nature plants

ARTICLES

https://doi.org/10.1038/s41477-018-0337-0

OPEN

Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution

Shu-Miaw Chaw[1,6*], Yu-Ching Liu[1], Yu-Wei Wu[2], Han-Yu Wang[1], Chan-Yi Ivy Lin[1], Chung-Shien Wu[1], Huei-Mien Ke[1], Lo-Yu Chang[1,3], Chih-Yao Hsu[1], Hui-Ting Yang[1], Edi Sudianto[1], Min-Hung Hsu[1,4], Kun-Pin Wu[4], Ling-Ni Wang[1], James H. Leebens-Mack[5] and Isheng J. Tsai[1,6*]

We present reference-quality genome assembly and annotation for the stout camphor tree (*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 than 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 the terpenoid synthase gene subfamilies within the Laurales spawned the diversity of *Cinnamomum* monoterpenes and sesquiterpenes.

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). For example, camphor, a bicyclic monoterpene ketone ($C_{10}H_{16}O$) that can be obtained from many members of this genus, has important industrial and pharmaceutical applications[1]. *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 America[2]. Among them, avocado (*Persea americana*), bay laurel (*Laurus nobilis*), camphor tree or camphor laurel (*Cinnamomum camphora*), cassia (*Cinnamomum cassia*) and cinnamon (including several *Cinnamomum* 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 (Magnoliidae), containing about 9,000 species, are characterized by three-merous flowers with diverse volatile secondary compounds, one-pored pollen and insect pollination[3]. 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. However, the phylogenetic position of magnoliids has been uncertain. They were considered to be (1) sister to the Chloranthaceae[4], (2) sister to the monocots[5], (3) sister to the clade containing monocots and eudicots[6], (4) sister to the clade composing Chloranthaceae and Ceratophyllaceae[7], or (5) sister to the clade including eudicots and Chloranthaceae–Ceratophyllaceae[8], based on plastid genes, plastomic inverted repeat regions, four mitochondrial genes, inflorescence and floral structures, and 59 conserved nuclear genes, respectively. Similar to the Angiosperm Phylogeny Group (APG) III system, the APG IV system[9] placed Magnoliidae and Chloranthaceae together as sister to a robust clade, including monocots and Ceratophyllales + eudicots. Furthermore, 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)[10], 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 m[11]. *Cinnamomum*, including SCT and six congeneric species, contributed to Taiwan's position as the largest producer and exporter of camphor in the nineteenth century, and the value of their wood was further enhanced by their massive trunk diameters—the largest diameters among flowering plants of Taiwan—and their aromatic, decay-resistance quality that has been attributed to the essential oil D-terpinenol[12]. *Antrodia cinnamomea* is a parasitic fungus that infects the trunks of SCT causing heart rot[13]. The fungus produces several medicinal triterpenoids that impede the growth of liver cancer cells[14] and act as antioxidants that protect against atherosclerosis[15]. Owing 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[16].

Here, we report a chromosome-level genome assembly of SCT. Comparative analyses of the SCT genome with those of ten other

[1]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan. [2]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan. [3]School of Medicine, National Taiwan University, Taipei, Taiwan. [4]Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan. [5]Plant Biology Department, University of Georgia, Athens, GA, USA. [6]These authors contributed equally: Shu-Miaw Chaw, Isheng J. Tsai. *e-mail: smchaw@sinica.edu.tw; ijtsai@sinica.edu.tw

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 seem 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)[17]. 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 the ground with pollinators, pathogens, herbivorous insect and soil microorganisms[17]. 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 *C. kanehirae*. SCT is diploid (2n = 24; Supplementary Fig. 1a) with an estimated genome size of 823.7 ± 58.2 Mb/1 C (Supplementary Figs. 1b and 2). We produced an assembly derived solely from 85× PacBio long reads (read N50 = 11.1 kb; contig N50 = 0.9 Mb) spanning 728.3 Mb. The consensus sequences of the assembly were corrected using 141× Illumina reads and further scaffolded with 207× 'Chicago' reconstituted chromatin and 204× Hi-C paired-end reads using the HiRise pipeline (Supplementary Fig. 3). A final, integrated assembly of 730.7 Mb was produced in 2,153 scaffolds, comprising 91.3% of the flow cytometry genome size estimate. The final scaffold N50 was 50.4 Mb with more than 90% in 12 pseudomolecules (Supplementary Table 1), presumably corresponding to the 12 SCT chromosomes.

Using a combination of reference plant protein homology support and transcriptome sequencing derived from various tissues (Supplementary Fig. 1c and Table 2) and ab initio gene prediction, 27,899 protein-coding gene models were annotated using the MAKER2 pipeline[18] (Supplementary Table 1). Of these, 93.7% were found to be homologous to proteins in the TrEMBL database and 50% could be assigned Gene Ontology terms using eggNOG-mapper[19]. The proteome was estimated to be at least 89% complete based on BUSCO[20] (benchmarking universal single-copy orthologs) assessment, which is comparable to other sequenced plant species (Supplementary Table 1). Orthofinder[21] clustering of SCT gene models with those from 12 diverse seed plant genomes yielded 20,658 orthologous groups (Supplementary Table 3). 24,148 SCT genes (86.56%) were part of orthologous groups with orthologues from at least one other plant species. 3,744 gene models were not orthologous to others, and only 210 genes were part of the 48 SCT-specific orthologous groups. Altogether, they suggest that the phenotypic diversification in magnoliids may be fuelled by de novo birth of species-specific genes and expansion of existing gene families.

Genome characterization. We identified 3,950,027 biallelic heterozygous sites in the SCT genome, corresponding to an average heterozygosity of 0.54% (one heterozygous single nucleotide polymorphism (SNP) per 185 bp). The alternative (non-reference) allele frequencies at these sites had a major peak around 50% consistent with the fact that SCT is diploid with no evidence for recent aneuploidy (Supplementary Fig. 4). The spatial distribution of heterozygous sites was highly variable, with 23.9% of the genome exhibiting less than 1 SNP locus per kb compared to 10% of the genome with at least 12.6 SNP loci per kb. Runs of homozygosity regions appeared to be distributed randomly across SCT chromosomes, reaching a maximum of 20.2 Mb in scaffold 11 (Fig. 1a). Such long runs of homozygosity regions have equal sequence coverage than the rest of the genome (Supplementary Fig. 5) and may be associated with selective sweeps, inbreeding or recent population bottlenecks. Genes located in these runs of homozygosity regions were found to be enriched in lignin biosynthetic process and galactose metabolism (Supplementary Table 4), which suggest some potential roles in the formation of lignin–carbohydrate complexes[22]. Pairwise sequentially Markovian coalescent[23] (PSMC) analysis based on heterozygous SNP densities implicated a continuous reduction of effective population size over the past 9 million years (Fig. 1b), with a possible bottleneck coincident with the mid-Pleistocene climatic shift 0.9 million years ago (Ma). Such patterns may reflect a complex population history of SCT associated with the geological history of Taiwan, including uplift and formation of the island in the late Miocene (9 Ma) followed by mountain building 5–6 Ma, respectively[24].

Transposable elements and interspersed repeats made up 48% of the genome assembly (Supplementary Table 5). The majority of the transposable elements belonged to long terminal repeat (LTR) retrotransposons (25.53%), followed by DNA transposable elements (12.67%). Among the LTRs, 40.75% and 23.88% retrotransposons belonged to Ty3/Gypsy and Ty1/Copia, respectively (Supplementary Table 5). Phylogeny of the 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, whereas Ty1/Copia elements were grouped into two sister clades (7,229 and 2,950 copies) (Supplementary Fig. 6). With the exception of two scaffolds, both Ty3/Gypsy and Ty1/Copia LTR transposable elements were clustered within the pericentromeric centres of the 12 largest scaffolds (Fig. 1c and Supplementary Fig. 7). In addition, the LTR-enriched regions (defined by 100 kb with an excess of 50% comprising LTR class transposable elements) had on average 35% greater coverage than the rest of the genome (Fig. 1c and Supplementary Fig. 8), 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 1), whereas introns are slightly longer in SCT owing to a higher density of transposable elements ($P < 0.001$, Wilcoxon rank-sum test; Supplementary Fig. 9).

