bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Chromosome-level genome assembly of *Torreya grandis* provides insights into the origin and evolution of gymnosperm-specific sciadonic acid biosynthesis

Heqiang Lou[1,6], Lili Song[1,6], Xiaolong Li[2,3,6], Weijie Chen[1], Yadi Gao[1], Shan Zheng[1], Zhangjun Fei[4,5,]*, Xuepeng Sun[2,3,]*, Jiasheng Wu[1,]*

[1]State Key Laboratory of Subtropical Silviculture, Zhejiang A&F University, Hangzhou, Zhejiang 311300, China.

[2]Collaborative Innovation Center for Efficient and Green Production of Agriculture in Mountainous Areas of Zhejiang Province, Zhejiang A&F University, Hangzhou, Zhejiang 311300, China.

[3]Key Laboratory of Quality and Safety Control for Subtropical Fruit and Vegetable, Ministry of Agriculture and Rural Affairs, Hangzhou, Zhejiang 311300, China.

[4]Boyce Thompson Institute, Cornell University, Ithaca, NY 14853, USA.

[5]U.S. Department of Agriculture-Agricultural Research Service, Robert W. Holley Center for Agriculture and Health, Ithaca, NY 14853, USA.

[6]These authors contributed equally.

*Email: zf25@cornell.edu; xs57@zafu.edu.cn; wujs@zafu.edu.cn

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

Species in genus *Torreya* are nut trees that produce dry fruits with a wide assortment of functions. Here, we report the 19-Gb chromosome-level genome assembly of *T. grandis*. The genome is shaped by an ancient whole genome duplication and recurrent LTR retrotransposon bursts. Comparative genomic analyses reveal key genes involved in reproductive organ development, cell wall biosynthesis and seed storage. Two genes encoding a C_{18} Δ^9-elongase and a C_{20} Δ^5-desaturase are identified in *T. grandis* to be responsible for sciadonic acid biosynthesis and both are present in diverse plant lineages except angiosperms. We demonstrate that the histidine-rich boxes of the Δ^5-desaturase are crucial for its catalytic activity. Methylome analysis reveals that methylation valleys of the *T. grandis* seed genome harbor genes associated with important seed activities, including cell wall and lipid biosynthesis. Moreover, seed development is accompanied by DNA methylation changes that possibly fuel energy production. This study provides important genomic resource for gymnosperms and unravels key enzymes for biosynthesis of sciadonic acid as a hallmark metabolite of gymnosperms.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Introduction

The emergence of seed plants consisting of angiosperms and gymnosperms marks a momentous event in the evolution of land plants and the change of earth environments. Angiosperms and gymnosperms diverged in Lower Mississippian[1], followed by rapid radiation of flowering plants resulting in approximately 352,000 extant species on earth compared to only 1000 species of gymnosperms. There is an apparent morphological/anatomical diversity and metabolic versatility between angiosperms and gymnosperms but the underlying genetic and biochemical mechanisms are largely elusive.

Torreya grandis, a gymnosperm species belonging to a small genus of the yew family (Taxaceae), is a useful multipurpose tree, providing timber, medicine, edible nuts and oil[2] (**Fig. 1a**). The first credible record of *T. grandis* as a medicinal source appears in Classic of the Materia Medica during the Three Kingdoms of China and dates back to the beginning of the 3rd century AD[3]. *T. grandis* is the only species in Taxaceae with edible seeds, which have been used as food for thousands of years in China due to the attractive aromatic flavor, beneficial nutrients and bioactive compounds[4-10]. Oils are enriched in seeds of *T. grandis* with an average content of 45.80-53.16%[6]. In all kinds of oils, sciadonic acid (SCA), a non-methylene-interrupted ω6 fatty acid, has been found to be enriched in *T. grandis*, and the content of SCA is over 10% in the kernel oil[10]. SCA has positive effects on human health, and functions in reducing inflammation, lowering triglycerides, preventing blood clots and regulating lipid metabolism[11-14]. Production of SCA has been detected in different lineages of gymnosperms and a handful of algae and ferns[15]. However, SCA is generally absent in flowering plants, with the exception of a few lower eudicots (e.g. Ranunculaceae)[16], thus leaving a puzzle on its origin and evolution in green plants.

Genome sequences are key to address critical questions of plant evolution. However, taxonomical sampling of published plant genomes is overwhelmed by flowering plants[17], while only a dozen of gymnosperm genomes have been sequenced largely due to their high heterozygosity and complexity. Here, we assembled a chromosome-scale reference genome for *T. grandis* using PacBio HiFi and Illumina sequencing. The large and highly contiguous assembly was anchored into 11 chromosomes, which lays a solid foundation for comparative genomic analysis through which genomic footprints underlying key morphological diversity of land plant lineages can be revealed. This study not only adds to our understanding on the adaptive evolution of land plants but also discovers two crucial enzymes that are required for SCA biosynthesis.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Information provided by this work will be useful for strategic design on improvement of SCA production, possibly through genetic engineering or synthetic biology approaches, and promote the utilization of *Torreya* genetic resources in multiple aspects in the future.

RESULTS AND DISCUSSION

Genome assembly and annotation

We generated a total of 1.93 Tb Illumina and 463.7 Gb PacBio HiFi reads for *T. grandis* (**Supplementary Table 1**), representing about 96.5× and 23.2× coverage, respectively, of the *T. grandis* genome that had an estimated size of ~20 Gb according to the *k*-mer analysis of the Illumina reads (**Supplementary Fig. 1a**). The final assembly had a size of 19,050,820,213 bp, comprising 11,811 contigs with an N50 size of 2.82 Mb (**Supplementary Table 2**). Using Hi-C reads of approximately 106.2× coverage, 18.87 Gb (99.1%) of the assembled contigs were grouped into 11 chromosomes (**Fig. 1b** and **Supplementary Fig. 1b**). All 11 chromosomes were found to be enriched with the 101-bp repetitive sequence unit that resembles the tandem centromeric satellite repeat known as the landmark of centromeres (**Fig. 1b**). Two and nine chromosomes were found with telomeric sequences (5'-TTTAGGG-3') at both and single ends, respectively (**Fig. 1b**). Assessment of the *T. grandis* genome using Merqury[18] revealed a consensus quality score of 46.9, equivalent to 99.998% base accuracy. BUSCO[19] evaluation indicated that 1,386 out of 1,614 land plant conserved orthologs were successfully captured by the *T. grandis* assembly, which was comparable with that of other gymnosperm genome assemblies (**Supplementary Table 3**). The LTR assembly index (LAI) for the *T. grandis* genome was 10.7, which was higher than the proposed standard for a reference genome[20]. These together with the high DNA (99.48%) and RNA (96.54%) read mapping rates suggested the high quality of the *T. grandis* genome assembly. The *T. grandis* genome harbored 11.4 Gb (59.8%) repetitive sequences, of which LTR retrotransposons (LTR-RTs; 87.0%) were predominant, followed by DNA transposons (7.1%) and long interspersed nuclear elements (LINEs; 3.1%) (**Supplementary Table 4**). The proportion of *Copia* LTR-RTs (11.6%) was relatively higher in *T. grandis* than in other gymnosperms, possibly due to recent species-specific bursts occurring in multiple subfamilies of LTR-RTs (**Fig. 1c**). Most of LTR-RT expansions in gymnosperms took place between 25-7 million years ago (mya; **Supplementary Fig. 2a**), overlapping with the geological time of Miocene epoch (23.03-5.33 mya)

