Stout camphor tree genome fills gaps in understanding of flowering plant genome and gene family evolution.

Shu-Miaw Chaw^{¶*}, Yu-Ching Liu¹, Han-Yu Wang¹, Yu-Wei Wu², Chan-Yi Ivy Lin¹, Chung-Shien Wu¹, Huei-Mien Ke¹, Lo-Yu Chang^{1,3}, Chih-Yao Hsu¹, Hui-Ting Yang¹, Edi Sudianto¹, Ming-Hung Hsu^{1,4}, Kun-Pin Wu⁴, Ning-Ni Wang¹, Jim Leebens-Mack⁵ and Isheng. J. Tsai^{¶*}

¹Biodiversity Research Center, Academia Sinica, Taipei 11529, Taiwan

²Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei 11031, Taiwan

³School of Medicine, National Taiwan University, Taipei 10051, Taiwan

⁴Institute of Biomedical Informatics, National Yang-Ming University, Taipei 11221, Taiwan

⁵Plant Biology Department, University of Georgia, Athens, GA30602, USA

[¶] These authors contributed equally to this work

Correspondence: Shu-Miaw Chaw (smchaw@sinica.edu.tw) and Isheng Jason Tsai (ijtsai@sinica.edu.tw)

Abstract

We present reference-quality genome assembly and annotation for the stout camphor tree (SCT; *Cinnamomum kanehirae* [Laurales, Lauraceae]), the first sequenced member of the Magnoliidae comprising four orders (Laurales, Magnoliales, Canellales, and Piperales) and over 9,000 species. Phylogenomic analysis of 13 representative seed plant genomes indicates that magnoliid and eudicot lineages share more recent common ancestry relative to monocots. Two whole genome duplication events were inferred within the magnoliid lineage, one before divergence of Laurales and Magnoliales and the other within the Lauraceae. Small scale segmental duplications and tandem duplications also contributed to innovation in the evolutionary history of *Cinnamomum*. For example, expansion of terpenoid synthase subfamilies within the Laurales spawned the diversity of *Cinnamomum* monoterpenes and sesquiterpenes.

Introduction

Aromatic medicinal plants have long been utilized as spices or curative agents throughout human history. In particular, many commercial essential oils are derived from flowering plants in the tree genus *Cinnamomum* L. (Lauraceae)¹⁻³. For example, camphor, a bicyclic monoterpene ketone (C₁₀H₁₆O) that can be obtained from many members of this genus, has important industrial and pharmaceutical applications⁴. *Cinnamomum* includes approximately 250 species of evergreen aromatic trees belonging to Lauraceae (laurel family), which is an economically and ecologically important family that includes 2,850 species distributed mainly in tropical and subtropical regions of Asia and South America⁵. Among them, avocado (*Persea americana*), bay laurel (*Laurus nobilis*), camphor tree or camphor laurel (*C. camphora*), cassia (*C. cassia*), and cinnamon (including several *C. spp.*) are important spice and fruit species. Lauraceae has traditionally been classified as one of the seven families of Laurales, which together with Canellales, Piperales and Magnoliales constitute the Magnoliidae (“magnoliids” informally).

The magnoliids, containing about 9,000 species, are characterized by 3-merous flowers with diverse volatile secondary compounds, 1-pored pollen, and insect-pollination⁶. Many magnoliids – such as custard apple (*Annonaceae*), nutmeg (*Myristica*), black pepper (*Piper nigrum*), magnolia, and tulip tree (*Liriodendron tulipifera*) – produce economically important fruits, spices, essential oils, drugs, perfumes, timber, and horticultural ornamentals. The phylogenetic position of magnoliids, however, has been uncertain. Further, there are also unresolved questions about genome evolution within the Magnoliidae. Analysis of transcriptome sequences has implicated two rounds of genome duplication in the ancestry of *Persea* (Lauraceae) and one in the ancestry of *Liriodendron* (Magnoliaceae)⁷, but the relative timing of these events remains ambiguous.

Cinnamomum kanehirae, commonly known as the stout camphor tree (SCT), a name referring to its bulky, tall and strong trunk, is endemic to Taiwan and under threat of extinction. It has a restricted distribution in broadleaved forests in an elevational band between 450 and 1,200 meters⁸. *Cinnamomum*, including SCT and six congeneric species contributed to Taiwan’s position as the largest producer and exporter of camphor in the 19th century, and its value was further enhanced due to its valuable

wood, with trunks exhibiting the largest diameters among flowering plants of Taiwan, and aromatic, decay-resistance attributed to the essential oil D-terpinenol⁹. *Antrodia cinnamomea*, a parasitic fungus that infects the trunks of SCT causing heart rot¹⁰. The fungus produces several medicinal triterpenoids that impede the growth of liver cancer cells^{10,11} and act as antioxidants that protect against atherosclerosis¹². Due to intensive deforestation in the past half century, followed by poor seed germination and illegal logging to cultivate the fungus, natural populations of SCT are fragmented and threatened^{13,14}.

Here we report a chromosome-level genome assembly of SCT. Comparative analyses of the SCT genome with those of 10 other angiosperms and two gymnosperms (ginkgo and Norway spruce) allow us to resolve the phylogenetic position of the magnoliids and shed new light on flowering plant genome evolution. Several gene families appear to be uniquely expanded in the SCT lineage, including the terpenoid synthase superfamily. Terpenoids play vital primary roles as photosynthetic pigments (carotenoids), electron carriers (plastoquinone and ubiquinone side chains), and regulators of plant growth (the phytohormone gibberellin and phytol side chain in chlorophyll)¹⁵. Specialized volatile or semi-volatile terpenoids are also important biological and ecological signals that protect plants against abiotic stress and promote beneficial biotic interactions above and below ground with pollinators, pathogens, herbivorous insect, and soil microbes¹⁵⁻¹⁸. Analyses of the SCT genome inform understanding of gene family evolution contributing to terpenoid biosynthesis, shed light on early events in flowering plant diversification, and provide new insights into the demographic history of SCT with important implications for future conservation efforts.

Results

Assembly and annotation of SCT

SCT is diploid (2n=24; Supplementary Fig. 1a) with an estimated genome size of 800 to 846 Mb (Supplementary Figs. 1b, 2). An initial assembly with 141x and 50x Illumina paired-end and mate-pair reads, respectively (Supplementary Table 1), produced 48,650 scaffolds spanning 714.7 Mb (scaffold N50 = 594 kb and N90 = 3 kb; Table 1). A second, long-read assembly derived solely from 85x Pacbio long reads (read N50 = 11.1 kb; contig N50 = 0.9 Mb) was scaffolded with 207x “Chicago” reconstituted-chromatin and 204x Hi-C paired-end reads using the HiRise pipeline¹⁹ (Table 1; Supplementary Fig. 3). A final, integrated assembly of 730.7 Mb was

113 produced in 2,153 scaffolds, comprising 91.3% of the flow cytometry genome size
 114 estimate. The final scaffold N50 was 50.4 Mb with more than 90% in 12
 115 pseudomolecules, presumably corresponding to the 12 SCT chromosomes.
 116 Using a combination of reference plant protein homology support and transcriptome
 117 sequencing derived from a variety of tissues (Supplementary Fig. 1c and Table 2) and
 118 *ab initio* gene prediction, 27,899 protein-coding genes models were annotated using
 119 the MAKER2 pipeline²⁰ (Table 1). Of these, 93.7% were found to be homologous to
 120 proteins in the TrEMBL database and 50% could be assigned gene ontology terms
 121 using eggNOG-mapper²¹. The proteome was estimated to be at least 89% complete
 122 based on BUSCO²² (Benchmarking Universal Single-Copy Orthologs) assessment
 123 which is comparable to other sequenced plant species (Supplementary Table 3).
 124 Orthofinder²³ clustering of SCT gene models with those from twelve diverse seed
 125 plant genomes yielded 20,658 orthologous groups (OGs) (Supplementary Table 4).
 126 24,148 SCT genes (85.8%) were part of OGs with orthologues from at least one other
 127 plant species. 3,744 gene models were not orthologous to others, and only 210 genes
 128 were part of the 48 SCT specific OGs. Altogether, they suggest that the phenotypic
 129 diversification in magnoliids may be fueled by *de novo* birth of species-specific genes
 130 as well as expansion of existing gene families.