As has been described for other plant genomes[25], the chromosome-level scaffolds of SCT exhibit low protein-coding gene density and high transposable element density in the centres of chromosomes, and increased gene density towards the chromosome ends (Fig. 1c). We identified clusters of a putative subtelomere heptamer, TTTAGGG, extending as long as 2,547 copies, which implicate telomeric repeats in plants[26] (Supplementary Table 6). In addition, 687 kb of nuclear plastid DNA-like sequences (NUPTs), 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. 10).

Phylogenomic placement of *C. kanehirae* sister to 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 (that is, one and only one homologue in all species) identified through OrthoFinder[21] gene family circumscription of all gene models from the SCT and 12 other seed plant genomes (see Methods). A single species tree was recovered through maximum likelihood analysis[27] of a concatenated supermatrix of the single-copy gene alignments and coalescent-based analysis using the 211

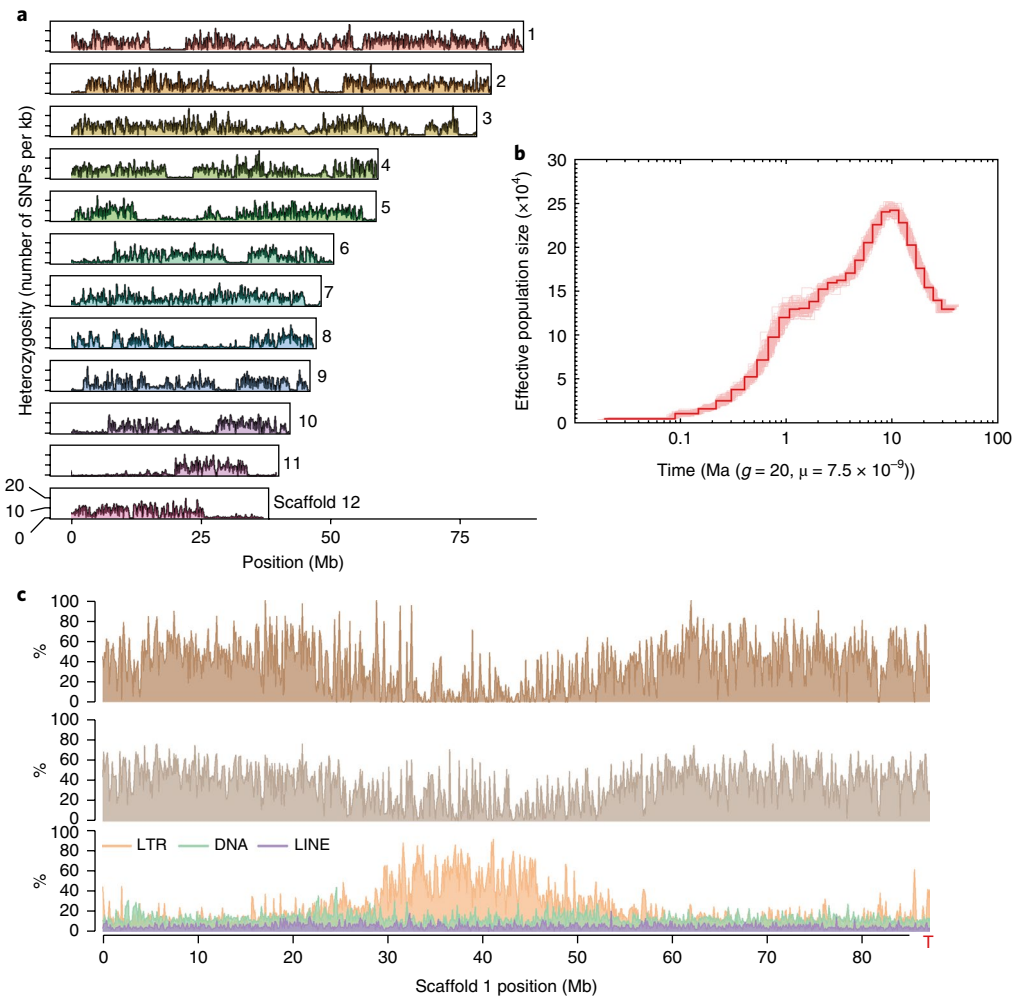

Fig. 1 | SCT genome heterozygosity. a, Number of heterozygous biallelic SNPs per 100-kb non-overlapping windows is plotted along the largest 12 scaffolds. Indels were excluded. **b**, The history of effective population size was inferred using the PSMC method. One hundred bootstraps were performed and the margins are shown in light red. **c**, For every non-overlapping 100-kb window, the distribution is shown from top to bottom: gene density (percentage of nucleotides with predicted model), transcriptome (percentage of nucleotides with evidence of transcriptome mapping) and three different classes of repetitive sequences (percentage of nucleotides with transposable element annotation). The red T letter denotes the presence of a telomeric repeat cluster at the scaffold end. LINE, long interspersed nuclear element.

gene trees[28] (Fig. 2 and Supplementary Fig. 11). SCT, representing the magnoliid lineage, was placed as sister to the eudicot clade (Fig. 2). This topology remained robust when we included a transcriptome data set of an additional 22 species of magnoliids order from the 1,000 plants initiative[29] (1KP), although lower bootstrap support was obtained (Supplementary Fig. 12). Using MCMCtree[30] with fossil calibrations, we calculated a 95% confidence interval for the time of divergence between magnoliids and eudicots to be 136.0–209.4 Ma (Fig. 2), which overlaps with two other recent estimates (114.8–164.1 Ma[31] and 118.9–149.9 Ma[32]).

Synteny analysis/whole-genome duplication. Previous investigations of expressed sequence tags data inferred a genome-wide duplication within the magnoliids before the divergence of the Magnoliales and Laurales[10], 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 whole-genome duplication (WGD) occurred in the ancestry of SCT (Fig. 3a). Two rounds of ancient WGD were implicated by extensive synteny between pairs of chromosomal regions and significant but less syntenic pairing of each region with two additional genomic segments (Supplementary Fig. 13). Synteny blocks of SCT's 12 largest scaffolds were assigned to five clusters that may correspond to pre-WGD ancestral chromosomes (Fig. 3a, Supplementary Fig. 13 and Supplementary 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 lineages[33]. To confirm that two rounds of WGD took place in the ancestry of SCT after divergence of lineages leading to SCT and *A. trichopoda*, we assessed synteny between the two genomes. Consistent with our hypothesis, one to four segments of the SCT genome were aligned to a single region in the *A. trichopoda* genome (Fig. 3b and Supplementary Fig. 14).

To more precisely infer the timing of the two rounds of WGD evident in the SCT genome, intragenomic and interspecies homologue Ks (synonymous substitutions per synonymous site) distributions were estimated. SCT intragenomic duplicates showed two peaks around 0.46 and 0.76 (Fig. 4a), congruent with the two WGD