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

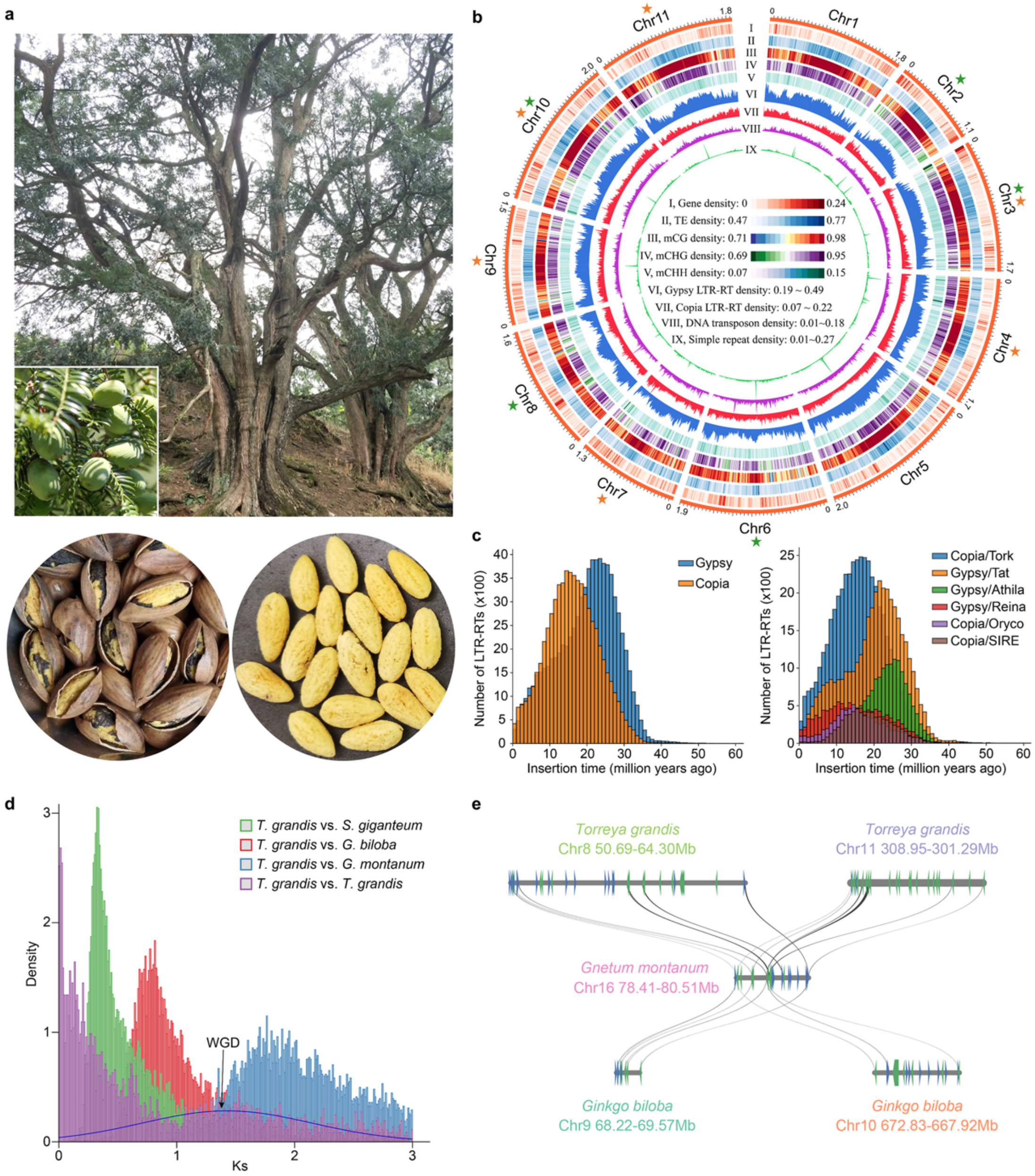

Figure 1. Genome of *T. grandis*. (**a**) Tree and fruit set of *T. grandis*. Lower panel shows the processed dry seed and its edible part (endosperm). (**b**) Circos plot of *T. grandis* genome and genomic features encoded by the chromosomes. Each feature was calculated based on a 10-Mb window across the chromosomes. Colored star indicates the presence of telomeric sequences on 5'- (green) or 3'-end (orange) of the chromosome. (**c**) Distribution of LTR-RT insertion time. Left panel shows all members of *Gypsy* and *Copia* families and the top six most abundant subfamilies are shown on the right panel. (**d**) *Ks* distribution of orthologues among *T. grandis*, *Sequoiadendron giganteum*, *Ginkgo biloba* and *Gnetum montanum*. *Ks* of paralogues in *T. grandis* was fitted with Guassian mixture model and the putative ancient WGD is indicated. (**e**) Micro-collinearity between genomes of *T. grandis*, *G. biloba* and *G. montanum*.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

when the earth cooled down towards ice ages[21], suggesting a potential environmental effect on the genome size evolution of gymnosperms.

A total of 47,089 protein-coding genes were predicted in the *T. grandis* genome, of which 46,338 were supported by homology and/or transcriptome evidence (**Supplementary Table 2**). In concordance with the large genome size, genes of *T. grandis* were longer than most of other gymnosperms, which can be attributed to the bloating of introns (**Supplementary Fig. 2b**). Intron size is more dynamic in gymnosperms than in angiosperms, in line with the fact that 70% of introns encoded LTR-RTs. In plants, LTR-RTs can be eliminated through unequal recombination, creating abundant solo-LTRs in the genome. The solo:intact LTR ratio in *T. grandis* was 4.3, higher than that in tomato (1.6) and Arabidopsis (2.6), suggesting a stronger force for LTR-RT removal in *T. grandis*. Moreover, epigenetic silencing of transposons and pericentromeric repeats is mediated by RNA-directed DNA methylation (RdDM) and 24-nt hetsiRNAs[22]. The *T. grandis* genome encoded homologues of key components of the RdDM pathway (**Supplementary Table 5**); however, small RNA profiling showed that 21-nt sRNAs were the most abundant in *T. grandis* (**Supplementary Fig. 3**), contrasting to 24-nt sRNAs in angiosperms, implying the presence of alternative pathways for transposon silencing in *T. grandis* and other gymnosperms[23].

Ancient whole-genome duplication

Whole-genome duplications (WGDs) have occurred across the breadth of eukaryote phylogeny[24]. In gymnosperms, several WGDs have been recognized although some of them remain in controversy[25-29]. The *Ks* distribution of 3,859 paralog groups within *T. grandis* indicated the absence of recent WGDs. However, we observed a peak of *Ks* ranging from 1 to 2 and a summit at 1.4, representing a potential ancient WGD that occurred in the common ancestor of conifers and ginkgophytes, a lineage diverged from gnetophytes (**Fig. 1d**). We then used a tree-based approach[30], which calculates the frequency of gene duplication on every branch of a phylogeny by reconciliation of gene tree and species tree, to cross validate the WGD event. Analysis of 19,649 gene trees from eight selected species led to the discovery of four ancient WGD signals including three (zeta, omega and another) reported previously[25,26,31] and one that was consistent with the *Ks* analysis (**Supplementary Fig. 4**). Whole genome comparison showed high collinearity among genomes of *T. grandis* and two evolutionarily distant gymnosperms, *Sequoiadendron giganteum* and *Ginkgo biloba* (**Supplementary Fig. 5**), but also revealed traces of collinear blocks that were

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

duplicated in both *T. grandis* and *G. biloba* but not in *Gnetum montanum*, agreeing with the timing at which the newly discovered WGD occurred (**Fig. 1e**).