131

132 **Genome characterization**

133 We identified 3,950,027 bi-allelic heterozygous sites in the SCT genome,
 134 corresponding to an average heterozygosity of 0.54% (one heterozygous SNP per 185
 135 bp). The minor allele frequency of these sites had a major peak around 50% consistent
 136 with the fact that SCT is diploid with no evidence for recent aneuploidy
 137 (Supplementary Fig. 4). The spatial distribution of heterozygous sites was highly
 138 variable with 23.9% of the genome exhibiting less than 1 SNP loci per kb compared to
 139 10% of the genome with at least 12.6 SNP loci per kb. Runs of homozygosity (ROH)
 140 regions appeared to be distributed randomly across SCT chromosomes reaching a
 141 maximum of 20.2 Mb in scaffold 11 (Fig. 1a). Such long ROH regions may be
 142 associated with selective sweeps, inbreeding or recent population bottlenecks.
 143 Pairwise sequentially Markovian coalescent²⁴ (PSMC) analysis based on
 144 heterozygous SNP densities implicated a continuous reduction of effective population
 145 size over the last 9 Ma (Fig. 1b) with a possible bottleneck coincident with the
 146 mid-Pleistocene climatic shift at 0.9 Ma. Such patterns may reflect a complex
 147 population history of SCT associated with the geologic history of Taiwan including
 148 uplift and formation of the island in the late Miocene (9 Ma) followed by mountain
 149 building 5–6 Ma, respectively²⁵.

150

Transposable elements (TEs) and interspersed repeats made up 48% of the genome assembly (Supplementary Table 5). The majority of the TEs belonged to LTR retrotransposons (25.53%), followed by DNA transposable elements (12.67%). Among the LTR, 40.75% and 23.88% of retrotransposons belonged to Ty3/Gypsy and Ty1/Copia, respectively (Supplementary Table 5). Phylogeny of reverse transcriptase domain showed that the majority of Ty3/Gypsy copies formed a distinct clade (20,092 copies) presumably as a result of recent expansion and proliferation, while Ty1/Copia elements were grouped into two sister clades (7,229 and 2,950 copies; Supplementary Fig. 5). With the exception of two scaffolds, both Ty3/Gypsy and Ty1/Copia LTR TEs were clustered within the pericentromeric centers of the 12 largest scaffolds (Fig 2; Supplementary Fig. 6). Additionally, the LTR enriched regions (defined by 100 kb with excess of 50% comprising LTR class TEs) had on average 35% greater coverage than rest of the genome (Fig 2; Supplementary Fig. 7), suggesting that these repeats were collapsed in the assembly and may have contributed to the differences in flow cytometry and k-mer genome size estimates. The coding sequence content of SCT is similar to the other angiosperm genomes included in our analyses (Supplementary Table 3), while introns are slightly longer in SCT due to a higher density of TEs ($P < 0.001$, Wilcoxon rank sum test; Supplementary Fig. 8).

As has been described for other plant genomes²⁶, the chromosome-level scaffolds of SCT exhibit low protein-coding gene density and high TE density in the centers of chromosomes, and increased gene density towards the chromosome ends (Fig. 2). We identified clusters of putative subtelomere heptamer TTTAGGG extending as long as 2,547 copies, which implicate telomeric repeats in plants²⁷ (Supplementary Table 6). Additionally, 687 kb of nuclear plastid DNAs (NUPT) averaging around 202.8 bp were uncovered (Supplementary Table 7). SCT NUPTs were overwhelmingly dominated by short fragments with 96% of the identified NUPTs less than 500 bp (Supplementary Table 8). The longest NUPT is ~20 kb in length and syntenic with 99.7% identity to a portion of the SCT plastome that contains seven protein-coding and five tRNA genes (Supplementary Fig. 9).

Phylogenomic placement of *C. kanehirae* sister to eudicots

The magnoliids have been hypothesized as the sister lineage to (1) the Chloranthaceae, (2) a clade including eudicots, Chloranthaceae, Ceratophyllaceae, (3) the monocots, (4) a monocot + eudicot clade, or (5) a Chloranthaceae + Ceratophyllaceae clade, based on phylogenetic analyses of plastid genes, plastomic IR regions, four mitochondrial genes, inflorescence and floral structures, and low copy nuclear genes^{7,28}. Similar to the APG III, the APG IV system²⁹ placed Magnoliidae and Chloranthaceae together

as sister to a robust clade comprising monocots and Ceratophyllales + eudicots. To resolve the long-standing debate over the phylogenetic placement of magnoliids relative to other major flowering plant lineages, we constructed a phylogenetic tree based on 211 strictly single copy orthologue sets shared among the 13 genomes included in our analyses. A single species tree was recovered through maximum likelihood analysis³⁰ of a concatenated supermatrix of the single copy gene alignments and coalescent-based analysis using the 211 gene trees³¹ (Fig. 3; Supplementary Fig. 10). SCT, representing the magnoliid lineage was placed as sister to the eudicot clade (Fig. 3). Using MCMCtree³², we calculated a 95% confidence interval for the time of divergence between magnoliids and eudicots to be 139.41–191.57 million years (Ma; Supplementary Fig. 11), which overlaps with two other recent estimates (114.75–164.09 Ma³³ and 118.9–149.9 Ma³⁴).

Synteny analysis / whole genome duplication (WGD)

Previous investigations of EST data inferred a genome-wide duplication within the magnoliids before the divergence of the Magnoliales and Laurales⁷, but synteny-based testing of this hypothesis has not been possible without an assembled magnoliid genome. A total of 16,498 gene pairs were identified in 992 syntenic blocks comprising 72.7% of the SCT genome assembly. Of these intragenomic syntenic blocks, 72.3% were found to be syntenic to more than one location on the genome, suggesting that more than one WGD occurred in the ancestry of SCT (Fig. 4a). Two rounds of ancient WGD were implicated by extensive synteny between pairs of chromosomal regions and significantly but less syntenic paring of each region with two additional genomic segments (Supplementary Fig. 12). Synteny blocks of SCT's 12 largest scaffolds were assigned to five clusters that may correspond to pre-WGD ancestral chromosomes (Fig. 4a; Supplementary Fig. 12 and Note).

Amborella trichopoda is the sole species representing the sister lineage to all other extant angiosperms, and it has no evidence of WGD since divergence from the last common ancestor extant flowering plant lineages³⁵. To confirm two rounds of WGD took place in ancestry of SCT after divergence of lineages leading to SCT and *A. trichopoda*, we assessed synteny between the two genomes. Consistent with our hypothesis, four segments of the SCT genome aligned with a single region in the *A. trichopoda* genome (Fig. 4b; Supplementary Fig. 13).

In order to more precisely infer the timing of the two rounds of WGD evident in the SCT genome, intragenomic and interspecies homolog Ks (synonymous substitutions per synonymous site) distributions were estimated. SCT intragenomic duplicates

showed two peaks around 0.46 and 0.76 (Fig. 5a), congruent with the two WGD events. Based on these two peaks, we were able to infer the karyotype evolution by organizing the clustered synteny blocks further into four groups presumably originating from one of the five pre-WGD chromosomes (Supplementary Fig. 14). Comparison between *Aquilegia coerulea* (Ranunculales, a sister lineage to all other extant eudicots³⁵) and SCT orthologs revealed a prominent peak around $K_s = 1.41$ (Fig. 5a), while the *Aquilegia* intra-genomic duplicate was around $K_s = 1$, implicating independent WGDs following the divergence of lineages leading to SCT and *Aquilegia*. The availability of the transcriptome of 17 Laurales + Magnoliales from 1,000 plants initiative³⁶ allowed us to test the hypothesized timing of the WGDs evident in the SCT genome⁸. K_s distribution of all species from Lauraceae have shown apparent two peaks, but only one peak was observed in other Laurales and Magnoliales samples, suggesting a WGD predating divergence of these two orders followed by a second recent WGD in the early ancestry of the Lauraceae (Fig. 5b). The K_s peak seen in *Aquilegia* data is likely attributable to WGD within the Ranunculales well after the divergence of eudicots and magnoliids (Supplementary Fig. 15).

244

245 **Specialization of the magnoliids proteome**

We sought to identify genes and protein domains specific to SCT by annotating protein family (Pfam) domains³⁷ and assessing their distribution across the 13 seed plant genomes included in our phylogenomic analyses. Consistent with the observation that there were very few SCT-specific OGs, principal component analysis of Pfam domain content clustered SCT with the monocots and eudicots, with the first two principal components separating gymnosperms and *A. trichopoda* from this group (Supplementary Fig. 16a). There were considerable overlaps between SCT, eudicot and monocot species, suggesting significant functional diversification since these three lineages split. SCT also showed a significant enrichment and reduction of 111 and 34 protein domains compared to other plant species, respectively (Supplementary Fig. 16b and Table 9). Gain of protein domains included the terpene synthase C terminal domain involved in defense responses and the leucine-rich repeats (628 vs 334.4) in plant transpiration efficiency³⁸. Interestingly, we found that SCT possesses 21 copies of EIN3/EIN3-like (EIL) transcription factor, more than the previously reported maximum of 17 copies in the banana genome (*Musa acuminata*)³⁹. EILs initiate an ethylene signaling response by activating ethylene response factors (ERF), which we also found to be highly expanded in SCT (150 copies versus an average of 68.3 copies from nine species reported in ref³⁹; Supplementary Fig. 17). Ethylene

signaling in plants was reported to be associated with fruit ripening³⁹ and secondary growth in wood formation⁴⁰ and may be involved in either processes in SCT.

