bohnert, H.J., et al. (2011). The genome of the extremophile crucifer Thellungiella parvula. Nat. Genet. **43**:913–918.

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.

Didi, V., Jackson, P., and Hejátko, J. (2015). Hormonal regulation of secondary cell wall formation. J. Exp. Bot. **66**:5015–5027.

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., 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 Syst. **3**:99–101.

Dyrmose, 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. Hist. Bull. Siam Soc. **62**.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinf. **5**:113–119.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinf. **9**:18–24.

Falara, V., Akhtar, T.A., Nguyen, T.T.H., Spyropoulou, E.A., Bleeker, P.M., Schauvinhold, I., Matsuba, Y., Bonini, M.E., Schilmiller, A.L., Last, R.L., et al. (2011). The tomato terpene synthase gene family. Plant Physiol. **157**:770–789.

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

Ghazoul, J. (2016). Dipterocarp Biology, Ecology, and Conservation (Oxford University Press).

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., et al. (2012). Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. **40**:D1178–D1186.

Grover, C.E., and Wendel, J.F. (2010). Recent insights into mechanisms of genome size change in plants. J. Bot. **2010**:1–8.

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

Heckenhauer, J., Samuel, R., Ashton, P.S., Abu Salim, K., and Paun, O. (2018). Phylogenomics resolves evolutionary relationships and provides insights into floral evolution in the tribe Shoreeae (Dipterocarpaceae). Mol. Phylogenet. Evol. **127**:1–13.

Heckenhauer, J., Samuel, R., Ashton, P.S., Turner, B., Barfuss, M.H.J., Jang, T.-S., Temsch, E.M., Mccann, J., Salim, K.A., Attanayake, A.M.A.S., et al. (2017). Phylogenetic analyses of plastid DNA suggest a different interpretation of morphological evolution than those used as the basis for previous classifications of Dipterocarpaceae (Malvales). Bot. J. Linn. Soc. **185**:1–26.

Higuchi, T. (2006). Look back over the studies of lignin biochemistry. J. Wood Sci. **52**:2–8.

Huelsenbeck, J.P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**:754–755.

Hurd, M.C., Kwon, M., and Ro, D.-K. (2017). Functional identification of a Lippia dulcis bornyl diphosphate synthase that contains a duplicated, inhibitory arginine-rich motif. Biochem. Biophys. Res. Commun. **490**:963–968.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449**:463–467.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature **473**:97–100.

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J.E., McKain, M.R., McNeal, J., Rolf, M., Ruzicka, D.R., Wafula, E., Wickett, N.J., et al. (2012). A genome triplication associated with early diversification of the core eudicots. Genome Biol. **13**:R3–R14.

Kaloustian, J., Chevalier, J., Mikail, C., Martino, M., Abou, L., and Vergnes, M.-F. (2008). Étude de six huiles essentielles: composition chimique et activité antibactérienne. Phytothérapie **6**:160–164.

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

Kato, H., Ishizaki, K., Kouno, M., Shirakawa, M., Bowman, J.L., Nishihama, R., and Kohchi, T. (2015). Auxin-mediated transcriptional system with a minimal set of components is critical for morphogenesis through the life cycle in Marchantia polymorpha. PLoS Genet. **11**:e1005084.

Kent, W.J. (2002). BLAT—the BLAST-like alignment tool. Genome Res. **12**:656–664.

Kim, S., Park, J., Yeom, S.-I., Kim, Y.-M., Seo, E., Kim, K.-T., Kim, M.-S., Lee, J.M., Cheong, K., Shin, H.-S., et al. (2017). New reference genome sequences of hot pepper reveal the massive evolution of

plant disease-resistance genes by retroduplication. Genome Biol. **18**:210–211.

Klein, S.J., and O'Neill, R.J. (2018). Transposable elements: genome innovation, chromosome diversity, and centromere conflict. Chromosome Res. **26**:5–23.

Kulkarni, M., Sawant, N., Kolapkar, A., Huprikar, A., and Desai, N. (2021). Borneol: a Promising monoterpenoid in enhancing drug delivery across various Physiological Barriers. AAPS PharmSciTech **22**:1–17.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. **33**:1870–1874.