Comparative genomics

Gene family evolution

We identified 19,362 orthologous groups (gene families) in 19 plant species comprising 7 gymnosperms and 12 representative species in major green plant lineages. Phylogeny and molecular dating using 219 low-copy gene families indicated that *T. grandis* separated from *T. wallichiana* around 68.5 mya (**Fig. 2a**). Expansion of gene families has been implicated in tight associations with morphological innovations[32,33]. Through reconstruction of gene family evolution, we found that bursts of gene family expansions coincided with major transitions of plant adaption (**Fig. 2a**). Massive gene family expansion (n=417, $P < 0.05$) was observed in the common ancestor of land plants, and subsequently in the extinct ancestors leading to seed plants (n=575), angiosperms (n=432) and various lineages of gymnosperms (n=428-818). Functions of expanded gene families were mostly associated with plant organ development, response to biotic (e.g., bacteria and fungi) and abiotic (e.g., water deprivation, light, temperature and salt) stresses, and biosynthesis and signaling of plant hormones. Many of the gene families were expanded continuously towards the evolution of higher plants, suggesting that gene duplication, possibly followed by sub/neo-functionalization, provides genetic foundation for morphological diversity and environmental adaptation in plants. Among the gene families that were significantly expanded in *T. grandis*, many of them encoded pfam domains associated with important biological functions, including lipid transfer (Oleosin and PF14368), biotic and abiotic stress responses (PF00201 and PF03018) and secondary metabolism (PF00067) (**Supplementary Table 6**).

Flower organ genes

Gymnosperms have unenclosed or naked seeds on the surface of scales or leaves, while flowers and fruits are angiosperm innovations. Phylogeny-based homologue search using well-studied flower development genes[34] showed sporadic distribution of the homologues in gymnosperms and non-seed plants (**Supplementary Table 7**), indicating the stepwise emergence accompanied with secondary loss of flower development genes during the evolution of land plants, as exemplified by *NOP10* (required for female gametophyte formation in flowers)[35] and *WUS* (required for shoot

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

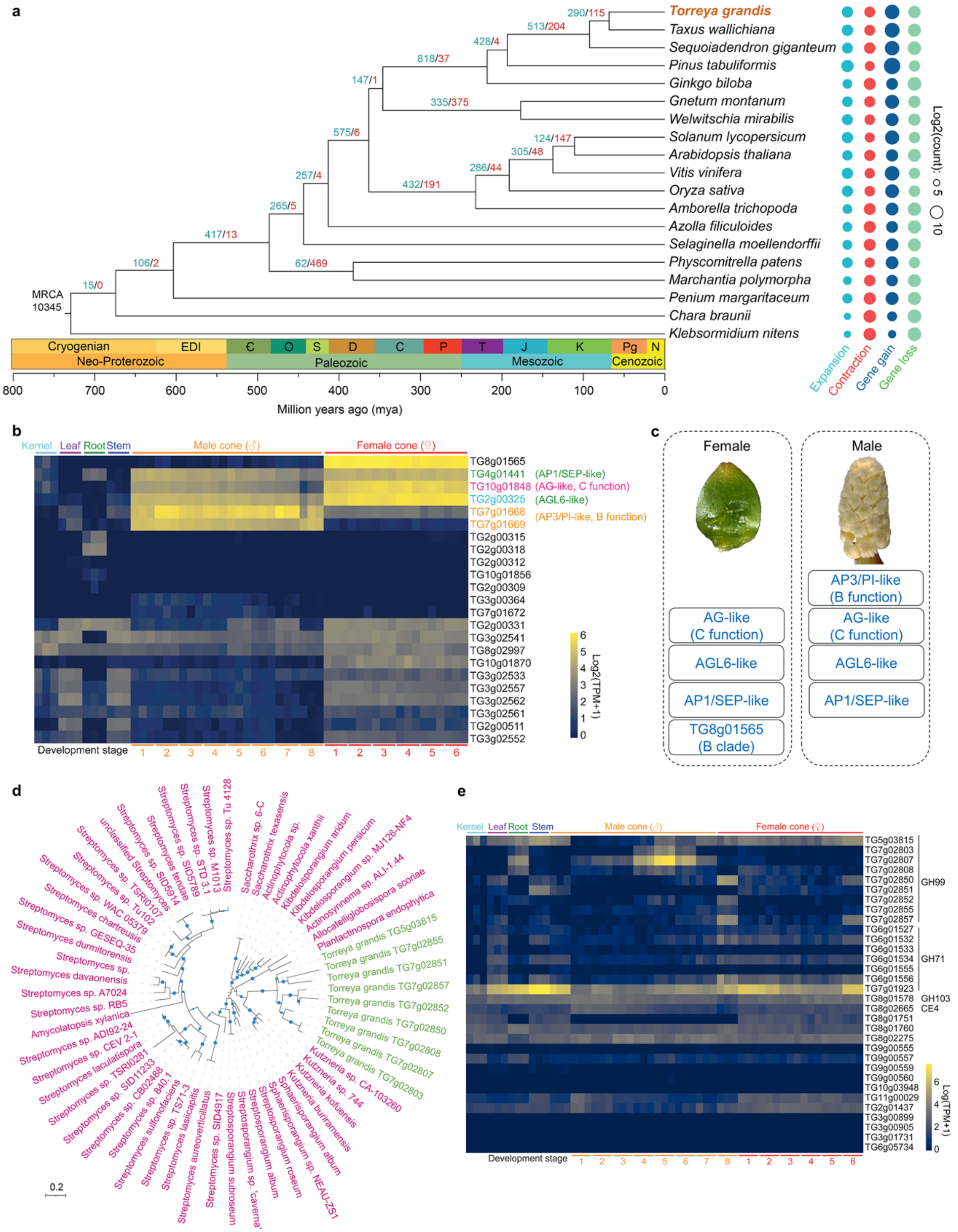

Figure 2. Gene family evolution in plants. (**a**) Gene family expansion and contraction during the evolution of green plants. The maximum likelihood phylogeny was built with 219 low-copy orthologous groups. Gene family analysis was started with 10,345 orthologous groups that were shared by the most recent common ancestor (MRCA) of green plants. Numbers on branches are the sizes of expanded (blue) and

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

contracted (red) gene families at each node. Colored pies on the right represent the sizes of expanded/contracted gene families as well as gained/lost genes for each leaf node of the tree. (**b**) Expression of MIKC^C type MADS-box genes in vegetative and reproductive tissues of *T. grandis*. Samples of male cone were collected from *T. grandis* tree at 8 different stages during February and April of 2021 with a time interval of 7 days. Female cones were collected at 6 different stages during January and April of 2021 with a time interval of 16 days. Each sampling was performed with three biological replicates. (**c**) Proposed reproductive organ identity genes in *T. grandis*. The AP3/PI-like genes (*TG7g01668* and *TG7g01669*) and *TG8g01565* were predominantly expressed in male and female cones, respectively. The AG-like (*TG10g01848*), AGL6-like (*TG2g00325*) and AP1/SEP-like (*TG4g01441*) genes were expressed in both female and male cones, with the first two showing a pattern biased to female cones. (**d**) Maximum likelihood phylogeny showing the bacterial origin of *T. grandis* genes. Blue pies indicate bootstrap support greater than 80% at the corresponding branches. (**e**) Expression of putative horizontally transferred genes in different tissues.

and floral meristem integrity)[36] genes that emerged early in land plants and were subsequently lost in both *T. grandis* and *T. wallichiana* (**Supplementary Table 7**).