CAFE⁴¹ was used to assess OG expansions and contractions across (Fig. 3) the seed plant phylogeny. Gene family size evolution was dynamic across the phylogeny, and the branch leading to SCT did not exhibit significantly different numbers of expansions and contractions. Enrichment of gene ontology terms revealed either various different gene families sharing common functions or single gene families undergoing large expansions (Supplementary Table 10 and 11). For example, the expanded members of plant resistance (R) genes add up to “plant-type hypersensitive response” (Supplementary Table 10). In contrast, the enriched gene ontology terms from the contracted gene families of SCT branch (Supplementary Table 11) contains members of ABC transporters, indole-3-acetic acid-amido synthetase, xyloglucan endotransglucosylase/hydrolase and auxin-responsive protein, all of which are part of the “response to auxin”.

Resistance (R) genes

The SCT genome annotation included 387 resistance gene models, 82% of which belong to nucleotide-binding site leucine-rich repeat (NBS-LRR) or coiled-coil NBS-LRR (CC-NBS-LRR) types. This result is consistent with a previous report that LRR is one of the most abundant protein domains in plants and it is highly likely that SCT is able to recognize and fight off pathogen products of avirulence (Avr) genes⁴². Among the sampled 13 genomes, SCT harbors the highest number of R genes among non-cultivated plants (Supplementary Fig. 18). The phylogenetic tree constructed from 2,465 NBS domains also suggested that clades within the gene family have diversified independently within the eudicots, monocots and magnoliids. Interestingly, the most diverse SCT NBS gene clades were sister to depauperate eudicot NBS gene clades (Supplementary Fig. 19).

Terpene synthase gene family

One of the most striking features of the SCT genome is the large number of terpene synthase (TPS) genes (*CkTPSs*). A total of 101 *CkTPSs* were predicted and annotated, the largest number for any other genome to date. By including transcriptome dataset of two more species from magnoliids (*Persea americana* and *Saruma henryi*), phylogenetic analyses of TPS from 15 species were performed to place *CkTPSs* among six of seven TPS subfamilies that have been described for seed plants⁴³⁻⁴⁵ (Fig. 6, Table 2 and Supplementary Fig. 20–25). *CkTPS* genes placed in the TPS-c (2) and TPS-e (5)

subfamilies likely encode diterpene synthases such as copalyl diphosphate synthase (CPS) and *ent*-kaurene synthase (KS)⁴⁶. These are key enzymes catalyzing the formation of the 20-carbon isoprenoids (collectively termed diterpenoids; C20), which was thought to be eudicot-specific⁴⁵ and serve primary functions like regulating plant primary metabolism. The remaining 94 predicted *CkTPS*s likely code for the 10-carbon monoterpene (C10) synthases, 15-carbon sesquiterpene (C15) synthases, and additional 20-carbon diterpene (C20) synthases (Table 2). With 25 and 58 homologs, respectively, TPS-a and TPS-b subfamilies are most diverse in SCT, presumably contributing to the mass and mixed production of volatile C15s and C10s⁴⁷. *CkTPS*s are not uniformly distributed throughout the chromosomes (Supplementary Table 12) and clustering of members from individual subfamilies were observed as tandem duplicates (Supplementary Fig. 26). For instance, scaffold 7 contains 29 *CkTPS* genes belonging to several subfamilies including all of the eight *CkTPS-a*, 12 *CkTPS-b*, five *CkTPS-e* and three *CkTPS-f* (Supplementary Fig. 26). In contrast, only two members of *CkTPS-c* reside in scaffold 1. Twenty-four *CkTPS*s locate in other smaller scaffolds, 22 of which code for subfamily TPS-b (Supplementary Fig. 21).

It is noteworthy that the TPS gene tree resolved Lauraceae-specific TPS gene clades within the TPS-a, -b, -f, and -g subfamilies (Supplementary Fig. 20–23). This pattern of TPS gene duplication in a common ancestor of *Persea* and *Cinnamomum* and subsequent retention may indicate subfunctionalization or neofunctionalization of duplicated *TPS*s within the Lauraceae. A magnoliids-specific subclade in the TPS-a subfamily was also identified in analyses including more magnoliid TPS genes with characterized functions (Supplementary Fig. 20). Indeed, we detected positive selection in the Lauraceae-specific TPS-f -I and -II subclades implying functional divergence (Supplementary Table 13). Together, these data suggest increasing diversification of magnoliid TPS genes both before and after the origin of the Lauraceae. The distribution of TPS genes in the SCT genome suggests that both segmental (including WGD) and tandem duplication events contributed to diversification of TPS enzymes in the SCT lineage and the terpenoids they produce.

Discussion

It is now challenging to find a wild SCT population making the conservation and basic study of this tree a priority. SCTs have been intensively logged since the 19th

century initially for hardwood properties and association with fungus *Antrodia cinnamomea*. The apparent runs of homozygosity have been observed due to anthropogenic selective pressures or inbreeding in several livestock⁴⁷, though inbreeding as a result of recent population bottleneck may be a more likely explanation for SCT. Interestingly, continuous decline in effective population size was inferred since 9 Ma. These observations may reflect a complex population history of SCT and Taiwan itself after origination and mountain building of the island that occurred around late Miocene (9 Ma) and 5–6 Ma, respectively²⁵. The availability of the SCT genome will help the development of precise genetic monitoring and tree management for the survival of SCT's natural populations.

The placement of SCT as sister to the eudicots has important implications for comparative genomic analyses of evolutionary innovations within the eudicots, which comprise ca. 75% of extant flowering plants⁴⁸. For example, the SCT genome will serve as an important reference outgroup for reconstructing the timing and nature of polyploidy event that gave rise to the hexaploid ancestor of all core eudicots (Pentapetalae)^{49,50}. Within the magnoliids we identified the timing of two independent rounds of WGD events that contributed to gene family expansions and innovations in pathogen, herbivore and mutualistic interactions.

Gene tree topologies for each of the six angiosperm TPS subfamilies revealed diversification of TPS genes and gene function in the ancestry of SCT. The C20s producing TPS-f genes were suggested to be eudicot-specific because both rice and sorghum lack genes in this subfamily⁴⁵. Our data clearly indicate that this subfamily was present in the last common ancestor of all but was lost from the grass family (Table 2). Massive diversification of the TPS-a and TPS-b subfamilies within the Lauraceae is consistent with a previous report that the main constituents of 58 essential oils produced in *Cinnamomum* leaves are C10s and C15s⁴⁷. These findings are in congruent with the fact that fruiting bodies of the SCT-specific parasitic fungus, *Antrodia cinnamomea*, can produce 78 kinds of terpenoids, including 31 structure-different triterpenoids (C30s)⁵¹, many of which are synthesized via the mevalonate pathway as are C10s and C15s followed by cyclizing squalenes (C₃₀H₅₀) into the skeletons of C30s⁵². It is reasonable to suggest that this fungus obtained intermediate compounds through decomposing trunk matters from SCT.

The 101 *CkTPSs* identified in the SCT genome are unevenly distributed across the 12 chromosomal scaffolds, and tandem arrays include gene clusters from the same subfamily (Supplementary Fig. 26). In the *Drosophila melanogaster* genome, “tandem

duplicate overactivity” has been observed with tandemly duplicated *Adh* genes showing 2.6-fold greater expression than single copy *Adh* genes⁵³.

In summary, the availability of SCT genome establishes a valuable genomic foundation that will help unravel the genetic diversity and evolution of other magnoliids, and a better understanding of flowering plant genome evolution and diversification. At the same time, the reference-quality SCT genome sequence will enable efforts to conserve genome-wide genetic diversity in this culturally and economically important tree species.

Methods

Plant Materials

All plant materials used in this study were collected from a 12-year-old SCT growing in Ershui Township, Changhua County, Taiwan (23°49'25.9"N, 120°36'41.2"E) during April to July of 2014–2016. The tree was grown up from a seedling obtained from Forestry Management Section, Department of Agriculture, Taoyuan City. The specimen (voucher number: Chaw 1501) was deposited in the Herbarium of Biodiversity Research Center, Academia Sinica, Taipei, Taiwan (HAST).