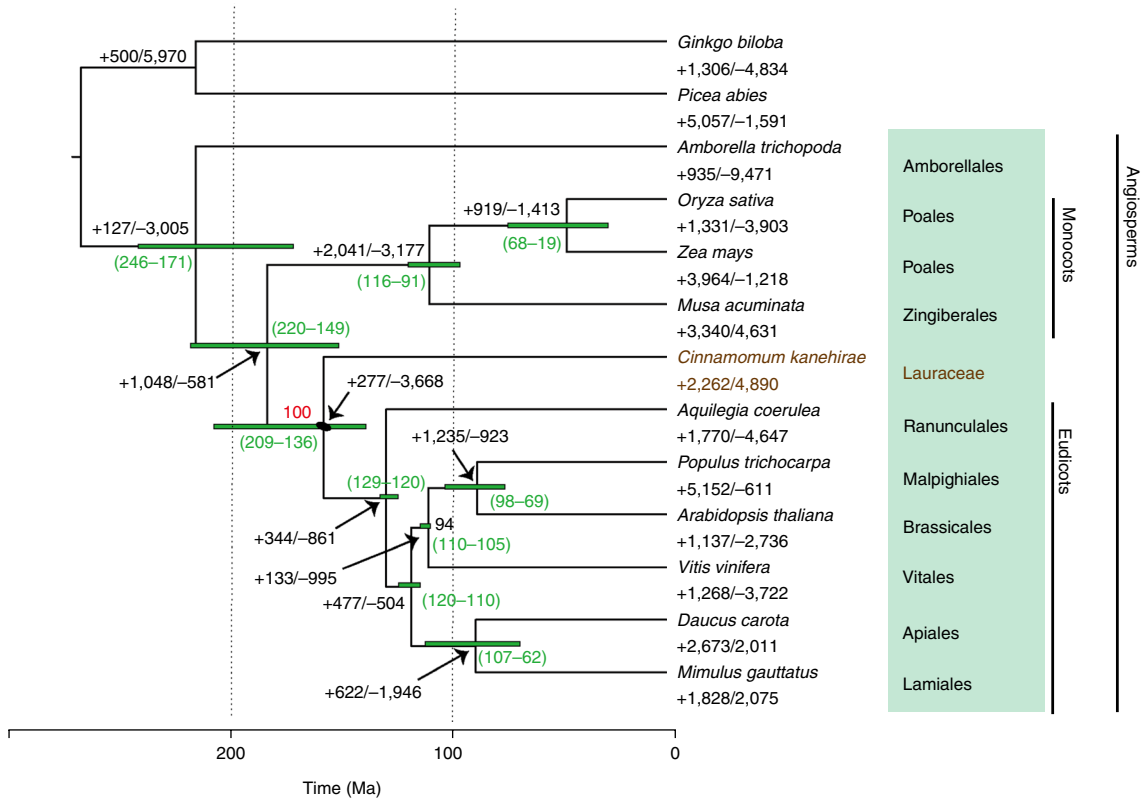

Fig. 2 | A species tree on the basis of 211 single-copy orthologues from 13 plant species. Gene family expansion and contraction are denoted in the numbers next to the plus and minus signs, respectively. The green numbers in the brackets denote divergence time estimates. All nodes' bootstrap support was 100 unless stated otherwise.

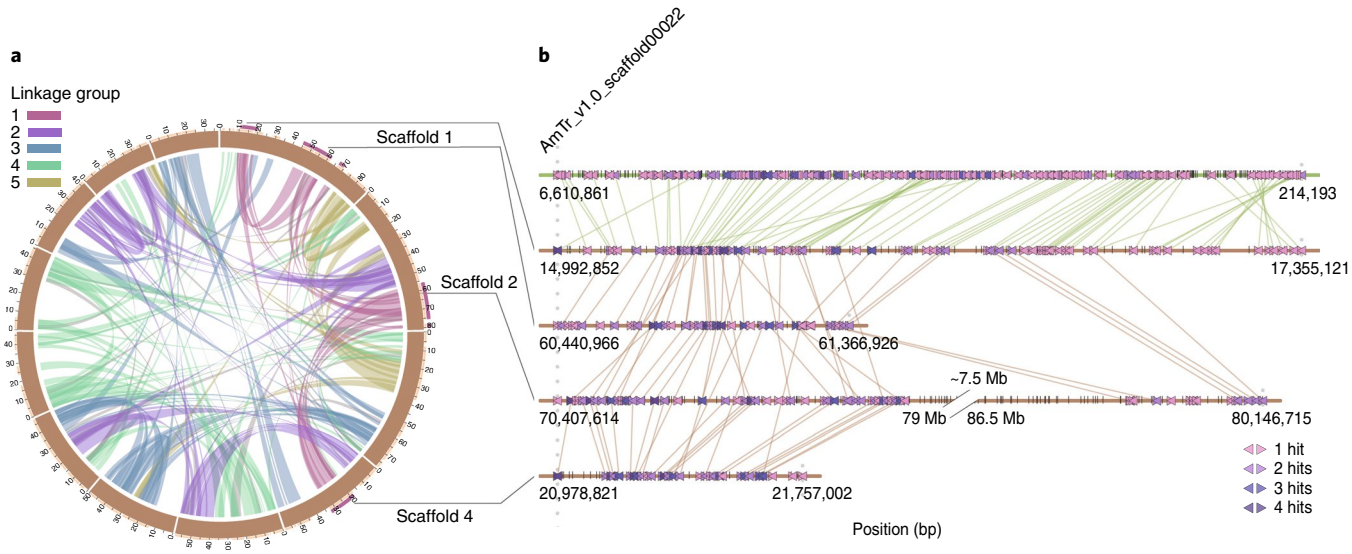

Fig. 3 | Evolutionary analysis of the SCT genome. a, Schematic representation of the intragenomic relationship among the 637 synteny blocks in the SCT genome. Synteny blocks (denoted by peach blocks) were assigned unambiguously into five linkage clusters representing ancient karyotypes and are colour coded. Purple blocks denote the synteny block assigned in the first linkage group (see also Supplementary Fig. 13). **b**, Schematic representation of the first linkage group within the SCT genome and their corresponding relationship in *A. trichopoda*.

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. 15). Comparison between *Aquilegia coerulea* (Ranunculales, a sister lineage to all other extant eudicots[33]) and SCT orthologues revealed a prominent peak around Ks = 1.41 (Fig. 4a), whereas the *Aquilegia* intragenomic duplicate was around Ks = 1, implicating independent WGDs following the

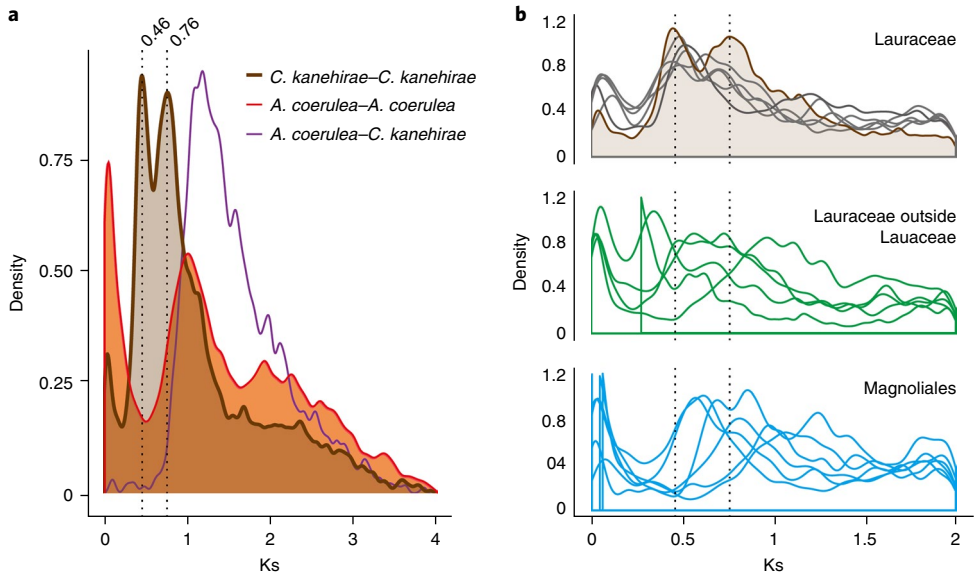

Fig. 4 | Density plots of synonymous substitutions (Ks) of the SCT genome and other plant species. a, Pairwise orthologue duplicates identified in synteny blocks within SCT, *A. coerulea* and between SCT and *A. coerulea*. **b**, Ks of intragenomic pairwise duplicates of the Lauraceae and the Magnoliales in the 1KP project[29]. Dashed lines denote the two Ks peaks observed in SCT. Brown and grey lines denote SCT and other Lauraceae's Ks distribution, respectively.

divergence of lineages leading to SCT and *Aquilegia*. The availability of the transcriptome of 17 Laurales + Magnoliales from the 1KP[29] allowed us to test the hypothesized timing of the WGDs evident in the SCT genome[8]. Ks distribution of five out of six available species from Lauraceae revealed two peaks (Fig. 4b and Supplementary Fig. 16), as was seen in the SCT Ks distribution (Fig. 4a) and corresponding to two synteny-based inferences of WGDs in the ancestry of SCT (Fig. 3 and Supplementary Fig. 15). Only one Ks peak was observed in the remaining Laurales and Magnoliales species, suggesting only one WGD event occurred in the ancestry of these species (Supplementary Figs. 17 and 18). The Ks peak seen in *Aquilegia* data is probably attributable to WGD within the Ranunculales well after the divergence of eudicots and magnoliids (Fig. 4a).