The MADS-box family genes are a class of transcription factors involved in regulation of floral organ specificity, flowering time, and fruit development. We identified 23 MIKC^C MADS-box genes in *T. grandis*, including homologues of genes in the ABCE model of floral organ identity[37]. These included one AP1/SEP-like gene (A or E function), two AP3/PI-like genes (B function) and six AG-like genes (C function) (**Supplementary Fig. 6**). Transcriptome analysis of 18 samples from vegetative and reproductive organs revealed six MADS-box genes that were highly expressed in male and/or female cones of *T. grandis*, among which the two tandemly duplicated AP3/PI-like genes (*TG7g01668* and *TG7g01669*) were predominantly expressed in the male cones, while one AG-like gene (*TG10g01848*) was expressed in male but upregulated by 6.6-fold in the female cones (**Fig. 2b**). Recent studies suggest that *AGL6*, member of an ancient subfamily of MADS-box genes, is involved in the E function of floral development in rice, maize and wheat[38,39], while participates in A function in the basal angiosperm *Nymphaea colorata*[40]. In *T. grandis*, the AGL6-like gene (*TG2g00325*) showed an expression pattern similar to that of the C function genes, whereas the *AP1/SEP*-like gene (*TG4g01441*) was expressed at a moderately high level in both male and female cones, resembling an ancestral role of E function. Interestingly, the most highly expressed MADS-box gene (*TG8g01565*) was exclusively activated in the female

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

cones (**Fig. 2b**). This gene was phylogenetically clustered with B clade genes comprising *AP3*, *PI* and B-sister genes *TT16* and *GOA* (**Supplementary Fig. 6**); however, its expression pattern was opposite to that of the *AP3*/*PI*-like genes. In conclusion, our finding on the involvement of additional MADS-box genes in seed development of gymnosperms supports the basic "BC" model, where the C function genes are generally expressed in reproductive male and female organs and the B function genes are restricted to male reproductive organs[41], and suggests a more sophisticated regulatory system for reproductive organ development in gymnosperms **(Fig. 2c)**.

Seed storage proteins

The protein content of *T. grandis* seeds ranges from 10.34% to 16.43% depending on cultivars (He et al., 2016, Li et al., 2005). Genes encoding seed storage proteins (SSPs) including 2S albumins (n=0-7), 7S globulins (n=1-9) and 11S globulins (n=2-14) were identified in *T. grandis* and other gymnosperms but not in the earlier forms of plants (**Supplementary Table 8),** suggesting their origin in seed plants. Transcriptome analysis showed that genes encoding 2S albumins and 7S globulins were expressed at an exceptionally high level (average transcripts per million (TPM)=14,125) in the kernel of *T. grandis* seeds and the expression was increased during seed development (**Supplementary Fig. 7a**). In contrast, all SSP genes including 11S globulin genes, which were moderately expressed in the kernel, remained transcriptionally inactive in the vegetative tissues (**Supplementary Fig. 7a**). The 2S albumin proteins harbor numerous cysteine residues to form disulfide bridge within and between subunits[42]. We found that all of these residues were conserved in *T. grandis* although the whole protein sequences were considerably divergent from the angiosperm counterparts (**Supplementary Fig. 7b**). Homology modeling revealed a high degree of protein structure conservation between the 2S albumin proteins from *T. grandis* (e.g., TG11g02972) and sunflower, particularly in the region where α-helices are formed (**Supplementary Fig. 7c**). Likewise, most of residues involved in trimer formation and stabilization as well as in correct globular folding of 11S globulins from flowering plants[43] were conserved in *T. grandis* (**Supplementary Fig. 8**). Overall, the gene expression and structural analyses suggest a conservative role of the major SSPs in both gymnosperms and angiosperms.

Cell wall biosynthesis and horizontal gene transfer

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Gymnosperms are mainly woody plants and their genomes encode a large set of carbohydrate active enzymes (CAZymes) whose functions are closely associated with cell wall biosynthesis. Among the 19 selected representative plant species, *T. grandis* harbored more CAZymes than most others, particularly in the families of glycoside hydrolases (e.g., GH1, GH16, GH18, GH19, GH27, GH71, GH99, and GH152), GT61 glycosyltransferases, and PL1 polysaccharide lyases (**Supplementary Table 9**), many of which were also expanded in other gymnosperms. In contrast to most CAZyme families that were universally present in plants, we identified four families comprising 18 genes, GH71 (n=7), GH99 (n=9), GH103 (n=1), and CE4 (n=1), that were present only in gymnosperms and prior lineages but not in angiosperms (**Supplementary Table 9**). Phylogenetic analysis showed that these families were of possible bacterial origin (**Fig. 2d** and **Supplementary Fig. 9**). Through systematic analysis, we identified 14 additional *T. grandis* genes that were derived from horizontal gene transfers (HGTs; **Supplementary Table 10**). Most of these genes were expressed in different tissues of the plant (**Fig. 2e**), reinforcing the contribution of HGTs in the evolution of land plants[44].

Lignin is a major component of plant secondary cell wall and is derived from p-hydroxyphenyl (H), guaiacyl (G), and syringyl (S) monolignols. S-lignin is restricted to flowering plants and some lycophytes, whereas G- and H-lignin are fundamental to all vascular plants[45]. Consistently, two key genes for S-lignin biosynthesis, *F5H* and *COMT*, were found only in angiosperms but not in gymnosperms. Unlike angiosperms in which vessels comprise major water-conducting elements in xylem, gymnosperm woods are mainly composed of tracheids[46]. Vessel differentiation is regulated by VASCULAR-RELATED MAC-DOMAIN (VND) proteins[47], while fibre development is associated with NAC SECONDARY WALL THICKENING PROMOTING FACTOR (NST)/SECONDARY WALL-ASSOCIATED NAC DOMAIN PROTEIN (SND) proteins[48]. The *T. grandis* genome encoded genes homologous to *VND4/5/6*, but lacked homologues of *VND1/2/3*, *NST* and *SND1* (**Supplementary Fig. 10**) that are required for secondary cell wall biogenesis, suggesting that gain of these genes in angiosperms may have contributed to their vessel formation.

Paclitaxel biosynthesis

Paclitaxel (or taxol) is a well-known anticancer chemotherapy drug originally discovered in the family of Taxaceae. Core biosynthetic genes of paclitaxel are positionally clustered and include

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

those encoding taxadiene synthase (TS), cytochrome P450s (CYP450s) and acetyltransferases[28]. Paclitaxel and relevant metabolites were not detected in *T. grandis*. Concordantly, orthologues of two taxadiene synthases were absent in the *T. grandis* genome (**Supplementary Fig. 11**). Although *T. grandis* harbored acetyltransferases and CYP725 family genes, those functioning in paclitaxel biosynthesis were derived from gene duplication and neofunctionalization after speciation of *Torreya* and *Taxus* (**Supplementary Fig. 12** and **13**). Thus, key genes and biosynthetic cluster of paclitaxel were not present in *T. grandis*.

Discovery of key genes for sciadonic acid biosynthesis

Sciadonic acid (SCA) is a Δ^5-olefinic fatty acid and its biosynthesis requires the activity of C_{18} Δ^9-elongase and C_{20} Δ^5-desaturase that uses 18:2-phosphatidylcholine (PC) as the initial substrate (**Fig. 3a**). Δ^5-desaturases are known as the 'front-end' desaturases[49], which usually encode a cytochrome b5-like heme/steroid binding domain (PF00173) and a fatty acid desaturase domain (PF00487), whereas the Δ^9-elongases encode a GNS1/SUR4 family domain (PF01151) for long chain fatty acid elongation. The *T. grandis* genome encoded four desaturase genes and four elongase genes based on the domain search. However, only one desaturase (TgDES1) showed high similarity with the previously reported Δ^5-desaturase in *Anemone leveillei*[50], while two elongases were considered as putative Δ^9-elongases but only one (*TgELO1*) was highly expressed in seed kernels (**Supplementary Fig. 14**). Since the unsaturated fatty acids are abundant component of seed oils, we investigated the expression of *TgDES1* and *TgELO1* during the course of seed maturation. We found that SCA was accumulated in mature seeds, accompanied with the increased expression of *TgDES1*. Similar trend was observed for the expression of *TgELO1* and the content of its putative product cis-11,14-eicosadienoic acid (**Fig. 3b,c**). Study of subcellular localization showed that both TgELO1 and TgDES1 were co-localized with the marker of endoplasmic reticulum (ER) in *N. benthamiana* leaves (**Fig. 3d**), suggesting that they were bound to ER membrane, consistent with the subcellular location of known desaturases[51]. To further verify their function in SCA biosynthesis, we overexpressed both *TgELO1* and *TgDES1* in *A. thaliana*, which neither encodes orthologues of *TgELO1* and *TgDES1* nor produces SCA or its precursor $20:2^{\Delta 11,14}$-PC. Gas chromatography analysis showed that SCA was successfully synthesized in seeds of the transgenic line expressing *TgDES1* and *TgELO1*, demonstrating that *TgELO1* and *TgDES1* are required for SCA biosynthesis in *T. grandis* (**Fig. 3e**).