Genomic DNA extraction and sequencing

We used a modified high-salt method⁵⁴ to eliminate the high content of polysaccharides in SCT leaves, followed by total DNA extraction with a modified CTAB method⁵⁵. Three approaches were employed in DNA sequencing. First, paired-end and mate-pair libraries were constructed using the Illumina TruSeq DNA HT Sample Prep Kit and Illumina Nextera Mate Pair Sample Prep Kit following the kit's instructions, respectively. All obtained libraries were sequenced on an Illumina NextSeq 500 platform to generate ca. 278.8 Gb of raw data. Second, SMRT libraries were constructed using the PacBio 20-Kb protocol (<https://www.pacb.com/>). After loading on SMRT cells (SMRT™ Cell 8 Pac), these libraries were sequenced on a PacBio RS-II instrument using P6 polymerase and C4 sequencing reagent (Pacific Biosciences, Menlo Park, California). Third, a Chicago library was prepared by Dovetail Genomics (Santa Cruz, California) and sequenced on an Illumina HiSeq 2500 to generate 150 bp read pairs. Supplementary Table 1 summarizes the coverage and information for the sequencing data.

RNA extraction and sequencing

Opening flowers, flower buds (two stages), immature leaves, young leaves, mature leaves, young stems, and fruits were collected from the same individual (Supplementary Fig. 1c) and their total RNAs were extracted⁵⁶. The extracted RNA was purified using poly-T oligo-attached magnetic beads. All transcriptome libraries were constructed using Illumina TruSeq library Stranded mRNA Prep Kit and sequenced on an Illumina HiSeq 2000 platform. A summary of transcriptome data is shown in Supplementary Table 2.

Chromosome number assessment

Root tips from cutting seedlings were used to examine the chromosome number based on Suen *et al.*'s method⁵⁷. The stained samples were observed under a Nikon Eclipse 90i microscope (Supplementary Fig. 1a).

Genome size estimation

Fresh leaves of SCT were cut into tiny pieces and mixed well with 1 mL isolation buffer (200 mM Tris, 4 mM MgCl₂-6H₂O, and 0.5% Triton X-100)⁵⁸. The mixture was filtered through a 42 µm nylon mesh, followed by incubation of the filtered suspensions with a DNA fluorochrome (50 µg/ml propidium iodide and 50 µg/ml RNase). The genome size was estimated using a MoFlo XDP flow cytometry (Beckman Coulter Life Science, Indianapolis, Indiana) with chicken erythrocyte and rice nuclei (BioSure, Grass Valley, California) as the internal standards (Supplementary Fig. 1b). Estimate of genome size from Illumina paired end sequences was inferred using Genomescope⁵⁹ (based on k-mer 31).

De novo assembly of SCT

Illumina paired end and mate pair reads were trimmed with Trimmomatic⁶⁰ (ver. 0.32; options LEADING:30 TRAILING:30 SLIDINGWINDOW:4:30 MINLEN:50) and subsequently assembled using Platanus⁶¹. Pacbio reads were assembled using the FALCON⁶² assembler and the consensus sequences were improved using Quiver⁶³. The Pacbio assembly was scaffolded using HiRISE scaffolder and consensus sequences were further improved using Pilon with one iteration⁶⁴. The genome completeness was assessed using plant dataset of BUSCO²² (ver. 3.0.2). To identify putative telomeric repeats, the assembly was searched for high copy number repeats less than 10 base pairs using tandem repeat finder⁶⁵ (ver. 4.09; options: 2 7 7 80 10 500). The heptamer TTTAGGG was identified (Supplementary Table 6).

449

450 **Gene predictions and functional annotation**

451 Transcriptome paired end reads were aligned to the genome using STAR⁶⁶.
452 Transcripts were identified using two approaches: i) assembled *de novo* using
453 Trinity⁶⁷, ii) reconstructed using Stringtie⁶⁸ or CLASS2⁶⁹. Transcripts generated from
454 Trinity were remapped to the reference using GMAP⁷⁰. The three sets of transcripts
455 were merged and filtered using MIKADO (<https://github.com/lucventurini/mikado>).
456 Proteomes from representative reference species (Uniprot plants; Proteomes of
457 *Amborella trichopoda* and *Arabidopsis thaliana*) were downloaded from Phytozome
458 (ver. 12.1; <https://phytozome.jgi.doe.gov/>). The gene predictor Augustus⁷¹ (ver. 3.2.1)
459 and SNAP⁷² were trained either on the gene models data using BRAKER1⁷³ or
460 MAKER2²⁰. The assembled transcripts, reference proteomes, BRAKER1 and the
461 BUSCO predictions were combined as evidence hints for input of the MAKER2²⁰
462 annotation pipeline. MAKER2²⁰ invoked the two trained gene predictors to generate a
463 final set of gene annotation. Amino acid sequences of the proteome were functionally
464 annotated using Blast2GO⁷⁴ and egg-nog-mapper²¹. Nuclear plastid DNAs (NUPT) of
465 SCT was searched against its plastid genome (plastome; KR014245⁷⁵) using blastn
466 (parameters were followed from ref⁷⁶).

467

468 **Analysis of genome heterozygosity**

469 Paired end reads of SCT was aligned to reference using bwa mem⁷⁷ (ver.
470 0.7.17-r1188). PCR duplicates were removed using samtools⁷⁸ (ver. 1.8).
471 Heterozygous bi-allelic SNPs were called using samtools⁷⁸ and consensus sequences
472 were generated using bcftools⁷⁹ (ver. 1.7). Depth of coverage and minor allele
473 frequency plots were conducted using R ver. 3.4.2. Consensus sequence was fed to
474 the PSMC program²⁴ to infer past effective population size. All of the parameters used
475 for the PSMC program were at default with the exception of -u 7.5e-09 taken from *A.*
476 *thaliana*⁸⁰ and -g 20 taken from *Neolitsea sericea* (Lauraceae)⁸¹.

477

478 **Identification of repetitive elements**

479 Repetitive elements were firstly identified by modeling the repeats using
480 RepeatModeler⁸² and then searched and quantified repeats using RepeatMasker⁸³.
481 Repeat types modeled as “Unknown” by RepeatModeler were further annotated using
482 TEclass⁸⁴. Tandem Repeats were identified using Tandem Repeats Finder⁶⁵. The
483 proportions of different types of repeats were quantified by dissecting the 12 largest
484 scaffolds into 100,000 bp chunks and calculating the total lengths and percentages of
485 the repetitive elements within the chunks. LTR-RT domains were extracted following
486 Guan *et al.*'s method⁸⁵. Briefly, a two-step procedure was applied on the genomes.

The first was to find candidate LTR-RTs similar to known reverse transcriptase domains and second was to identify other LTR-RTs using the candidates identified in the first step. The identified LTR-RT domains were integrated with those downloaded from the Ty1/Copia and Ty3/Gypsy trees of Guan *et al.*⁸⁵. Trees were built by aligning the sequences using MAFFT⁸⁷ (ver. 7.310; --genafpair --ep 0) and applied FastTree⁸⁸ with JTT model on the aligned sequences, and were colored using APE package⁸⁹.

Gene family / Orthogroup inference and analysis of protein domains

The amino acid and nucleotide sequences of 12 representative plant species were downloaded from various sources: *Aquilegia coerulea*, *Arabidopsis thaliana*, *Daucus carota*, *Mimulus guttatus*, *Musa acuminata*, *Oryza sativa japonica*, *Populus trichocarpa*, *Vitis vinifera* and *Zea mays* from Phytozome (ver. 12.1; <https://phytozome.jgi.doe.gov/>), *Picea abies* from the Plant Genome Integrative Explorer Resource⁹⁰ (<http://plantgenie.org/>), *Ginkgo biloba* from GigaDB⁹¹, and *Amborella trichopoda* from Ensembl plants⁹² (Release 39; <https://plants.ensembl.org/index.html>). Gene families or orthologous groups of these species and SCT were determined by OrthoFinder²³ (ver. 2.2.0). Protein family domains (Pfam) of each species were calculated from Pfam website (ver. 31.0; <https://pfam.xfam.org/>). Pfam numbers of every species were transformed into z-scores. Significant expansion or reduction of Pfams in SCT were based on its z-score greater than 1.96 or less than -1.96, respectively. The significant Pfams were sorted by Pfam numbers (Supplementary Fig. 16). Gene family expansion and loss were inferred using CAFE⁴¹ (ver. 4.1 with input tree as the species tree inferred from the single copy orthologues).