Specialization of the magnoliids proteome. We sought to identify genes and protein domains specific to SCT by annotating protein family (Pfam) domains 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 orthologous groups, 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. 19a). 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. 19b and Supplementary Table 9). Gain of protein domains included the terpene synthase (TPS) carboxy-terminal domain involved in defence responses and the leucine-rich repeats (628 versus 334.4) in plant transpiration efficiency[34]. 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*)[35]. EILs initiate an ethylene signalling response by activating ethylene response factor (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 ref. [35]; Supplementary Fig. 20). ERF responds and positively modulates biosynthesis of phytohormonal signals, including ethylene[36]. Expression of ERF has been implicated in positively modulating plant development from fruit ripening[35] to secondary growth in wood formation[37], as well as in increased resistance to abiotic[38] or biotic[39] factors. Thus, expansion of EILs in SCT may stimulate ERF, leading to various regulation of downstream effectors that result in traits specific to SCT.

We next assessed orthologous group expansions and contractions across the seed plant phylogeny (Fig. 2). 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 different gene families sharing common functions or single-gene families undergoing large expansions (Supplementary Tables 10 and 11). For example, expanded members of plant resistance (*R*) genes add up to 'plant-type hypersensitive response' (Supplementary Table 10). By contrast, the enriched Gene Ontology terms from the contracted gene families of the SCT branch (Supplementary Table 11) contain 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'.

***R* genes.** The SCT genome annotation included 387 *R* gene models, 82% of which belong to nucleotide-binding site leucine-rich repeat (NBS-LRR) or coiled-coil 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*) genes[40]. Among the sampled 13 genomes, SCT harbours the highest number of *R* genes among non-cultivated plants (Supplementary Fig. 21). The phylogenetic tree constructed from 2,465 NBS domains also suggests 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. 22).

***TPS* gene family.** One of the most striking features of the SCT genome is the large number of *TPS* genes (Ck*TPS*). A total of 101

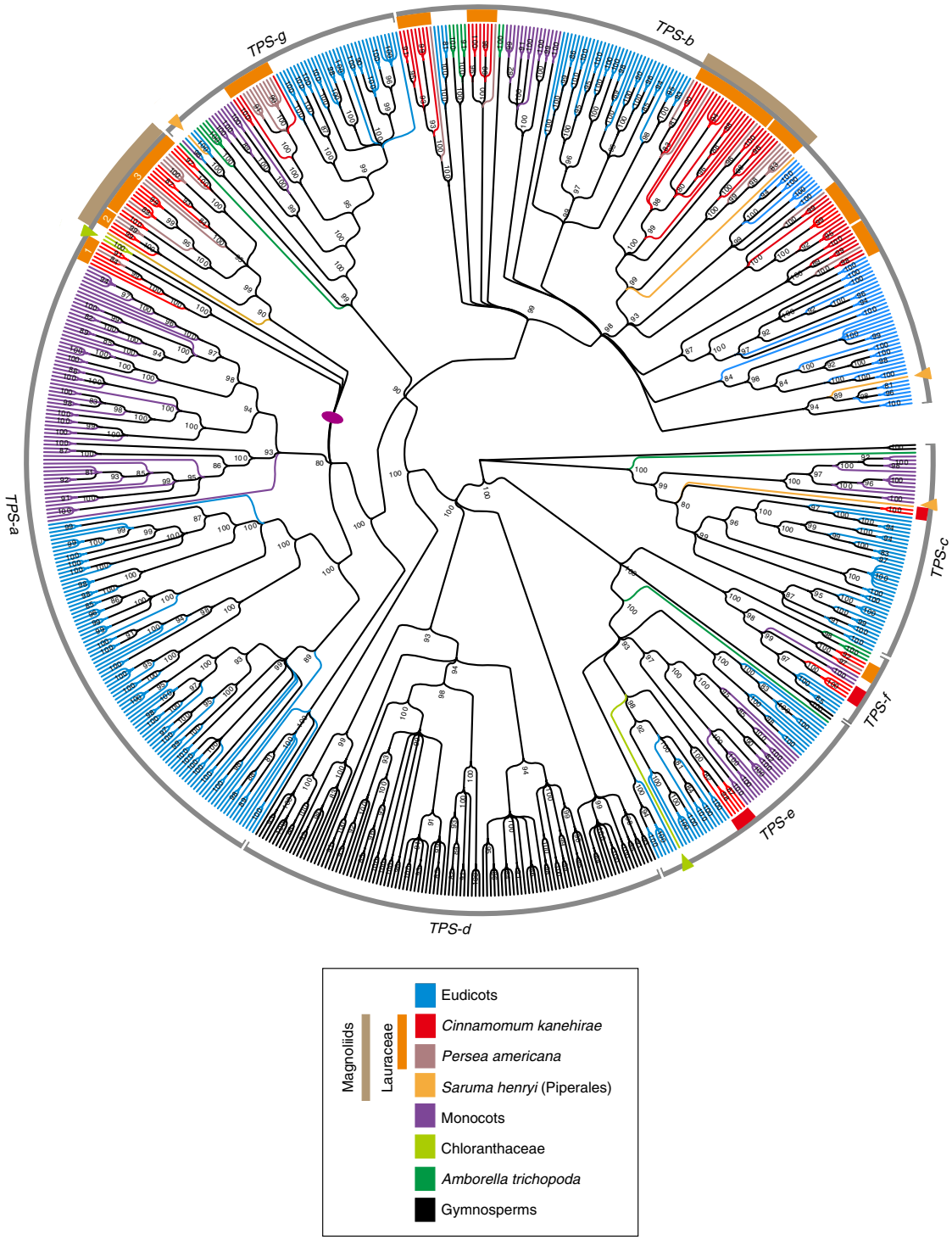

Fig. 5 | Phylogenetic placements of the 101 Ck*TPS* genes. The phylogenetic tree was constructed using putative or characterized *TPS* genes from 13 sequenced land plant genomes and two magnoliids with available transcriptomic data.

Ck*TPS* genes were predicted and annotated, the largest number for any other genome to date. By including a transcriptome data set of two more species from magnoliids (*P. americana* and *Saruma henryi*), phylogenetic analyses of *TPS* from 15 species placed Ck*TPS* genes among six of seven *TPS* gene subfamilies that have been described for seed plants[41] (Fig. 5, Table 1 and Supplementary Figs. 23–28). Ck*TPS* genes placed in the *TPS-c* (2) and *TPS-e* (5) subfamilies probably encode diterpene synthases, such as copalyl diphosphate synthase and *ent*-kaurene synthase[42]. These are key enzymes catalysing the formation of the 20-carbon isoprenoids (collectively termed diterpenoids; C20s), which were thought to be eudicot specific[41] and serve primary functions such as regulating plant primary metabolism. The remaining 94-predicted Ck*TPS* genes probably encode the 10-carbon monoterpene (C10) synthases, 15-carbon

Table 1 | Numbers of *TPS* subfamilies in the 13 genomes and three transcriptomes of major seed plant lineages

		Primary metabolism				Secondary metabolism			
***TPS* subfamilies**	**Genome size (Mb)**	**c**	**e**	**a**	**b**	**d**[a]	**f**	**g**	**Total no.**
Function species		**CPS, C20**	**KS, C20**	**C15**	**IspS, C10**	**C10, C15, C20**	**C20**	**C10**	
Gymnosperms									
G. biloba	10,609	1	1	–	–	49	–	–	51
P. abies	12,301	2	1	–	–	59	–	–	62
Angiosperms									
A. trichopoda	706	1	1	–	7	–	3	5	17
Chloranthaceae									
S. glabra[b]	–	–	1	2	–	–	–	–	3
Magnoliids									
Lauraceae									
C. kanehirae	731	2	5	25	58	–	7	4	101
P. americana[b]	–	–	–	11	12	–	1	9	33
Piperales									
S. henryi[b]	–	1	–	1	2	–	–	1	5
Monocots									
M. acuminata	473	2	2	21	13	–	3	3	44
O. sativa	375	3	10	19	–	–	–	1	33
Z. mays	2,068	8	6	30	2	–	–	5	51
Eudicots									
A. coerulea	307	15	13	12	34	–	–	8	82
Rosids									
A. thaliana	120	1	1	23	5	–	1	1	32
P. trichocarpa	473	2	2	16	14	–	1	3	38
V. vinifera[c]	434	2	1	29	10	–	2[c]	14	58
Asterids									
D. carota	422	3	2	1	15	–	1	7	29
M. guttatus	313	13	13	19	17	–	–	1	63