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

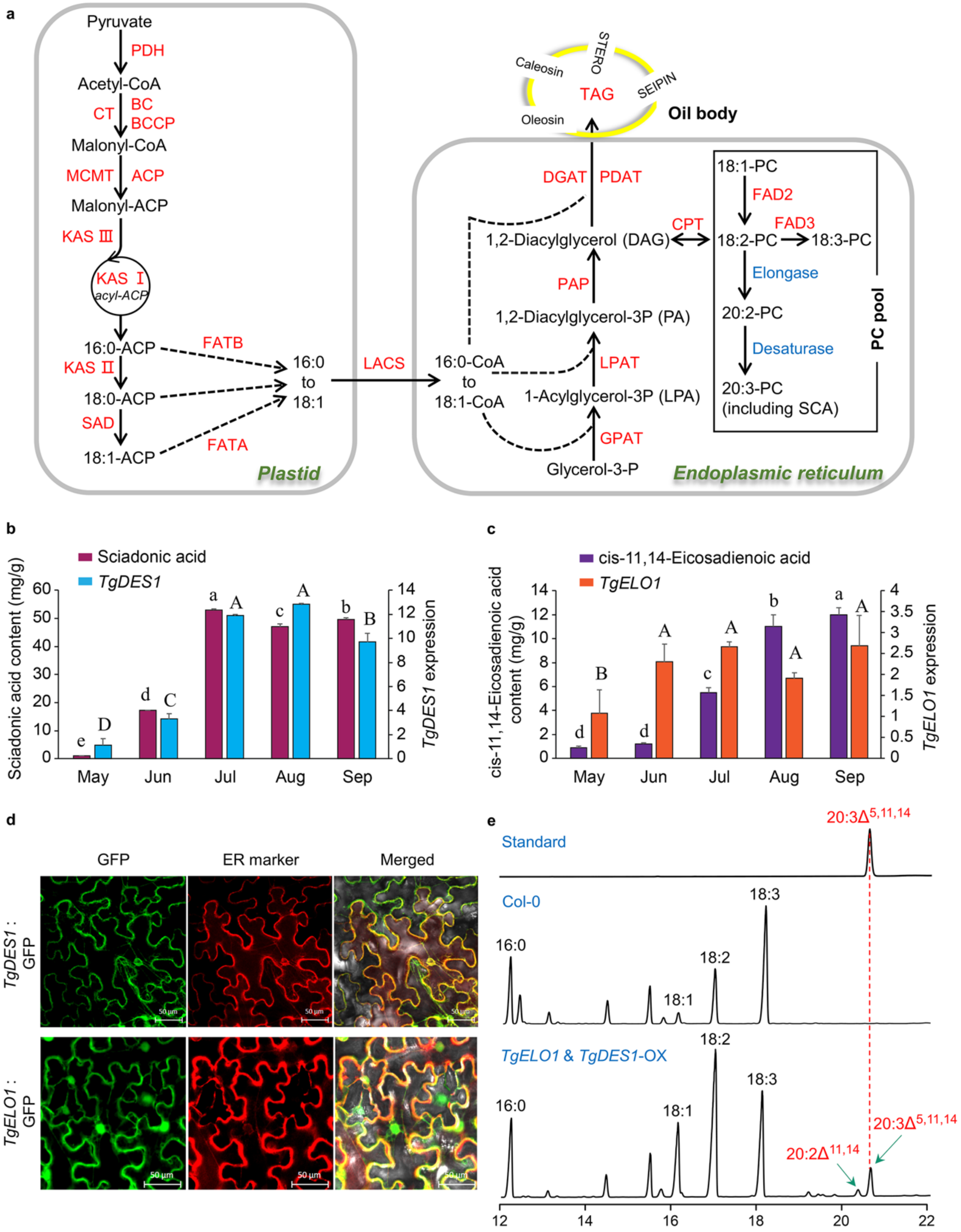

Figure 3. Characterization of genes responsible for SCA biosynthesis in *T. grandis*. (**a**) Overview of the pathway for fatty acid biosynthesis. PDH, pyruvate dehydrogenase. CT, carboxyltransferase. BC, biotin carboxylase. BCCP, biotin carboxyl carrier protein. MCMT, malonyl-CoA:ACP malonyltransferase. ACP, acyl carrier protein. KAS, ketoacyl-ACP synthase. SAD, stearoyl-ACP desaturase. FATA, acyl-ACP thioesterase A. FATB, acyl-ACP thioesterase B. LACS, long-chain acyl-CoA synthetase. DGAT,

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

diacylglycerol acyltransferase. PDAT, phospholipid:diacylglycerol acyltransferase. PAP, phosphatidic acid phosphatase. LPAT, lysophosphatidic acid acyltransferase. GPAT, glycerol-3-phosphate acyltransferase. CPT, cholinephosphotransferase. FAD2, oleate desaturase. FAD3, linoleate desaturase. PC, phosphatidyl choline. (**b**) *TgDES1* expression and SCA content in seeds from early development stage (May) to maturation stage (September). Different letters on the bars indicate statistical significance between samples at α=0.05. (**c**) Expression of *TgELO1* and the content of its product cis-11,14-Eicosadienoic acid in seeds. Different letters on the bars indicate statistical significance between samples at α=0.05. (**d**) Subcellular localization of TgDES1 and TgELO1 in *N. benthamiana* leaves. (**e**) Detection of SCA and its precursor in Arabidopsis Col-0 and the transgenic line overexpressing both *TgDES1* and *TgELO1*.

Origin and evolution of plant Δ^5-desaturases and Δ^9-elongases

Phylogenetic analysis of desaturases in green plants (Viridiplantae) showed that TgDES1 clustered with desaturases exclusively from non-angiosperm organisms, and this monophyletic clade was close to the family containing sphingolipid desaturases including AtSLDs from Arabidopsis (**Fig. 4a**). Interestingly, the TgDES1 clade clearly separated from the group harboring AL10 and AL21, two proteins that were found to be responsible for SCA biosynthesis in the basal eudicot *Anemone leveillei*[50]. Structure modeling of TgDES1, AtSLD2 and AL21 showed overall similar structures between TgDES1 and AtSLD2, particularly in the region where the active center was formed, whereas structure of AL21 was relatively diverged from TgDES1 (**Fig. 4b**). Since flowering plants rarely synthesize SCA, our phylogenetic and structural evidence suggested that this is possibly due to the loss of TgDES1 clade desaturases, while the ability of SCA biosynthesis in particular species of eudicots was largely attributed to the secondary gain of the Δ^5-desaturase activity of evolutionarily independent counterparts. Similarly, close homologues of TgELO1 were not found in flowering plants but present in early land plants and algae, suggesting the co-evolution of Δ^5-desaturase and Δ^9-elongase in plants (**Supplementary Fig. 15**).