Lei, D., Qiu, Z., Wu, J., Qiao, B., Qiao, J., and Zhao, G.-R. (2021). Combining metabolic and monoterpene synthase engineering for de novo production of monoterpene alcohols in Escherichia coli. ACS Synth. Biol. **10**:1531–1544.

Lewis, S.L. (2006). Tropical forests and the changing earth system. Philos. Trans. R. Soc. Lond. B Biol. Sci. **361**:195–210.

Li, L., Stoeckert, C.J., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. **13**:2178–2189.

Li, S., and Mu, B.C.G. (2004). People's Medical Publishing House. Beijing, China.

Little, A., Schwerdt, J.G., Shirley, N.J., Khor, S.F., Neumann, K., O'Donovan, L.A., Lahnstein, J., Collins, H.M., Henderson, M., Fincher, G.B., et al. (2018). Revised phylogeny of the cellulose synthase gene superfamily: insights into cell wall evolution. Plant Physiol. **177**:1124–1141.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D., and Fan, W. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. Preprint at arXiv. https://doi.org/10.48550/arXiv.1308.2012.

Ma, R., Su, P., Guo, J., et al. (2021). Bornyl Diphosphate Synthase From Cinnamomum burmanni and Its Application for (+)-Borneol Biosynthesis in Yeast. Frontiers in bioengineering and biotechnology **9**:631863.

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

Meents, M.J., Watanabe, Y., and Samuels, A.L. (2018). The cell biology of secondary cell wall biosynthesis. Ann. Bot. **121**:1107–1125.

Mellerowicz, E.J., and Sundberg, B. (2008). Wood cell walls: biosynthesis, developmental dynamics and their implications for wood properties. Curr. Opin. Plant Biol. **11**:293–300.

Morgan, J.L.W., Strumillo, J., and Zimmer, J. (2013). Crystallographic snapshot of cellulose synthesis and membrane translocation. Nature **493**:181–186.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

Ng, C.H., Lee, S.L., Tnah, L.H., Ng, K.K.S., Lee, C.T., and Madon, M. (2016). Genome size variation and evolution in Dipterocarpaceae. Plant Ecol. Divers. **9**:437–446.

Ng, K.K.S., Kobayashi, M.J., Fawcett, J.A., et al. (2021). The genome of Shorea leprosula (Dipterocarpaceae) highlights the ecological relevance of drought in aseasonal tropical rainforests. Commun. Biol. **4**:1–14.

Novák, P., Guignard, M.S., Neumann, P., Kelly, L.J., Mlinarec, J., Koblížková, A., Dodsworth, S., Kovařík, A., Pellicer, J., Wang, W., et al. (2020). Repeat-sequence turnover shifts fundamentally in species with large genomes. Nat. Plants **6**:1325–1329.

O'Brien, M.J., Leuzinger, S., Philipson, C.D., Tay, J., and Hector, A. (2014). Drought survival of tropical tree seedlings enhanced by non-structural carbohydrate levels. Nat. Clim. Chang. **4**:710–714.

O'Brien, T.E., Bertolani, S.J., Zhang, Y., Siegel, J.B., and Tantillo, D.J. (2018). Predicting productive binding modes for substrates and carbocation intermediates in terpene synthases—bornyl diphosphate synthase as a representative case. ACS Catal. **8**:3322–3330.

Polko, J.K., and Kieber, J.J. (2019). The regulation of cellulose biosynthesis in plants. Plant Cell **31**:282–296.

Popper, Z.A., Michel, G., Hervé, C., Domozych, D.S., Willats, W.G.T., Tuohy, M.G., Kloareg, B., and Stengel, D.B. (2011). Evolution and diversity of plant cell walls: from algae to flowering plants. Annu. Rev. Plant Biol. **62**:567–590.

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

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic Acids Res. **33**:W116–W120.

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 (Berl.) **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.

Roach, M.J., Schmidt, S.A., and Borneman, A.R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinf. **19**:1–10.

Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: assessing genome assembly and annotation completeness. In Gene Prediction (Springer), pp. 227–245.

Sodhi, N.S., Posa, M.R.C., Lee, T.M., Bickford, D., Koh, L.P., and Brook, B.W. (2010). The state and conservation of Southeast Asian biodiversity. Biodivers. Conserv. **19**:317–328.