[a]*TPS-d* subfamily is gymnosperm specific. [b]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 are for reference only. [c]These two *TPS-f* were previously characterized from grape floral cDNA without identical genomic Vv*TPS* genes (Martin et al.[103]); Vv*TPS* sequences labelled as unknown (Martin et al.[103]) in the *TPS* gene tree were not counted. CPS, copalyl diphosphate synthase; KS, kaurene synthase; IspS, isoprene synthase.

sesquiterpene (C15) synthases and additional 20-carbon diterpene (C20) synthases (Table 1). With 25 and 58 homologues, respectively, *TPS-a* and *TPS-b* subfamilies are most diverse in SCT, presumably contributing to the mass and mixed production of volatile C15s and C10s[43].

It is noteworthy that the *TPS* gene tree resolved Lauraceae-specific *TPS* gene clades within the *TPS-a*, *TPS-b*, *TPS-f* and *TPS-g* gene subfamilies (Supplementary Figs. 23, 24, 27 and 28). 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* genes 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. 23). Indeed, we detected positive selection in the Lauraceae-specific *TPS-f* -I and -II subclades, implying functional divergence (Supplementary Table 13). Together, these data indicate increasing diversification of magnoliid *TPS* genes both before and after the origin of the Lauraceae.

Ck*TPS* genes are not uniformly distributed throughout the chromosomes (Supplementary Table 12) and clustering of members from individual subfamilies was observed as tandem duplicates (Supplementary Fig. 29). Seventy-six *TPS* genes were observed in the largest 12 scaffolds of SCT. Of those, 60.5% (46 copies) belonging to different subfamilies were found in the 0.5–15 Mb and 22.0–24.5 Mb region of scaffolds 7 and 10, respectively (Supplementary Fig. 29). Scaffold 7 contains 29 Ck*TPS* genes belonging to several subfamilies, including all of the eight Ck*TPS-a*, 12 Ck*TPS-b*, five Ck*TPS-e* and three Ck*TPS-f* (Supplementary Fig. 29). By contrast, only two members of Ck*TPS-c* reside in scaffold 1. Twenty-four Ck*TPS* genes are located in other smaller scaffolds, 22 of which encode the subfamily TPS-b (Supplementary Fig. 24). Some of these subfamilies located on scaffolds 7 and 10 are physically in proximity of each other (Supplementary Fig. 29). For instance, 3 out of 11 TPS-b-Lau III subfamily members were located adjacent to 4 out of 11 TPS-b-Lau V subfamily (Supplementary Fig. 29), whereas other subfamily members were found not in corresponding syntenic regions but elsewhere in the genome (Supplementary Fig. 30). Genes belonging to this cluster were not grouped together in their corresponding subfamily phylogeny (Supplementary Fig. 30), suggesting that their arrangement might have occurred more recently than the last WGD event.

Discussion

It is currently challenging to find wild SCT populations, making the conservation and basic study of this tree a priority. Camphor trees have been intensively logged since the nineteenth century, initially

for hardwood properties and association with the fungus *A. cinnamomea*. The apparent runs of homozygosity have been observed due to anthropogenic selective pressures or inbreeding in several livestock[44], although 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, respectively[24]. 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.

Our phylogenomic analyses of 211 single-copy orthologues from 13 representative seed plant genomes, including the first magnoliid representative, SCT, resolve magnoliids to be closer to eudicots than to monocots. This result disagrees with APG IV's resolution placing magnloliids as an outgroup to a clade containing monocots, Ceratophyllales and eudicots, but is in good agreement with a recent analysis of 59 orthologous nuclear genes based on transcriptome data of 26 seed plants[8]. Unfortunately, no complete genomic data of either Chloranthaceae or Ceratophyllacae are currently available for further re-examining the relationships of these two taxa, magnoliids, monocots, eudicots and the Amborella–Nymphaeles–Austrobaileyales grade. However, the placement of SCT as a sister to the eudicots in our analysis has important implications for comparative genomic analyses of evolutionary innovations within the eudicots, which comprise ~75% of extant flowering plants[8]. Consistent to early isozyme analysis[45], within the Lauraceae, we identified the timing of two rounds of independent WGD events that contributed to gene family expansions and innovations in pathogen, herbivore and mutualistic interactions. Large Ks peak ranges in the Laurales and Magnoliales from the 1KP transcriptome data set may be due to variation of synonymous substitution rates in the different lineages[29]. Complete genome assemblies for representatives of additional magnoliid lineages are needed to pinpoint the exact timing of these WGD events. The SCT genome will serve as an important reference outgroup for reconstructing the timing and nature of polyploidy events that gave rise to the hexaploid ancestor of all core eudicots (Pentapetalae)[46,47].

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 this subfamily[41]. Our data clearly indicate that this subfamily was present in the last common ancestor of all angiosperms but was lost from the grass family (Table 1). 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 C15s[43]. These findings are in congruent with the fact that fruiting bodies of the SCT-specific parasitic fungus *Antrodia cinamomea* can produce 78 kinds of terpenoids, including 31 structure-different triterpenoids (C30s)[48], many of which are synthesized via the mevalonate pathway, as are C10s and C15s followed by cyclizing squalenes ($C_{30}H_{50}$) into the skeletons of C30s[49]. It is reasonable to suggest that this fungus obtained intermediate compounds through decomposing trunk matters from SCT.

The 101 Ck*TPS* genes identified in the SCT genome are unevenly distributed across the 12 chromosomal scaffolds and include gene clusters from multiple subfamilies (Supplementary Fig. 30). 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* genes[50]. These rearrangement events may have also contributed to diversification of TPS enzymes in the SCT lineage and subsequent clustering of genes associated with mass production of terpenoids.

In summary, the availability of the SCT genome establishes a valuable genomic foundation that will help to unravel the genetic diversity and evolution of other magnoliids, and to give 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 broadleaved forest 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–July of 2014–2016. The tree was grown up from a seedling obtained from the Forestry Management Section, Department of Agriculture, Taoyuan City, Taiwan. The specimen (voucher number: Chaw 1501) was deposited in the Herbarium of Biodiversity Research Center, Academia Sinica, Taipei, Taiwan.

Genomic DNA extraction and sequencing. We used a modified high-salt method[51] to eliminate the high content of polysaccharides in SCT leaves, followed by total DNA extraction with a modified cetyltrimethylammonium bromide (CTAB) method[52]. 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 ~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 8Pac), these libraries were sequenced on a PacBio RS-II instrument using P6 polymerase and C4 sequencing reagent (Pacific Biosciences). Third, a Chicago and a Hi-C library were prepared by Dovetail Genomics (Santa Cruz) and sequenced on an Illumina HiSeq 2500 to generate 150-bp read pairs.

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 extracted[53]. The extracted RNA was purified using poly-T oligo-attached magnetic beads. All transcriptome libraries were constructed using the 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 method[54]. The stained samples were observed under a Nikon Eclipse 90i microscope (Supplementary Fig. 1a).