Characterization of protein sequences revealed the conservation of an N-terminal cytochrome b5-like domain and three histidine-rich boxes of TgDES1 clade desaturases (clade 1) and their two closely related groups (group 1 and group 2 of clade 2), whereas striking variation was observed in the first two histidine-rich boxes among different groups (**Fig. 4c**). A previous study reported that site-directed substitution of histidine boxes could influence substrate chain-length specificity and selectivity[52]. The single amino acid substitution likely directs the outcome

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

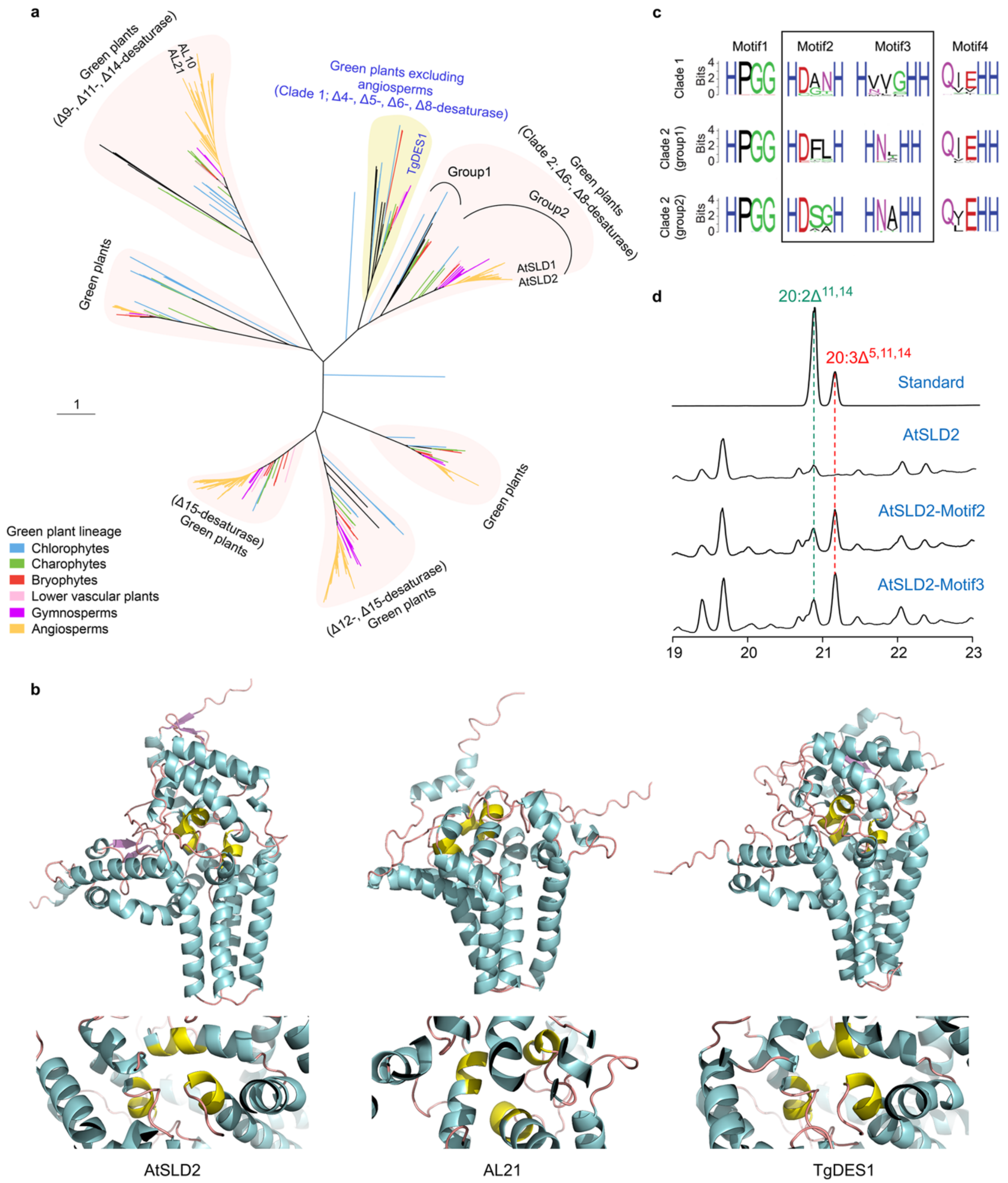

Figure 4. Origin and evolution of plant Δ^5-desaturases. (**a**) Maximum likelihood phylogeny of plant desaturases. TgDES1 is clustered within a group (clade 1) close to a sister clade (clade 2) comprising Δ^6- and Δ^8- desaturases. (**b**) Structure modeling of TgDES1 and the desaturases from Arabidopsis (AtSLD2) and *Anemone leveillei* (AL21). Protein structures were modeled with AlphaFold2 and the bioactive center of each protein comprising three histidine-rich motifs is marked with yellow. (**c**) Comparison of conserved motifs in different desaturase groups showing in (**a**). (**d**) Detection of SCA and its precursor in *N. benthamiana* leaves expressing Arabidopsis *AtSLD2* with conserved histidine-rich motifs (motif2 and

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

motif3) replaced by those from *TgDES1* of *T. grandis*. AtSLD2, AtSLD2-Motif2, AtSLD2-Motif3, are lines bearing Arabidopsis wild-type *AtSLD2* gene, or *AtSLD2* gene with motif2 or motif3 from *TgDES1*, respectively.

of the desaturation reaction by modulating the distance between substrate fatty acyl carbon atoms and active center metal ions[53]. To test whether sequence variation of histidine-rich domains determined substrate specificity that led to the success of SCA biosynthesis, we replaced the histidine-rich domain of Arabidopsis desaturase AtSLD2 with that of TgDES1, and transiently expressed the construct in *N. benthamiana* leaves. We noted that *TgELO1* was not coexpressed with the engineered desaturase gene because $20{:}2^{\Delta 11,14}$-PC, the product of Δ^9-elongase catalysis, could be detected in leaves of the wild-type tobacco. SCA was undetectable in *N. benthamiana* leaves expressing wild-type AtSLD2; however, switch of either of the two histidine-rich boxes from TgDES1 was sufficient to synthesize SCA in *N. benthamiana* leaves (**Fig. 4d**). Taken together, our data suggest that mutations in these two histidine-rich motifs of desaturases have led to the alternation of substrate specificity and consequently the evolution of specific clade for SCA biosynthesis, loss of which marks the significant metabolic diversity between gymnosperms and angiosperms.

Dynamics of DNA methylation during seed development

Seed development in gymnosperms is a long process spanning multiple years[54]. To understand whether and how DNA methylation participates in seed development of *T. grandis*, as is evident in flowering plants[55], we profiled seed methylomes at three developmental stages (**Fig. 5a**). Genes involved in DNA methylation of all three cytosine contexts (CG, CHG, CHH) were identified in the *T. grandis* genome (**Supplementary Table 5**). The global average methylation levels of mCG, mCHG, mCHH in *T. grandis* seed genome were 83%, 69% and 4%, respectively. Both mCG and mCHG methylation levels were higher than those in most of previously studied angiosperms[56], coinciding with the proposal of positive correlation between genome sizes and mCG/mCHG methylation levels[57]. mC of all sequence contexts was enriched at centromeric and peri-centromeric regions, despite that both mCG and mCHG were also broadly distributed in chromosome arms (**Fig. 1b**). In flowering plants, CG methylation is enriched within transcribed