Phylogenetic analysis

MAFFT⁸⁷ (ver. 7.271; option --maxiterate 1000) was used to align 13 sets of amino acid sequences of 211 single-copy OGs. Each OG alignment was used to compute a maximum likelihood phylogeny using RAxML³⁰ (ver. 8.2.11; options: -m PROTGAMMAILGF -f a) with 500 bootstrap replicates. The best phylogeny and bootstrap replicates for each gene were used to infer a consensus species tree using ASTRAL-III³¹. A maximum likelihood phylogeny was constructed with the concatenated amino acid alignments of the single copy OGs (ver. 8.2.11; options: -m PROTGAMMAILGF -f a) also with 500 bootstrap replicates.

Estimation of divergence time

Divergence time of each tree node was inferred using MCMCtree of PAML³² package (ver. 4.9g; options: correlated molecular clock, JC69 model and rest being default). The final species tree and the concatenated translated nucleotide alignments of 211 single-copy-orthologs were used as input of MCMCtree. The phylogeny was calibrated using various fossil records or molecular divergence estimate by placing soft bounds at split node of: i) *A. thaliana*-*V. vinifera* (115–105 Ma)⁹³, ii) *M. acuminata*-*Z. mays* (115–90 Ma)⁹³, iii) Ranunculales (128.63–119.6 Ma)³⁴, iv) Angiospermae (247.2–125 Ma)³⁴, v) Acrogymnospermae (365.629–308.14 Ma)³⁴, and v) a hard bound of 420 Ma of outgroup *P. patens*⁹⁴.

Analysis of genome synteny and whole genome duplication

Dot plots between SCT and *A. trichopoda* assemblies were produced using SynMap from Comparative Genomics Platform (Coge⁹⁵) to visualize the paleoploidy level of SCT. Synteny blocks within SCT and between *A. trichopoda* and *A. coerulea* were identified using DAGchainer⁹⁶ (same parameters as Coge⁹⁵: -E 0.05 -D 20 -g 10 -A 5). Ks between syntenic group pairs were calculated using the DECIPHER⁹⁷ package in R. Depth of the inferred syntenic blocks were calculated using Bedtools⁹⁸. Both the Ks distribution and syntenic block depth were used to determine the paleopolyploidy level⁹⁹ of SCT. Using the quadruplicate or triplicate orthologues in the syntenic blocks as backbones, as well as *A. trichopoda* regions showing up to four syntenic regions, we identified the start and end coordinates of linkage clusters (Supplementary Note).

Resistance (R) genes

R genes were identified based on the ref¹⁰⁰. Briefly, the predicted genes of the 13 sampled species were searched for the Pfam NBS (NB-ARC) protein family (PF00931) using HMMER ver. 3.1b2¹⁰¹ with an e-value cutoff of 1e-5. Extracted sequences were then checked for protein domains using InterproScan¹⁰² (ver. 5.19-58.0) to remove false positive NB-ARC domain hits. The NBS domains of the genes that passed both HMMER and InterproScan were extracted according to the InterproScan annotation and aligned using MAFFT⁸⁷ (ver. 7.310; --genafpair --ep 0); the alignment was then input into FastTree⁸⁸ with the JTT model and visualized using EvolView¹⁰³.

Terpene synthase genes

In addition to the 13 species' proteome dataset used in this study, transcriptome data from one Chloranthaceae species, *Sarcandra glabra* and two magnollids representatives, *Persea americana* (avocado) and *Saruma henryi* (saruma), were

downloaded from oneKP transcriptome database¹⁰⁴. Previously annotated TPS genes of four species: *Arabidopsis thaliana*¹⁰⁵, *Oryza sativa*⁴⁵, *Populus trichocarpa*¹⁰⁶, and *Vitis vinifera*¹⁰⁷ were retrieved. For species without *a priori* TPS annotations, two Pfam domains: PF03936 and PF01397, were used to identify against the proteomes using HMMER¹⁰⁸ (ver. 3.0; cut-off at e-values $< 10^{-5}$). Sequence lengths shorter than 200 amino acids were excluded from further analysis. 702 putative or annotated protein sequences of *TPS* were aligned using MAFFT⁸⁷ (ver. 7.310 with default parameters) and manually adjusted using MEGA¹⁰⁹ (ver. 7.0). The TPS gene tree was constructed using FastTree¹¹⁰ (ver. 2.1.0) with 1,000 bootstrap replicates. Subfamily TPS-c was designated as the outgroup. Branching nodes with bootstrap values $< 80\%$ were treated as collapsed.

Figure and Table legends

Figure 1| Stout camphor tree genome heterozygosity. **a**, Number of heterozygous bi-allelic SNPs per 100 kb non-overlapping windows is plotted along the largest 12 scaffolds. Indels were excluded. **b**, The history of effective population size was inferred using the PSMC method. 100 bootstraps were performed and the margins are shown in light red.

Figure 2| Genomic landscape of stout camphor tree chromosome 1. For every non-overlapping 100 kb window distribution is shown from top to bottom: gene density (percent of nucleotide with predicted model), transcriptome (percent of nucleotides with evidence of transcriptome mapping), three different classes repetitive sequences (percent of nucleotides with TE annotation) and heterozygosity (number of bi-allelic SNPs). The red T letter denote presence of telomeric repeat cluster at scaffold end.

Figure 3| A species tree on the basis of 211 single copy orthologues from 13 plant species. Gene family expansion and contraction are denoted in numbers next to plus and minus signs, respectively. Unless stated, bootstrap support of 100 is denoted as blue circles.

Figure 4| Evolutionary analysis of the stout camphor tree genome. **a**, Schematic representation of intragenomic relationship amongst the 637 syntenic blocks in the stout camphor tree genome. Syntenic blocks assigned unambiguously into 5 linkage clusters representing ancient karyotypes are color coded. **b**, Schematic representation

of the first linkage group within the stout camphor tree genome and their corresponding relationship in *A. trichopoda*.

Figure 5| Density plots of synonymous substitutions (Ks) of stout camphor tree genome and other plant species. a, Pairwise orthologue duplicates identified in synteny blocks within SCT, within *A. coerulea* and between SCT and *A. coerulea*. **b,** Ks of intragenomic pairwise duplicates of the Lauraceae and the Magnoliales in the 1KP project¹⁰⁴. Dashed lines denote the two Ks peaks observed in SCT.

Figure 6| Phylogenetic tree of putative or characterized TPS genes from the 13 sequenced land plant genomes and two magnoliids with available transcriptomic data.

Table 1| Statistics of stout camphor tree genome assemblies using different sequencing technologies and final gene predictions.

Table 2| Comparison of the known/predicted seven TPS subfamilies among 14 known genomes and three available transcriptomes of major seed plant lineages.

Authors contribution

Conceived the study: S.M.C

Genome assembly and annotation: I.J.T and H.M.K

Repeat Analysis: L.Y.C and Y.W.W

Plastid DNA analysis: E.S.

Conducted the experiments: C.S.W, L.N.W, H.T.Y., C.Y.H and S.M.C.

Comparative genomics analysis: I.J.T, Y.L, H.M.K, C.Y.I.L and J.L.M

Analysis of R genes: Y.W.W, M.H.H, K.P.W, S.M.C

Analysis of terpene gene family: H.Y.W, S.M.C, C.Y.H and Y.W.W

Wrote the manuscript: I.J.T, J.LM and S.M.C

Data availability

All of raw sequence reads used in this study have been deposited in NCBI under the BioProject accession number PRJNA477266. The assembly and annotation of SCT is available under the accession number SAMN09509728.

Acknowledgement

Chi-yuan Tsai for plant materials; Chih-Ming Hung for PSMC analysis. S.M.C was funded by Investigators' Award and Central Academic Committee, Academia Sinica. I.J.T was funded by Career Development Award, Academia Sinica. H.M.K, C.S.W and C.Y.H were funded by postdoctoral fellowship, Academia Sinica.