Genome size estimation. Fresh leaves of SCT were finely chopped with a new razor blade in 250 µl isolation buffer (200 mM Tris, 4 mM $MgCl_2$·$6H_2O$ and 0.5% Triton X-100) and mixed well, following the protocol of Dolezel et al.[55]. The mixture was filtered through a 40-µm nylon mesh, followed by incubation of the filtered suspensions with a DNA fluorochrome (50 µg ml^{-1} propidium iodide containing RNase A). Samples were analysed on the MoFlo XDP Cell Sorter (Beckman Coulter Life Science) and the Attune NxT Flow Cytometer (Thermo Fisher Scientific) in the Institute of Plant and Microbial Biology Flow Cytometry Analysis and Sorting Services at Academia Sinica, Taipei, Taiwan. Two and one replicates were performed on the former and latter machines, respectively, using chicken erythrocyte (BioSure) as an internal reference (Supplementary Fig. 1b). The 1 C genome size for SCT was estimated to be 781–890 Mb (Supplementary Figs. 1b and 2). Estimates of genome size from Illumina paired-end sequences were inferred using Genomescope[56] (version 1.0; based on k-mer 31).

De novo assembly of SCT. PacBio reads were assembled using the FALCON[57] (version 0.5.0) assembler. The consensus sequences of the assembly were further corrected using PacBio reads using Quiver[58] and Illumina reads using Pilon[59] (version 1.22). The PacBio assembly was scaffolded using the HiRISE scaffolder[60] (version July2015_GR), and consensus sequences were further improved using Pilon with one iteration[59]. The genome completeness was assessed using a plant data set of BUSCO[20] (version 3.0.2). To identify putative telomeric repeats, the assembly was searched for high copy number repeats less than 10 bp using tandem repeat finder[61] (version 4.09; options: 2 7 7 80 10 50 500). The heptamer TTTAGGG was identified (Supplementary Table 6).

Gene predictions and functional annotation. Transcriptome paired-end reads were aligned to the genome using STAR[62] (version 2.5.3a). Transcripts were identified using two approaches: (1) assembled de novo using Trinity[63] (version 2.3.2) and (2) reconstructed using Stringtie[64] (version 1.3.1c) as well as CLASS2 (ref. [65]) (version 2.1.7). Transcripts generated from Trinity were remapped to the reference using GMAP[66]. The three sets of transcripts were merged and filtered

using MIKADO[67] (version 1.1). Proteomes from representative reference species (Uniprot plants; proteomes of *Amborella trichopoda* and *Arabidopsis thaliana*) were downloaded from Phytozome (version 12.1; https://phytozome.jgi.doe.gov/). The gene predictor Augustus[68] (version 3.2.1) and SNAP[69] were trained either on the gene models predicted using BRAKER1 (ref. [70]) or MAKER2 (ref. [18]) (version 2.31.9). The assembled transcripts, reference proteomes, BRAKER1 and the BUSCO predictions were combined as evidence hints for input of the MAKER2 (ref. [18]) annotation pipeline. MAKER2 (ref. [18]) invoked the two trained gene predictors to generate a final set of gene annotation. Amino acid sequences of the proteome were functionally annotated using Blast2GO[71] and eggNOG-mapper[19] (version 1.0.3). NUPTs of SCT were searched against its plastid genome (plastome; KR014245 (ref. [72])) using blastn (parameters were followed from Smith et al.[73]).

Analysis of genome heterozygosity. Paired-end Illumina reads of SCT were aligned to reference using bwa mem[74] (version 0.7.17-r1188). PCR duplicates were removed using SAMtools[75] (version 1.8). Heterozygous biallelic SNPs were called using SAMtools[75] and consensus sequences were generated using bcftools[76] (version 1.7). Depth of coverage and alternative allele frequency plots were conducted using R version 3.4.2. The consensus sequence was fed to the PSMC program[23] to infer past effective population size. All of the parameters used for the PSMC program were at default with the exception of -u 7.5×10^{-9} taken from *A. thaliana*[77] and -g 20 taken from *Neolitsea sericea* (Lauraceae)[78].

Identification of repetitive elements. Repetitive elements were first identified by modelling the repeats using RepeatModeler[79] and then searched and quantified repeats using RepeatMasker[80]. Repeat types modelled as 'unknown' by RepeatModeler were further annotated using TEclass[81]. Tandem repeats were identified using Tandem Repeats Finder[61]. The proportions of different types of repeats were quantified by dissecting the 12 largest scaffolds into 100,000-bp chunks and calculating the total lengths and percentages of the repetitive elements within the chunks. LTR retrotransposons (LTR-RT) domains were extracted following Guan et al.'s method[82]. 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 the 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.[82]. Trees were built by aligning the sequences using MAFFT[83] (version 7.310; --genafpair --ep 0) and applied FastTree[84] with the Jones, Taylor and Thornton (JTT) model on the aligned sequences, and were coloured using the APE package[85].

Gene family or orthogroup inference and analysis of protein domains. The amino acid and nucleotide sequences of 12 representative plant species were downloaded from various sources: *A. coerulea*, *A. thaliana*, *Daucus carota*, *Mimulus guttatus*, *M. acuminata*, *Oryza sativa japonica*, *Populus trichocarpa*, *Vitis vinifera* and *Zea mays* from Phytozome (version 12.1; https://phytozome.jgi.doe.gov/), *Picea abies* from the Plant Genome Integrative Explorer Resource[86] (http://plantgenie.org/), *Ginkgo biloba* from GigaDB[87] and *A. trichopoda* from Ensembl plants[88] (release 39). Gene families or orthologous groups of these species and SCT were determined by OrthoFinder[21] (version 2.2.0). Pfams of each species were calculated from the Pfam website (version 31.0; https://pfam.xfam.org/). Pfam numbers of every species were transformed into z-scores. Significant expansions or reductions of Pfams in SCT were based on a z-score greater than 1.96 or less than −1.96, respectively. The significant Pfams were sorted by Pfam numbers (Supplementary Fig. 19). Gene family expansion and loss were inferred using CAFE[89] (version 4.1, with an input tree as the species tree inferred from the single-copy orthologues).

Phylogenetic analysis. MAFFT[83] (version 7.271; option --maxiterate 1000) was used to align 13 sets of amino acid sequences of 211 single-copy orthologous groups. Each orthologous group alignment was used to compute a maximum likelihood phylogeny using RAxML[27] (version 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-III[28]. A maximum likelihood phylogeny was constructed with the concatenated amino acid alignments of the single-copy orthogroups (version 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 the PAML[30] package (version 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 orthologues 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:(1) *A. thaliana*–*V. vinifera* (115–105 Ma)[90], (2) *M. acuminata*–*Z. mays* (115–90 Ma)[90], (3) Ranunculales (128.63–119.6 Ma)[32], (4) Angiospermae (247.2–125 Ma)[32], (5) Acrogymnospermae (365.629–308.14 Ma)[32] and (6) a hard bound of 420 Ma of outgroup *Physcomitrella patens*[91].

Analysis of genome synteny and WGD. Dot plots between SCT and *A. trichopoda* assemblies were produced using SynMap from the Comparative Genomics Platform (Coge[92]) to visualize the paleoploidy level of SCT. Synteny blocks within SCT and between *A. trichopoda* and *A. coerulea* were identified using DAGchainer[93] (same parameters as Coge:[92] -E 0.05 -D 20 -g 10 -A 5). Ks between syntenic group pairs were calculated using the DECIPHER[94] package in R. Depth of the inferred syntenic blocks were calculated using Bedtools[95]. Both the Ks distribution and the syntenic block depth were used to determine the paleopolyploidy level[96] 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).

***R* genes.** *R* genes were identified based on ref. [97]. Briefly, the predicted genes of the 13 sampled species were searched for the Pfam NBS (NB-ARC) protein family (PF00931) using HMMER version 3.1b2 (ref. [98]) with an *e*-value cut-off of 1×10^{-5}. Extracted sequences were then checked for protein domains using InterproScan[99] (version 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 MAFFT[83] (version 7.310; --genafpair --ep 0); the alignment was then input into FastTree[84] with the JTT model and visualized using EvolView[100].