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

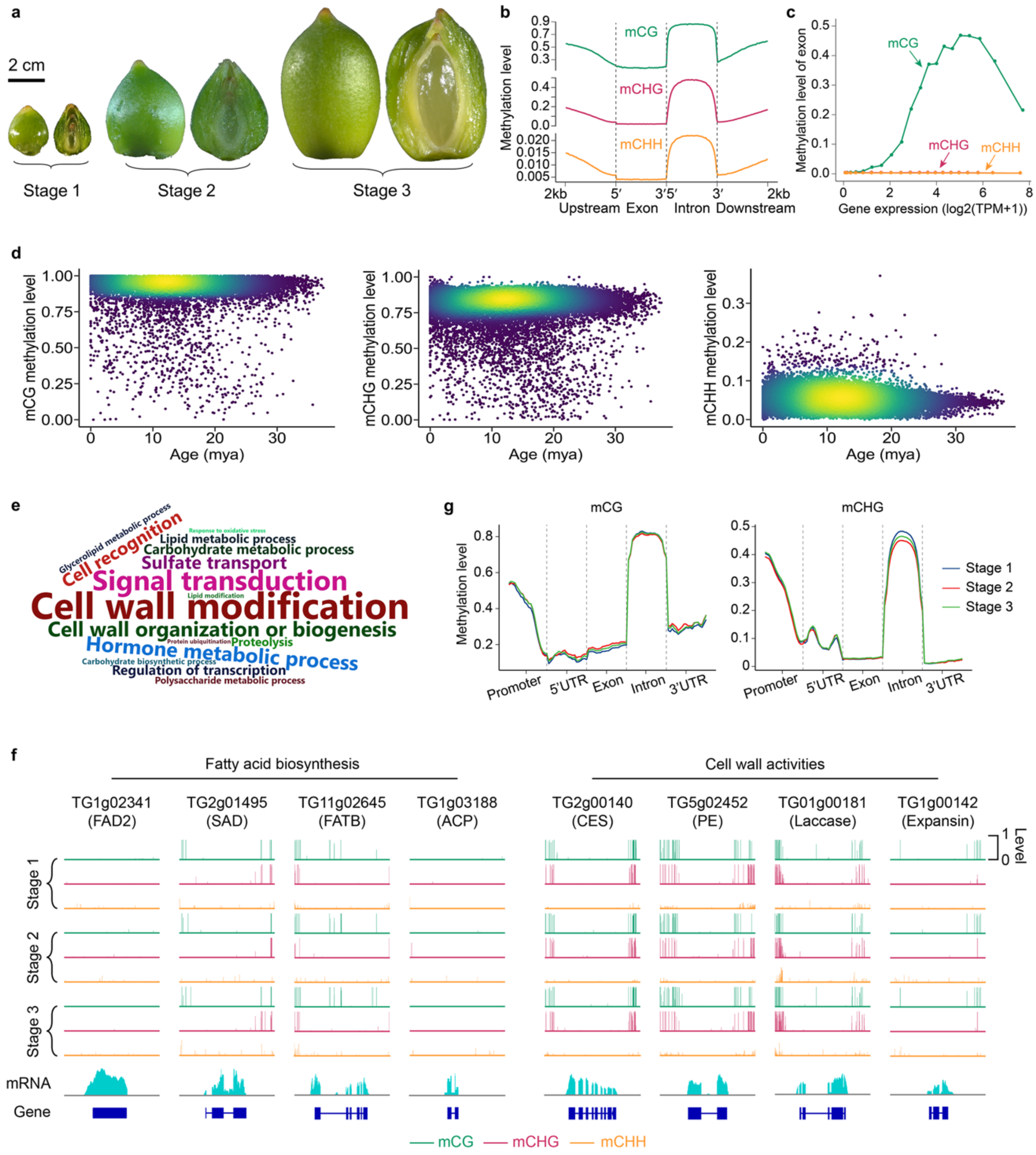

Figure 5. *T. grandis* seed methylomes. (**a**) Sampled seeds for methylome profiling. Pictures show the outer side and inner side (through longitudinal section) of the seeds. (**b**) Methylation levels of exon, intron and gene flanking regions in seeds. (**c**) Methylation levels at three cytosine contexts on exonic region of genes. Genes are categorized into 20 groups based on ordered expression levels. For each group, the median value of gene expression and the average methylation level across all exonic regions of genes are recorded. (**d**) Methylation levels of intact LTR-RTs in the *T. grandis* genome. (**e**) GO terms enriched in genes overlapping with demethylation valleys shared by seeds of all three developmental stages. GO terms with adjusted *P* value < 0.05 are plotted and sizes of GO terms in the word cloud figure correlate with their

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

statistical significance. (**f**) View of methylation levels of selected genes overlapping with demethylation valleys. (**g**) Comparison of mCG and mCHG methylation levels in different genomic regions of seeds at three stages.

regions and depleted at the transcription start (TSS) and end sites (TES), which is referred to as gene body methylation (gbM)[58]. We observed a depletion of DNA methylation at both TSS and TES of *T. grandis*, but the enrichment of mC in exons did not occur (**Fig. 5b** and **Supplementary Fig. 16a**). However, when genes were categorized based on their expression, we observed a clear enrichment of mCG instead of mCHG/mCHH on moderately expressed genes, for which the expression was positively correlated with methylation levels (**Fig. 5c** and **Supplementary Fig. 16b**), demonstrating the presence and potentially functional similarity of gbM in the sister lineage of angiosperms. Contrasting to exons, introns of *T. grandis* were highly methylated (**Fig. 5b**), agreeing with the dense distribution and heavy methylation of TEs within introns (**Fig. 5d**).

Seed development and germination genes are frequently localized within demethylation valleys (DMVs), where methylation level was low (e.g., <5%) for any of the cytosine contexts[59]. We identified 5,099 common DMVs in the seed genome of the three samples, which spanned 30 Mb including the largest interval extending to 144 kb. The DMVs intersected with 4,200 protein-coding genes, many of which encoded important classes of seed proteins, such as storage proteins, transcriptional factors, and enzymes for cell wall modification, hormone homeostasis and fatty acid biosynthesis (**Fig. 5e**). *T. grandis* seed is coated with a specialized outgrowth, called aril (**Fig. 5a**). During development, seed coat develops heavily lignified secondary cell walls to reinforce the outer surface of the seed[60], while preserves a soft inner one that directly surrounds the endosperm. Consistently, genes encoding laccases (n=38), which function in cell wall lignification[61], and expansins (n=13) that are associated with cell wall loosening[62], were frequently found in DMVs. Notably, 18% of *T. grandis* transcription factor (TF) genes (n=370) were located within seed DMV regions, representing a significant enrichment (χ^2 test; $P < 0.0001$; **Supplementary Table 11**). These TFs belonged to diverse gene families but were particularly abundant in MYB, NAC and AP2 families, which are known to regulate plant growth and development. The methylation of mCHH varied more remarkably than that of mCG and mCHG during seed development (**Supplementary Fig. 17**). We identified differentially methylated regions (DMRs) for each of the three cytosine contexts (**Supplementary Tables 12-14**). Among

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

genes overlapping with DMRs, 12% of them were differentially expressed, suggesting the translation of epigenetic variation to gene expression flexibility during seed development (**Supplementary Fig. 18**). GO enrichment analysis showed that DMR-associated genes were mainly enriched with those involved in photosynthesis and secondary metabolism (**Supplementary Table 15**), in line with the fact that photosynthesis fuels energy-generating biochemical pathways by contributing oxygen to seed tissues during the development of green seeds as developing seeds suffer from limited penetration of oxygen, particularly into inner tissues[63].