References

- 1 Jayaprakasha, G. K., Rao, L. J. & Sakariah, K. K. Chemical composition of volatile oil from *Cinnamomum zeylanicum* buds. *Z Naturforsch C* **57**, 990–993 (2002).
- 2 Joshi, R., Satyal, P. & Setzer, W. Himalayan Aromatic medicinal plants: a review of their ethnopharmacology, volatile phytochemistry, and biological activities. *Medicines* **3**, doi:10.3390/medicines3010006 (2016).
- 3 Kaul, P. N., Bhattacharya, A. K., Rajeswara Rao, B. R., Syamasundar, K. V. & Ramesh, S. Volatile constituents of essential oils isolated from different parts of cinnamon (*Cinnamomum zeylanicum* Blume). *J Sci Food and Agric* **83**, 53–55, doi:10.1002/jsfa.1277 (2003).
- 4 Shahlari, M., Hamidpour, M., Hamidpour, S. & Hamidpour, R. Camphor (*Cinnamomum camphora*), a traditional remedy with the history of treating several diseases. *Int J Case Rep Images* **4**, doi:10.5348/ijcri-2013-02-267-RA-1 (2013).
- 5 Christenhusz, M. J. M. & Byng, J. W. The number of known plants species in the world and its annual increase. *Phytotaxa* **261**, doi:10.11646/phytotaxa.261.3.1 (2016).
- 6 Palmer, J. D., Soltis, D. E. & Chase, M. W. The plant tree of life: an overview and some points of view. *Am J Bot* **91**, 1437–1445, doi:10.3732/ajb.91.10.1437 (2004).
- 7 Cui, L. *et al.* Widespread genome duplications throughout the history of flowering plants. *Genome Res* **16**, 738–749, doi:10.1101/gr.4825606 (2006).
- 8 Liu, Y. C., Lu, F. Y. & Ou, C. H. Trees of Taiwan. *Monographic Publication* **7**, 105–131 (1988).
- 9 Fujita, Y. Classification and phylogeny of the genus *Cinnamomum* viewed from the constituents of essential oils. *Shokubutsugaku Zasshi* **80**, 261–271, doi:10.15281/jplantres1887.80.261 (1967).
- 10 Chang, T. T. & Chou, W. N. *Antrodia cinnamomea* sp. nov. on *Cinnamomum kanehirai* in Taiwan. *Mycol Res* **99**, 756–758, doi:https://doi.org/10.1016/S0953-7562(09)80541-8 (1995).
- 11 Wu, S. H., Ryvarden, L. & Chang, T. T. *Antrodia camphorata*

676 ("niu-chang-chih"), new combination of a medicinal fungus in Taiwan. *Bot*
677 *Stud* **38**, 273–275 (1997).

678 12 Hseu, Y. C., Chen, S. C., Yech, Y. J., Wang, L. & Yang, H. L. Antioxidant
679 activity of *Antrodia camphorata* on free radical-induced endothelial cell
680 damage. *J Ethnopharmacol* **118**, 237–245, doi:10.1016/j.jep.2008.04.004
681 (2008).

682 13 Liao, P. C. *et al.* Historical spatial range expansion and a very recent
683 bottleneck of *Cinnamomum kanehirae* Hay. (Lauraceae) in Taiwan inferred
684 from nuclear genes. *BMC Evol Biol* **10**, 124, doi:10.1186/1471-2148-10-124
685 (2010).

686 14 Hung, K. H., Lin, C. H., Shih, H. C., Chiang, Y. C. & Ju, L. P. Development,
687 characterization and cross-species amplification of new microsatellite primers
688 from an endemic species *Cinnamomum kanehirae* (Lauraceae) in Taiwan.
689 *Conserv Genet Resour* **6**, 911–913, doi:10.1007/s12686-014-0239-z (2014).

690 15 Zerbe, P. & Bohlmann, J. Plant diterpene synthases: exploring modularity and
691 metabolic diversity for bioengineering. *Trends Biotechnol* **33**, 419–428,
692 doi:10.1016/j.tibtech.2015.04.006 (2015).

693 16 Loreto, F., Dicke, M., Schnitzler, J. P. & Turlings, T. C. Plant volatiles and the
694 environment. *Plant Cell Environ* **37**, 1905–1908, doi:10.1111/pce.12369
695 (2014).

696 17 Tholl, D. Biosynthesis and biological functions of terpenoids in plants. *Adv*
697 *Biochem Eng Biotechnol* **148**, 63–106, doi:10.1007/10_2014_295 (2015).

698 18 Gonzalez-Coloma, A., Reina, M., Diaz, C. E. & Fraga, B. M. in
699 *Comprehensive Natural Products II* (eds Liu, H. W. & Mander, L.) 237–268
700 (Elsevier, 2010).

701 19 Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro
702 method for long-range linkage. *Genome Res* **26**, 342–350,
703 doi:10.1101/gr.193474.115 (2016).

704 20 Holt, C. & Yandell, M. MAKER2: an annotation pipeline and
705 genome-database management tool for second-generation genome projects.
706 *BMC bioinformatics* **12**, 491, doi:10.1186/1471-2105-12-491 (2011).

707 21 Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through
708 orthology assignment by eggNOG-mapper. *Mol Biol Evol* **34**, 2115–2122
709 doi:10.1093/molbev/msx148 (2017).

710 22 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov,
711 E. M. BUSCO: assessing genome assembly and annotation completeness with
712 single-copy orthologs. *Bioinformatics* **31**, 3210–3212,
713 doi:10.1093/bioinformatics/btv351 (2015).

714 23 Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole
715 genome comparisons dramatically improves orthogroup inference accuracy.
716 *Genome Biol* **16**, 157, doi:10.1186/s13059-015-0721-2 (2015).

717 24 Li, H. & Durbin, R. Inference of human population history from individual
718 whole-genome sequences. *Nature* **475**, 493–496, doi:10.1038/nature10231
719 (2011).

720 25 Sibuet, J.-C. & Hsu, S.-K. How was Taiwan created? *Tectonophysics* **379**,
721 159–181, doi:10.1016/j.tecto.2003.10.022 (2004).

722 26 Dong, P. *et al.* 3D Chromatin architecture of large plant genomes determined
723 by local A/B Compartments. *Mol Plant* **10**, 1497–1509,
724 doi:10.1016/j.molp.2017.11.005 (2017).

725 27 Watson, J. M. & Riha, K. Comparative biology of telomeres: where plants
726 stand. *FEBS Lett* **584**, 3752–3759, doi:10.1016/j.febslet.2010.06.017 (2010).

727 28 Zeng, L. *et al.* Resolution of deep angiosperm phylogeny using conserved
728 nuclear genes and estimates of early divergence times. *Nat Commun* **5**, 4956,
729 doi:10.1038/ncomms5956 (2014).

730 29 An update of the Angiosperm Phylogeny Group classification for the orders
731 and families of flowering plants: APG IV. *Bot J Linean Soc* **181**, 1–20,
732 doi:doi:10.1111/boj.12385 (2016).

733 30 Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic
734 analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford,*
735 *England)* **22**, 2688–2690, doi:10.1093/bioinformatics/btl446 (2006).

736 31 Mirarab, S. & Warnow, T. ASTRAL-II: coalescent-based species tree
737 estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*
738 **31**, 44–52, doi:10.1093/bioinformatics/btv234 (2015).

739 32 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol*
740 *Evol* **24**, 1586–1591, doi:10.1093/molbev/msm088 (2007).

741 33 Massoni, J., Couvreur, T. L. & Sauquet, H. Five major shifts of diversification
742 through the long evolutionary history of Magnoliidae (angiosperms). *BMC*
743 *Evol Biol* **15**, 49, doi:10.1186/s12862-015-0320-6 (2015).

744 34 Morris, J. L. *et al.* The timescale of early land plant evolution. *Proc Natl Acad*
745 *Sci USA* **115**, E2274–E2283, doi:10.1073/pnas.1719588115 (2018).

746 35 Zhong, B. & Betancur-R, R. Expanded taxonomic sampling coupled with gene
747 genealogy interrogation provides unambiguous resolution for the evolutionary
748 root of angiosperms. *Genome Biol Evol* **9**, 3154–3161,
749 doi:10.1093/gbe/evx233 (2017).

750 36 Matasci, N. *et al.* Data access for the 1,000 Plants (1KP) project. *GigaScience*
751 **3**, doi:10.1186/2047-217x-3-17 (2014).

752 37 Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res* **42**,
753 D222–230, doi:10.1093/nar/gkt1223 (2014).

754 38 Lang, T. *et al.* Protein domain analysis of genomic sequence data reveals
755 regulation of LRR related domains in plant transpiration in Ficus. *PLoS One* **9**,
756 e108719, doi:10.1371/journal.pone.0108719 (2014).

757 39 Jourda, C. *et al.* Expansion of banana (*Musa acuminata*) gene families
758 involved in ethylene biosynthesis and signalling after lineage-specific
759 whole-genome duplications. *New Phytol* **202**, 986–1000,
760 doi:10.1111/nph.12710 (2014).

761 40 Seyfferth, C. *et al.* Ethylene-related gene expression networks in wood
762 formation. *Front Plant Sci* **9**, 272, doi:10.3389/fpls.2018.00272 (2018).