***TPS* genes.** In addition to the 13 species proteome data set used in this study, transcriptome data from one Chloranthaceae species, *Sarcandra glabra*, and two magnollids representatives, *P. americana* (avocado) and *S. henryi* (saruma), were downloaded from the 1KP transcriptome database[29]. Previously annotated *TPS* genes of four species: *A. thaliana*[101], *O. sativa*[41], *P. trichocarpa*[102] and *V. vinifera*[103] were retrieved. For species without a priori *TPS* annotations, two Pfam domains: PF03936 and PF01397, were used to identify against the proteomes using HMMER[104] (version 3.0; cut-off at $e<10^{-5}$). Pseudogenes and sequence lengths shorter than 200 amino acids were excluded from further analysis. Putative or annotated protein sequences of TPS ($n=702$) were aligned using MAFFT[83] (version 7.310 with default parameters) and manually adjusted using MEGA[105] (version 7.0). The *TPS* gene tree was constructed using FastTree[84] (version 2.1.0) with 1,000 bootstrap replicates. The subfamily *TPS-c* was designated as the outgroup. Branching nodes with bootstrap values of <80% were treated as collapsed.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All of the raw sequence reads used in this study have been deposited in NCBI under the BioProject accession number PRJNA477266. The assembly of SCT is available under the accession number GCA_003546025.1.

Received: 17 July 2018; Accepted: 26 November 2018;
Published online: 9 January 2019

References

1. Hamidpour, R., Hamidpour, S., Hamidpour, M. & Shahlari, M. Camphor (*Cinnamomum camphora*), a traditional remedy with the history of treating several diseases. *Int. J. Case Rep. Imag.* **4**, 86–89 (2013).
2. Christenhusz, M. J. M. & Byng, J. W. The number of known plants species in the world and its annual increase. *Phytotaxa* **261**, 201–217 (2016).
3. 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 (2004).
4. Moore, M. J., Bell, C. D., Soltis, P. S. & Soltis, D. E. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc. Natl Acad. Sci. USA* **104**, 19363–19368 (2007).
5. Endress, P. K. & Doyle, J. A. Reconstructing the ancestral angiosperm flower and its initial specializations. *Am. J. Bot.* **96**, 22–66 (2009).
6. Qiu, Y.-L. et al. Angiosperm phylogeny inferred from sequences of four mitochondrial genes. *J. Syst. Evol.* **48**, 391–425 (2010).
7. Zhang, N., Zeng, L. P., Shan, H. Y. & Ma, H. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. *New Phytol.* **195**, 923–937 (2012).
8. Zeng, L. et al. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nat. Commun.* **5**, 4956 (2014).
9. Byng, J. W. et al. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* **181**, 1–20 (2016).
10. Cui, L. et al. Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16**, 738–749 (2006).
11. Liu, Y. C., Lu, F. Y. & Ou, C. H. Trees of Taiwan. *Monograph. Pub.* **7**, 105–131 (1988).
12. Fujita, Y. Classification and phylogeny of the genus *Cinnamomum* viewed from the constituents of essential oils. *Bot. Mag. Tokyo* **80**, 261–271 (1967).

13. Chang, T. T. & Chou, W. N. *Antrodia cinnamomea* sp. nov. on *Cinnamomum kanehirai* in Taiwan. *Mycol. Res.* **99**, 756–758 (1995).
14. Wu, S. H., Ryvarden, L. & Chang, T. T. *Antrodia camphorata* ("niu-chang-chih"), new combination of a medicinal fungus in Taiwan. *Bot. Bull. Acad. Sinica* **38**, 273–275 (1997).
15. Hseu, Y. C., Chen, S. C., Yech, Y. J., Wang, L. & Yang, H. L. Antioxidant activity of *Antrodia camphorata* on free radical-induced endothelial cell damage. *J. Ethnopharmacol.* **118**, 237–245 (2008).
16. Liao, P. C. et al. Historical spatial range expansion and a very recent bottleneck of *Cinnamomum kanehirae* Hay. (Lauraceae) in Taiwan inferred from nuclear genes. *BMC Evol. Biol.* **10**, 124 (2010).
17. Zerbe, P. & Bohlmann, J. Plant diterpene synthases: exploring modularity and metabolic diversity for bioengineering. *Trends Biotechnol.* **33**, 419–428 (2015).
18. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
19. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
20. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
21. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
22. Azuma, J.-I. & Tetsuo, K. Lignin–carbohydrate complexes from various sources. *Methods Enzymol.* **161**, 12–18 (1988).
23. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
24. Sibuet, J.-C. & Hsu, S.-K. How was Taiwan created? *Tectonophysics* **379**, 159–181 (2004).
25. Dong, P. F. et al. 3D chromatin architecture of large plant genomes determined by local A/B compartments. *Mol. Plant* **10**, 1497–1509 (2017).
26. Watson, J. M. & Riha, K. Comparative biology of telomeres: where plants stand. *FEBS Lett.* **584**, 3752–3759 (2010).
27. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
28. Mirarab, S. & Warnow, T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**, i44–i52 (2015).
29. Matasci, N. et al. Data access for the 1,000 Plants (1KP) project. *Gigascience* **3**, 17 (2014).
30. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
31. Massoni, J., Couvreur, T. L. & Sauquet, H. Five major shifts of diversification through the long evolutionary history of Magnoliidae (angiosperms). *BMC Evol. Biol.* **15**, 49 (2015).
32. Morris, J. L. et al. The timescale of early land plant evolution. *Proc. Natl Acad. Sci. USA* **115**, E2274–E2283 (2018).
33. Zhong, B. J. & Betancur-R, R. Expanded taxonomic sampling coupled with gene genealogy interrogation provides unambiguous resolution for the evolutionary root of angiosperms. *Genome Biol. Evol.* **9**, 3154–3161 (2017).
34. Lang, T. G. et al. Protein domain analysis of genomic sequence data reveals regulation of LRR related domains in plant transpiration in *Ficus*. *PLoS ONE* **9**, e108719 (2014).
35. Jourda, C. et al. Expansion of banana (*Musa acuminata*) gene families involved in ethylene biosynthesis and signalling after lineage-specific whole-genome duplications. *New Phytol.* **202**, 986–1000 (2014).
36. Gu, C. et al. Multiple regulatory roles of AP2/ERF transcription factor in angiosperm. *Bot. Stud.* **58**, 6 (2017).
37. Seyfferth, C. et al. Ethylene-related gene expression networks in wood formation. *Front. Plant Sci.* **9**, 272 (2018).
38. Chen, T. et al. Expression of an alfalfa (*Medicago sativa* L.) ethylene response factor gene *MsERF8* in tobacco plants enhances resistance to salinity. *Mol. Biol. Rep.* **39**, 6067–6075 (2012).
39. Wu, L., Zhang, Z., Zhang, H., Wang, X. C. & Huang, R. Transcriptional modulation of ethylene response factor protein JERF3 in the oxidative stress response enhances tolerance of tobacco seedlings to salt, drought, and freezing. *Plant Physiol.* **148**, 1953–1963 (2008).
40. Dodds, P. N. et al. Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proc. Natl Acad. Sci. USA* **103**, 8888–8893 (2006).
41. Chen, F., Tholl, D., Bohlmann, J. & Pichersky, E. 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 (2011).
42. Martin, D. M., Faldt, J. & Bohlmann, J. Functional characterization of nine Norway spruce *TPS* genes and evolution of gymnosperm terpene synthases of the *TPS-d* subfamily. *Plant Physiol.* **135**, 1908–1927 (2004).
43. Cheng, S. S. et al. Chemical polymorphism and composition of leaf essential oils of *Cinnamomum kanehirae* using gas chromatography/mass spectrometry, cluster analysis, and principal component analysis. *J. Wood Chem. Technol.* **35**, 207–219 (2015).
44. Peripolli, E. et al. Runs of homozygosity: current knowledge and applications in livestock. *Anim. Genet.* **48**, 255–271 (2017).
45. Soltis, D. E. & Soltis, P. S. Isozyme evidence for ancient polyploidy in primitive angiosperms. *Syst. Bot.* **15**, 328–337 (1990).
46. Jiao, Y. et al. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **13**, R3 (2012).
47. Chanderbali, A. S., Berger, B. A., Howarth, D. G., Soltis, D. E. & Soltis, P. S. Evolution of floral diversity: genomics, genes and gamma. *Phil. Trans. R. Soc.* **372**, 20150509 (2017).
48. Geethangili, M. & Tzeng, Y. M. Review of pharmacological effects of *Antrodia camphorata* and its bioactive compounds. *Evid. Based Complement. Alternat. Med.* **2011**, 1–17 (2011).
49. Lu, M. Y. J. 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–E4752 (2014).
50. Loehlin, D. W. & Carroll, S. B. Expression of tandem gene duplicates is often greater than twofold. *Proc. Natl Acad. Sci. USA* **113**, 5988–5992 (2016).
51. 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 (1989).
52. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15 (1987).
53. 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).
54. Suen, D. F. et al. Assignment of DNA markers to *Nicotiana sylvestris* chromosomes using monosomic alien addition lines. *Theor. Appl. Genet.* **94**, 331–337 (1997).
55. Dolezel, J., Greilhuber, J. & Suda, J. Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* **2**, 2233–2244 (2007).
56. Vurture, G. W. et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
57. Chin, C. S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
58. Chin, C. S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
59. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
60. Putnam, N. H. et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
61. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
62. Dobin, A., Davis, C. & Schlesinger, F. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
63. 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 (2013).
64. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
65. Song, L., Sabunciyan, S. & Florea, L. CLASS2: accurate and efficient splice variant annotation from RNA-seq reads. *Nucleic Acids Res.* **44**, e98 (2016).
66. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
67. Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L. & Swarbreck, D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *Gigascience* **7**, giy093 (2018).
68. 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**, S11 (2006).
69. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
70. 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 (2016).
71. Conesa, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
72. Wu, C. C., Ho, C. K. & Chang, S. H. The complete chloroplast genome of *Cinnamomum kanehirae* Hayata (Lauraceae). *Mitochondrial DNA A DNA Mapp. Seq. Anal.* **27**, 2681–2682 (2016).