Soltis, P.S., and Soltis, D.E. (2016). Ancient WGD events as drivers of key innovations in angiosperms. Curr. Opin. Plant Biol. **30**:159–165.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics **24**:637–644.

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

Stone, B.C. (1983). Dipterocarpaceae. JSTOR.

Sun, P., Jiao, B., Yang, Y., Shan, L., Li, T., Li, X., Xi, Z., Wang, X., and Liu, J. (2021). WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. Preprint at bioRxiv. https://doi.org/10.1101/2021.04.29.441969.

Symington, C.F., Ashton, P.S., and Appanah, S. (2004). Foresters' Manual of Dipterocarps (Forest Research Institute Malaysia).

Tang, H., Wang, X., Bowers, J.E., Ming, R., Alam, M., and Paterson, A.H. (2008a). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res. **18**:1944–1954.

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

Tang, L., Liao, X., Tembrock, L.R., Ge, S., and Wu, Z. (2022). A chromosome-scale genome and transcriptomic analysis of the endangered tropical tree Vatica mangachapoi (Dipterocarpaceae). DNA Res. **29**:dsac005.

Vega, F.J., García-Barrera, P., Perrilliat, M.d.C., Coutiño, M.A., and Mariño-Pérez, R. (2006). El Espinal, a new plattenkalk facies locality

from the lower Cretaceous Sierra Madre formation, Chiapas, southeastern Mexico. Rev. Mex. Ciencias Geol. **23**:323–333.

Wang, H., Ma, D., Yang, J., Deng, K., Li, M., Ji, X., Zhong, L., and Zhao, H. (2018). An integrative volatile terpenoid profiling and transcriptomics analysis for gene mining and functional characterization of AvBPPS and AvPS involved in the monoterpenoid biosynthesis in Amomum villosum. Front. Plant Sci. **9**:846.

Wang, S., Liang, H., Wang, H., Li, L., Xu, Y., Liu, Y., Liu, M., Wei, J., Ma, T., Le, C., et al. (2022). The chromosome-scale genomes of Dipterocarpus turbinatus and Hopea hainanensis (Dipterocarpaceae) provide insights into fragrant oleoresin biosynthesis and hardwood formation. Plant Biotechnol. J. **20**:538–553.

Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.-h., Jin, H., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. **40**:e49.

Wendel, J.F., Jackson, S.A., Meyers, B.C., and Wing, R.A. (2016). Evolution of plant genome architecture. Genome Biol. **17**:37–44.

Whittington, D.A., Wise, M.L., Urbansky, M., Coates, R.M., Croteau, R.B., and Christianson, D.W. (2002). Bornyl diphosphate synthase: structure and strategy for carbocation manipulation by a terpenoid cyclase. Proc. Natl. Acad. Sci. USA **99**:15375–15380.

Xie, T., Zheng, J.-F., Liu, S., Peng, C., Zhou, Y.-M., Yang, Q.-Y., and Zhang, H.-Y. (2015). De novo plant genome assembly based on chromatin interactions: a case study of Arabidopsis thaliana. Mol. Plant **8**:489–492.

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.

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24**:1586–1591.

Yang, Z., An, W., Liu, S., Huang, Y., Xie, C., Huang, S., and Zheng, X. (2020). Mining of candidate genes involved in the biosynthesis of dextrorotatory borneol in Cinnamomum burmannii by transcriptomic analysis on three chemotypes. PeerJ **8**:e9311.

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

Yu, X.-J., Zheng, H.-K., Wang, J., Wang, W., and Su, B. (2006). Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. Genomics **88**:745–751.

Zhang, Q.-L., Fu, B.M., and Zhang, Z.-J. (2017). Borneol, a novel agent that improves central nervous system drug delivery by enhancing blood–brain barrier permeability. Drug Deliv. **24**:1037–1044.

Zhang, Z., Li, J., Zhao, X.-Q., Wang, J., Wong, G.K.-S., and Yu, J. (2006). KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Dev. Reprod. Biol. **4**:259–263.

Zou, L., Zhang, Y., Li, W., Zhang, J., Wang, D., Fu, J., and Wang, P. (2017). Comparison of chemical profiles, anti-inflammatory activity, and UPLC-Q-TOF/MS-based metabolomics in endotoxic fever rats between synthetic borneol and natural borneol. Molecules **22**:1446.