763 41 De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a
764 computational tool for the study of gene family evolution. *Bioinformatics* **22**,
765 1269–1271, doi:10.1093/bioinformatics/btl097 (2006).

766 42 Dodds, P. N. *et al.* Direct protein interaction underlies gene-for-gene
767 specificity and coevolution of the flax resistance genes and flax rust avirulence
768 genes. *Proc Natl Acad Sci USA* **103**, 8888–8893,
769 doi:10.1073/pnas.0602577103 (2006).

770 43 Trapp, S. C. & Croteau, R. B. Genomic organization of plant terpene synthases
771 and molecular evolutionary implications. *Genetics* **158**, 811–832 (2001).

772 44 Gershenzon, J. & Dudareva, N. The function of terpene natural products in the
773 natural world. *Nat Chem Biol* **3**, 408, doi:10.1038/nchembio.2007.5 (2007).

774 45 Chen, F., Tholl, D., Bohlmann, J. & Pichersky, E. The family of terpene
775 synthases in plants: a mid-size family of genes for specialized metabolism that
776 is highly diversified throughout the kingdom. *Plant J* **66**, 212–229,
777 doi:10.1111/j.1365-313X.2011.04520.x (2011).

778 46 Martin, D. M., Fäldt, J. & Bohlmann, J. Functional characterization of nine
779 Norway Spruce *TPS* genes and evolution of gymnosperm terpene synthases of
780 the *TPS-d* Subfamily. *Plant Physiol* **135**, 1908–1927,
781 doi:10.1104/pp.104.042028 (2004).

782 47 Cheng, S.-S. *et al.* Chemical polymorphism and composition of leaf essential
783 oils of *Cinnamomum kanehirae* using gas chromatography/mass spectrometry,
784 cluster analysis, and principal component analysis. *J Wood Chem Tech* **35**,
785 207–219, doi:10.1080/02773813.2014.924967 (2015).

786 48 Liping, Z. *et al.* Resolution of deep eudicot phylogeny and their temporal
787 diversification using nuclear genes from transcriptomic and genomic datasets.
788 *New Phytol* **214**, 1338–1354, doi:doi:10.1111/nph.14503 (2017).

789 49 Jiao, Y. *et al.* A genome triplication associated with early diversification of the

core eudicots. *Genome Biol* **13**, R3–R3, doi:10.1186/gb-2012-13-1-r3 (2012).

50 Chanderbali, A. S., Berger, B. A., Howarth, D. G., Soltis, D. E. & Soltis, P. S.
Evolution of floral diversity: genomics, genes and gamma. *Philos Trans R Soc
Lond B Biol Sci* **372**, doi:10.1098/rstb.2015.0509 (2017).

51 Geethangili, M. & Tzeng, Y. M. Review of pharmacological effects of
Antrodia camphorata and its bioactive compounds. *Evidence-Based
Complement Alternat Med* **2011**, 17, doi:10.1093/ecam/nep108 (2011).

52 Lu, M. Y. *et al.* Genomic and transcriptomic analyses of the medicinal fungus
Antrodia cinnamomea for its metabolite biosynthesis and sexual development.
Proc Natl Acad Sci USA **111**, E4743–4752, doi:10.1073/pnas.1417570111
(2014).

53 Loehlin, D. W. & Carroll, S. B. Expression of tandem gene duplicates is often
greater than twofold. *Proc Natl Acad Sci USA* **113**, 5988–5992,
doi:10.1073/pnas.1605886113 (2016).

54 Sandbrink, J. M., Vellekoop, P., Vanham, R. & Vanbrederode, J. A method for
evolutionary studies on RFLP of chloroplast DNA, applicable to a range of
plant species. *Biochem Syst Ecol* **17**, 45–49, doi:Doi
10.1016/0305-1978(89)90041-0 (1989).

55 Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small
quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**, 11–15,
doi:citeulike-article-id:678648 (1987).

56 Kolosova, N., Gorenstein, N., Kish, C. M. & Dudareva, N. Regulation of
circadian methyl benzoate emission in diurnally and nocturnally emitting
plants. *Plant Cell* **13**, 2333–2347 (2001).

57 Suen, D. F. *et al.* Assignment of DNA markers to *Nicotiana sylvestris*
chromosomes using monosomic alien addition lines. *Theor Appl Genet* **94**,
331–337, doi:DOI 10.1007/s001220050420 (1997).

58 Dolezel, J., Greilhuber, J. & Suda, J. Estimation of nuclear DNA content in
plants using flow cytometry. *Nat Protoc* **2**, 2233–2244,
doi:10.1038/nprot.2007.310 (2007).

59 Vurture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from
short reads. *Bioinformatics* **33**, 2202–2204, doi:10.1093/bioinformatics/btx153
(2017).

60 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for
Illumina sequence data. *Bioinformatics (Oxford, England)*, 1–7,
doi:10.1093/bioinformatics/btu170 (2014).

61 Kajitani, R. *et al.* Efficient de novo assembly of highly heterozygous genomes
from whole-genome shotgun short reads. *Genome res* **24**, 1384–1395,

doi:10.1101/gr.170720.113 (2014).

62 Chin, C. S. *et al.* Phased diploid genome assembly with single-molecule
real-time sequencing. *Nat Methods* **13**, 1050–1054, doi:10.1038/nmeth.4035
(2016).

63 Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from
long-read SMRT sequencing data. *Nat Methods* **10**, 563–569,
doi:10.1038/nmeth.2474 (2013).

64 Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial
variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963,
doi:10.1371/journal.pone.0112963 (2014).

65 Benson, G. Tandem repeats finder: a program to analyze DNA sequences.
Nucleic Acids Res **27**, 573–580 (1999).

66 Dobin, A., Davis, C. & Schlesinger, F. STAR: ultrafast universal RNA-seq
aligner. *Bioinformatics*, 1–7 (2013).

67 Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq
using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**,
1494–1512, doi:10.1038/nprot.2013.084 (2013).

68 Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome
from RNA-seq reads. *Nat Biotechnol* **33**, 290–295, doi:10.1038/nbt.3122
(2015).

69 Song, L., Sabunciyan, S. & Florea, L. CLASS2: accurate and efficient splice
variant annotation from RNA-seq reads. *Nucleic Acids Res* **44**, e98,
doi:10.1093/nar/gkw158 (2016).

70 Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment
program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875,
doi:10.1093/bioinformatics/bti310 (2005).

71 Stanke, M., Tzvetkova, A. & Morgenstern, B. AUGUSTUS at EGASP: using
EST, protein and genomic alignments for improved gene prediction in the
human genome. *Genome Biol* **7 Suppl 1**, S11.11–18,
doi:10.1186/gb-2006-7-s1-s11 (2006).

72 Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59,
doi:10.1186/1471-2105-5-59 (2004).

73 Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M.
BRAKER1: unsupervised RNA-Seq-Based genome annotation with
GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769,
doi:10.1093/bioinformatics/btv661 (2016).

74 Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and
analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676,

doi:10.1093/bioinformatics/bti610 (2005).

75 Wu, C. C., Ho, C. K. & Chang, S. H. The complete chloroplast genome of
Cinnamomum kanehirae Hayata (Lauraceae). *Mitochondr DNA* **27**, 2681–
2682, doi:10.3109/19401736.2015.1043541 (2016).

76 Smith, D. R., Crosby, K. & Lee, R. W. Correlation between nuclear plastid
DNA abundance and plastid number supports the limited transfer window
hypothesis. *Genome Biol Evol* **3**, 365–371, doi:10.1093/gbe/evr001 (2011).

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

78 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools.
Bioinformatics (Oxford, England) **25**, 2078–2079,
doi:10.1093/bioinformatics/btp352 (2009).

79 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**,
2156–2158, doi:10.1093/bioinformatics/btr330 (2011).

80 Buschiazzi, E., Ritland, C., Bohlmann, J. & Ritland, K. Slow but not low:
genomic comparisons reveal slower evolutionary rate and higher dN/dS in
conifers compared to angiosperms. *BMC Evol Biol* **12**, 8,
doi:10.1186/1471-2148-12-8 (2012).

81 Cao, Y. N. *et al.* Inferring spatial patterns and drivers of population divergence
of *Neolitsea sericea* (Lauraceae), based on molecular phylogeography and
landscape genomics. *Mol Phylogenet Evol* **126**, 162–172,
doi:10.1016/j.ympev.2018.04.010 (2018).

82 Smit, A. & Hubley, R. *RepeatModeler Open-1.0*, <http://www.repeatmasker.org>
(2008–2015).

83 Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-4.0*,
<http://www.repeatmasker.org> (2013–2015).