73. 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 (2011).
74. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
75. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
76. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
77. Buschiazzo, 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 (2012).
78. 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 (2018).
79. Smit, A. & Hubley, R. RepeatModeler Open-1.0 (Institute for Systems Biology, 2015).
80. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0 (Institute for Systems Biology, 2015).
81. Abrusan, G., Grundmann, N., DeMester, L. & Makalowski, W. TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329–1330 (2009).
82. Guan, R. et al. Draft genome of the living fossil *Ginkgo biloba*. *Gigascience* **5**, 49 (2016).
83. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
84. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
85. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
86. Sundell, D. et al. The plant genome integrative explorer resource: PlantGenIE.org. *New Phytol.* **208**, 1149–1156 (2015).
87. Sneddon, T. P., Li, P. & Edmunds, S. C. GigaDB: announcing the GigaScience database. *Gigascience* **1**, 11 (2012).
88. Bolser, D., Staines, D. M., Pritchard, E. & Kersey, P. in *Plant Bioinformatics: Methods and Protocols* (ed. Edwards, D.) 115–140 (Springer, New York, 2016).
89. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).
90. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
91. Pryer, K. M. et al. Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. *Nature* **409**, 618–622 (2001).
92. Lyons, E. et al. Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with Rosids. *Plant Physiol.* **148**, 1772–1781 (2008).
93. Haas, B. J., Delcher, A. L., Wortman, J. R. & Salzberg, S. L. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20**, 3643–3646 (2004).
94. Wright, E. S. Using DECIPHER v2.0 to analyze big biological sequence data in R. *R J.* **8**, 352–359 (2016).
95. Quinlan, A. R. & Hall, I. M. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
96. Ming, R. et al. The pineapple genome and the evolution of CAM photosynthesis. *Nat. Genet.* **47**, 1435–1442 (2015).
97. Lozano, R., Hamblin, M. T., Prochnik, S. & Jannink, J. L. Identification and distribution of the NBS-LRR gene family in the Cassava genome. *BMC Genomics* **16**, 360 (2015).
98. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
99. Finn, R. D. et al. InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199 (2017).
100. He, Z. et al. Evolviewv2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic Acids Res.* **44**, W236–W241 (2016).
101. Aubourg, S., Lecharny, A. & Bohlmann, J. Genomic analysis of the terpenoid synthase (*AtTPS*) gene family of *Arabidopsis thaliana*. *Mol. Genet. Genomics* **267**, 730–745 (2002).
102. Irmisch, S., Jiang, Y. F., Chen, F., Gershenzon, J. & Kollner, T. G. Terpene synthases and their contribution to herbivore-induced volatile emission in western balsam poplar (*Populus trichocarpa*). *BMC Plant Biol.* **14**, 270 (2014).
103. Martin, D. M. et al. Functional annotation, genome organization and phylogeny of the grapevine (*Vitis vinifera*) terpene synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays. *BMC Plant Biol.* **10**, 226 (2010).
104. Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487–2489 (2013).
105. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).

Acknowledgements

We thank C.-y. Tsai for plant materials, C.-M. Hung for PSMC analysis and W.-Y. Ko for discussion branch site test. S.-M.C. was funded by an Investigators Award, Taiwan BioGenome Program and Biodiversity Research Center, Academia Sinica. I.J.T. was funded by a Career Development Award AS-CDA-107-L01, Academia Sinica. H.-M.K., C.-S.W. and C.-Y.H. were funded by a postdoctoral fellowship, Academia Sinica.

Author contributions

S.-M.C. conceived and initiated the study. I.J.T. and H.-M.K. performed the assembly and annotation of the SCT genome. L.-Y.C. and Y.-W.W. carried out the repeat analysis. E.S. carried out analysis of plastid DNA. C.-S.W., L.-N.W., H.-T.Y., C.-Y.H. and S.-M.C. conducted the experiments. I.J.T., Y.-C.L., H.-M.K., C.-Y.I.L. and J.L.-M. carried out the comparative genomics analysis. Y.-W.W., M.-H.H., K.-P.W. and S.-M.C. analysed the *R* genes. H.-Y.W., S.-M.C., C.-Y.H. and Y.-W.W. analyszed the terpene gene families. I.J.T., S.-M.C., and J.L.-M. wrote the manuscript. S.-M.C. and I.J.T. supervised the study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41477-018-0337-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.-M.C. or I.J.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

nature research

Corresponding author(s): Shu-Miaw Chaw and Isheng Jason Tsai

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Pacbio reads were assembled using the FALCON (ver. 0.5.0) assembler. The consensus sequences of the assembly were further corrected using Pacbio reads using Quiver and Illumina reads using Pilon (ver. 1.22). The Pacbio assembly was scaffolded using HiRISE scaffolder (ver. July2015_GR) and consensus sequences were further improved using Pilon with one iteration. Genome prediction were performed using open-source tools : STAR (ver. 2.5.3a), Trinity (ver. 2.3.2), Stringtie (ver. 1.3.1c) , CLASS2 (ver. 2.1.7), GMAP (2017 -> 2018-07-04), MIKADO (ver 1.1). BRAKER1, MAKER2 (ver 2.31.9), Blast2GO and eggnog-mapper (ver 1.0.3).
Data analysis	R (3.3.4) was used for statistical analyses and figure generation. Commercial software were not used.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

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

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For comparative genome analyses, 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 Resource (http://plantgenie.org/), Ginkgo biloba from GigaDB, and Amborella trichopoda from Ensembl plants (Release 39).
Data exclusions	No data exclusions in this manuscript
Replication	No replication in this manuscript
Randomization	No randomization in this manuscript as genomes assemblies were not allocated into experimental groups.
Blinding	No blinding in this manuscript as the data were not allocated into groups

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Tissues of about 0.25 square centimeters were excised from freshly collected leaves, and placed in plastic petri dishes (35 mm x 10 mm) on ice, and then sliced to fine pieces using a new razor blade in extraction buffer for 1~2 min.
Instrument	The data were collected using MoFlo XDP Cell Sorter (Beckman Coulter Life Science, Indianapolis, IN) and Attune NxT Flow Cytometer (Thermo Fisher Scientific Inc., Waltham, MA).
Software	Summit (ver5.3)
Cell population abundance	Flow cytometry was used for quantification purposely only, and no post-sorting fraction was collected.
Gating strategy	Filter-575-25 was used in gating. The FSC-Area/SSC-Area gate (R1) method was used to eliminate debris, cell fragments, and dead cells. Single cell and double cells were discriminated by using RPE-Height/ RPE-Area (R4).

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.