84 Abrusan, G., Grundmann, N., DeMester, L. & Makalowski, W. TEclass--a tool
for automated classification of unknown eukaryotic transposable elements.
Bioinformatics **25**, 1329–1330, doi:10.1093/bioinformatics/btp084 (2009).

85 Guan, R. *et al.* Draft genome of the living fossil *Ginkgo biloba*. *Gigascience* **5**,
49, doi:10.1186/s13742-016-0154-1 (2016).

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

87 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately
maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490,
doi:10.1371/journal.pone.0009490 (2010).

904 88 Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and
905 evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

906 89 Sundell, D. *et al.* The plant genome integrative explorer resource:
907 PlantGenIE.org. *New Phytol* **208**, 1149–1156, doi:10.1111/nph.13557 (2015).

908 90 Sneddon, T. P., Li, P. & Edmunds, S. C. GigaDB: announcing the GigaScience
909 database. *Gigascience* **1**, 11, doi:10.1186/2047-217X-1-11 (2012).

910 91 Bolser, D., Staines, D. M., Pritchard, E. & Kersey, P. Ensembl Plants:
911 integrating tools for visualizing, mining, and analyzing plant genomics data.
912 *Methods Mol Biol* **1374**, 115–140, doi:10.1007/978-1-4939-3167-5_6 (2016).

913 92 Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for
914 timelines, timetrees, and divergence times. *Mol Biol Evol* **34**, 1812–1819,
915 doi:10.1093/molbev/msx116 (2017).

916 93 Pryer, K. M. *et al.* Horsetails and ferns are a monophyletic group and the
917 closest living relatives to seed plants. *Nature* **409**, 618–622,
918 doi:10.1038/35054555 (2001).

919 94 Lyons, E. *et al.* Finding and comparing syntenic regions among *Arabidopsis*
920 and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol*
921 **148**, 1772–1781, doi:10.1104/pp.108.124867 (2008).

922 95 Haas, B. J., Delcher, A. L., Wortman, J. R. & Salzberg, S. L. DAGchainer: a
923 tool for mining segmental genome duplications and synteny. *Bioinformatics*
924 (*Oxford, England*) **20**, 3643–3646, doi:10.1093/bioinformatics/bth397 (2004).

925 96 Wright, E. Using DECIPHER v2.0 to analyze big biological sequence data in
926 R. *The R Journal* **8**, 352–359 (2016).

927 97 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for
928 comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841–842,
929 doi:10.1093/bioinformatics/btq033 (2010).

930 98 Ming, R. *et al.* The pineapple genome and the evolution of CAM
931 photosynthesis. *Nat Genet* **47**, 1435–1442, doi:10.1038/ng.3435 (2015).

932 99 Lozano, R., Hamblin, M. T., Prochnik, S. & Jannink, J. L. Identification and
933 distribution of the NBS-LRR gene family in the Cassava genome. *BMC*
934 *Genomics* **16**, 360, doi:10.1186/s12864-015-1554-9 (2015).

935 100 Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput Biol* **7**,
936 e1002195, doi:10.1371/journal.pcbi.1002195 (2011).

937 101 Finn, R. D. *et al.* InterPro in 2017-beyond protein family and domain
938 annotations. *Nucleic Acids Res* **45**, D190–D199, doi:10.1093/nar/gkw1107
939 (2017).

940 102 He, Z. *et al.* Evolvview v2: an online visualization and management tool for
941 customized and annotated phylogenetic trees. *Nucleic Acids Res* **44**, W236–

942 241, doi:10.1093/nar/gkw370 (2016).

943 103 Matasci, N. *et al.* Data access for the 1,000 Plants (1KP) project. *Gigascience*
944 **3**, 17, doi:10.1186/2047-217X-3-17 (2014).

945 104 Aubourg, S., Lecharny, A. & Bohlmann, J. Genomic analysis of the terpenoid
946 synthase (*AtTPS*) gene family of *Arabidopsis thaliana*. *Mol Genet Genomics*
947 **267**, 730–745, doi:10.1007/s00438-002-0709-y (2002).

948 105 Irmisch, S., Jiang, Y., Chen, F., Gershenzon, J. & Köllner, T. G. Terpene
949 synthases and their contribution to herbivore-induced volatile emission in
950 western balsam poplar (*Populus trichocarpa*). *BMC Plant Biol* **14**, 270,
951 doi:10.1186/s12870-014-0270-y (2014).

952 106 Martin, D. M. *et al.* Functional annotation, genome organization and
953 phylogeny of the grapevine (*Vitis vinifera*) terpene synthase gene family based
954 on genome assembly, FLcDNA cloning, and enzyme assays. *BMC Plant Biol*
955 **10**, 226, doi:10.1186/1471-2229-10-226 (2010).

956 107 Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile
957 HMMs. *Bioinformatics* **29**, 2487–2489, doi:10.1093/bioinformatics/btt403
958 (2013).

959 108 Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular evolutionary
960 genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* **33**, 1870–1874,
961 doi:10.1093/molbev/msw054 (2016).

962 109 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum
963 evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**,
964 1641–1650, doi:10.1093/molbev/msp077 (2009).

965

966

967

a **b**

a **b**

	Illumina (v1)	Pacbio (v2)	Pacbio + HiC (v3)
Assembly feature			
Size (Mb)	714.7	728.3	730.7
Num. scaffolds	48,650	5,210	2,153
Average (Kb)	15	140	339
Largest scaffold (Mb)	6.2	12.8	87.0
N50 (Mb)	0.6	0.9	50.4
L50 (n)	254	185	6
N90 (Mb)	0.003	0.058	37.1
L90 (n)	6,123	1,709	12
Genome annotation			
Num. genes (n)			27,899
Gene model (Mb)			235.6
Exons (Mb)			36.4
Intron length (Mb)			199.2
Num. genes in 12 largest scaffolds			26,692
BUSCO completeness (%)			88.5

Table 1. Statistics of the stout camphor tree genome assemblies using different sequencing technologies and final gene predictions.

Table 2. Numbers of TPS subfamilies in the 14 genomes and three transcriptomes of major seed plant lineages.

TPS subfamilies	Genome size (Mb)	Primary Metabolism		Secondary Metabolism					Total no.
		c	e	a	b	d ²	f	g	
Function		CPS, C20 ¹	KS, C20	C15	IspS, C10	C10,15,20	C20	C10	
Species									
Gymnosperms									
<i>Ginkgo biloba</i>	10,609	1	1	-	-	49	-	-	51
<i>Picea abies</i>	12,301	2	1	-	-	59	-	-	62
Angiosperms									
<i>Amborella trichopoda</i>	706	1	1	-	7	-	3	5	17
Chloranthaceae									
<i>Sarcandra glabra</i> ³	-	-	1	2	-	-	-	-	3
Magnoliids									
Lauraceae									
<i>Cinnamomum kanehirae</i>	731	2	5	25	58	-	7	4	101
<i>Persea americana</i> ³	-	-	-	11	12	-	1	9	33
Piperales									
<i>Saruma henryi</i> ³	-	1	-	1	2	-	-	1	5
Monocots									
<i>Musa acuminata</i>	473	2	2	21	13	-	3	3	44
<i>Oryza sativa</i>	375	3	10	19	-	-	-	1	33
<i>Zea mays</i>	2,068	8	6	30	2	-	-	5	51
Eudicots									
<i>Aquilegia coerulea</i>	307	15	13	12	34	-	-	8	82
Rosids									
<i>Arabidopsis thaliana</i>	120	1	1	23	5	-	1	1	32

<i>Populus trichocarpa</i>	473	2	2	16	14	-	1	3	38
<i>Vitis vinifera</i> ⁴	434	2	1	29	10	-	2	14	58 ⁴
Asterids									
<i>Daucus carota</i>	422	3	2	1	15	-	1	7	29
<i>Mimulus guttatus</i>	313	13	13	19	17	-	-	1	63

¹ The “C” and “Arabic number” within the parenthesis designate C10: monoterpene; C15: sesquiterpenes; C20: diterpenes.

CPS: copalyl diphosphate synthase; KS: kaurene synthase; IspS: isoprene synthase.

² TPS-d subfamily is gymnosperm-specific.

³ Transcriptome data of these three taxa were highly likely incomplete for covering all TPS transcripts, so that their total numbers of TPS were not reliable but for reference only.

⁴ These two TPS-f were previously characterized from grape floral cDNA without identical genomic *VvTPS* genes (Martin et al., 2010); *VvTPS* sequences that labeled as unknown (Martin et al., 2010) in the TPS gene tree were not counted.