CONCLUSION

Gymnosperms are considered as a treasure trove of life history on the earth. Here, we assembled a chromosome-level reference genome for a gymnosperm species *T. grandis*. The genome size is huge and much larger than most of plant species ever sequenced. Based on this assembly and analysis of multi-omics data, we conclude that (1) recurrent LTR-RT bursts contribute to the bloating of the *T. grandis* genome, whereas *T. grandis* counteracts TE expansion through unequal recombination and epigenetic silencing with a mechanism potentially different from angiosperms; (2) ancient WGDs had occurred in *T. grandis,* while tandem duplications are frequently observed in gene families; (3) expansion and neofunctionalization of gene families are the driving force for plant adaptative evolution, and gain or loss of important gene families, e.g., MADS-box proteins and CAZymes, are likely associated with morphological diversity between gymnosperms and angiosperms; (4) SCA biosynthesis requires the Δ^9-elongase and Δ^5-desaturase, which are co-evolved and have been lost in flowering plants; (5) substrate specificity of Δ^5-desaturase is determined by the two histidine-rich boxes, mutation on which may lead to the alternation of substrate recognition, and subsequently the change of its product; (6) the seed genome of *T. grandis* comprises both heavily methylated repeat sequences and demethylation valleys, latter of which intersect with genes exerting important seed functions such as cell wall modification and fatty acid biosynthesis, as well as regulation of gene expression and hormone homeostasis. Overall, our high-quality reference genome coupled with comparative and functional genomic analyses fill an important gap in understanding of gymnosperm biology, particularly in the biosynthesis and evolution of SCA that features metabolic versatility between the major land plant lineages.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

METHODS

Plant materials and sequencing

Young leaves from a plant of *T. grandis* grown in Shaoxing, China, were collected in March 2018 and used for high-molecular-weight DNA extraction with DNeasy Plant Mini Kit (Qiagen). A paired-end (PE) library with insertion size of 350 bp was constructed using the Illumina Genomic DNA Sample Preparation kit following the manufacturer's instructions (Illumina), and sequenced on an Illumina NovaSeq system with a read length of 150 bp. A PacBio SMRTbell library was constructed using SMRTbell Express Template Prep Kit 2.0 and sequenced on a PacBio Sequel II platform. The circular consensus reads (HiFi reads) were generated using ccs software (https://github.com/pacificbiosciences/unanimity/) with parameter '--minPasses 3'. For Hi-C sequencing, libraries were prepared using leaf tissues fixed in 2% formaldehyde. Nucleus extraction and permeabilization, chromatin digestion and proximity-ligation treatments were performed following the protocol described elsewhere[64]. The restriction endonuclease DpnII was used to digest genomic DNA, followed by Hi-C DNA recovery and subsequent DNA manipulations as described previously[65]. The Hi-C libraries were sequenced on an Illumina NovaSeq platform with the read length of 150 bp.

To assist gene prediction, transcriptome sequencing was performed for samples collected from leaf, root, stem, young seed, aril, seed coat and kernel tissues of the same plant (**Supplementary Table 1**). Total RNA was extracted using TRIzol Reagent (Invitrogen) and quantified with NanoDrop ND-2000 spectrophotometer (NanoDrop Technologies). cDNA was synthesized from total RNA and used for library construction with the NEBNext Ultra II RNA Library Prep Kit for Illumina (NEB) following the manufacturer's instructions. The RNA-Seq libraries were sequenced on an Illumina NovaSeq platform under 2×150-bp mode. For PacBio Iso-seq, total RNA from leaf, root, stem, aril and kernel tissues were pooled equally and cDNA was synthesized using the SMARTer PCR cDNA Synthesis Kit (Clontech). Size fractionation and selection (1-2, 2-3 and 3-6 kb) were performed using the BluePippin Size Selection System (Sage Science). The SMRT libraries were generated using the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences) and sequenced on the PacBio RSII platform.

Genome assembly and quality assessment

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The HiFi reads were assembled using hifiasm[66] (version 0.8-dirty-r280) with default parameters and the assembled contigs were further polished by Racon (https://github.com/lbcb-sci/racon; v1.4.13) with Illumina reads. Purge Haplotigs[67] (version v1.1.0) was used to filter out redundant sequences in the assembly. Illumina reads from Hi-C libraries were processed with Trimmomatic[68] (v0.36) to remove adaptors and low-quality sequences. The cleaned reads were then used for scaffolding of the assembled contigs into chromosomes using ALLHiC[69] (version 0.9.8). Quality of the assembly was evaluated using the Illumina sequencing reads, which were mapped to the genome assembly using BWA-MEM[70].

Repeat annotation and LTR insertion time estimation

Repetitive sequences were identified using a combination of knowledge and *de novo* based approaches. A species-specific TE library for *T. grandis* was constructed to include LTR retrotransposons (LTR-RTs) and other TE elements identified by LTR_Finder[71] and RepeatModeler[72], respectively. This library was then combined with the Repbase library[73] for TE identification by RepeatMasker[74] (v.4.0.7). Repetitive elements were also predicted by RepeatProteinMask and the tandem repetitive sequences were identified by the TRF program[75]. To estimate LTR-RT insertion times, intact LTR-RTs were searched by LTR_Finder and LTR-harvest[76]. MUSCLE[77] was used to align LTR sequences of intact LTR-RTs, and the nucleotide distance (*K*) between them was calculated with the Kimura two-parameter criterion using the distmat program in the EMBOSS package (http://emboss.sourceforge.net). The insertion time (*T*) was calculated as $T = K/2r$, where the rate of nucleotide substitution (*r*) used for gymnosperm species was 2.2×10^{-9} per base per year[23].

Gene prediction and gene set assessment

Protein-coding genes were predicted using repeat-masked genome sequences. AUGUSTUS[78], GlimmerHMM[79], SNAP[80], geneid[81] and GENSCAN[82] were used for *ab initio* gene prediction. For homology-based prediction, protein sequences from one moss (*Physcomitrella patens*), one fern (*Selaginella moellendorffii*), seven angiosperms (*Amborella trichopoda, Arabidopsis thaliana, Oryza sativa, Phalaenopsis equestris, Populus trichocarpa, Vitis vinifera* and *Zea mays*) and four gymnosperms (*Ginkgo biloba, Gnetum montanum, Picea abies,* and *Pinus taeda)* were aligned to the *T. grandis* genome using tblastn[83] with an e-value cutoff of 1E-5. GenBlastA[84] was then applied

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

to cluster adjacent high-scoring pairs from the same protein alignments, and accurate gene structures were identified with GeneWise[85] (v.2.4.1). Raw RNA-Seq reads were cleaned with Trimmomatic and mapped to the *T. grandis* genome using TopHat2[86]. Subsequently, Cufflinks[87] (v.2.2.1) was employed to predict gene models. Finally, all genes predicted with the three approaches were integrated to generate a high-confidence gene set with EVidenceModeler[88].

To evaluate the accuracy of predicted genes, we examined the coverage of highly conserved genes using BUSCO[19]. We further performed functional annotation of the *T. grandis* predicted gene models by searching against the databases Kyoto Encyclopedia of Genes and Genomes (KEGG)[89], SwissProt and TrEMBL using BLASTP with an e-value cutoff of 1E-5, and the best alignment hits were used to assign homology-based gene functions. GO categories and InterPro entries were obtained via InterProScan[90].

Gene family evolution

The longest transcript of each of the protein-coding genes from 18 representative species (*Taxus wallichiana*, *Amborella trichopoda*, *Arabidopsis thaliana*, *Ginkgo biloba*, *Gnetum montanum*, *Welwitschia mirabilis*, *Oryza sativa*, *Solanum lycopersicum*, *Physcomitrella patens*, *Pinus tabuliformis*, *Selaginella moellendorffii*, *Vitis vinifera*, *Sequoiadendron giganteum*, *Azolla filiculoides*, *Klebsormidium flaccidum*, *Chara braunii*, *Marchantia polymorpha* and *Penium margaritaceum*) and *T. grandis* were selected to construct gene families based on all-against-all BLASTP alignments using OrthoFinder[91]. Phylogenetic analyses were conducted using IQ-TREE[92] (v. 2.1.3). Based on MRCA analysis using CAFE[93] (v.4.2.1), we determined the expansion and contraction of gene families between extant species and their last common ancestors.

Analysis of WGD events in the *T. grandis* genome

All-against-all BLASTP search were performed with an e-value cutoff of 1E-5. The top five alignments were selected for each gene and used to detect syntenic gene pairs located in collinear blocks with MCScanX[94]. Paralogous gene pairs were determined by the best reciprocal BLASTP alignments. *K*s of each syntenic or paralogous gene pair was calculated using YN00 in the package PAML 4.8a[95] with default parameters. Phylogeny based inference of WGD was carried out following the methods described previously[30].

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Small RNA sequencing

Total RNA (3 μg) from leaves was isolated for small RNA library construction using NEB Next® Multiplex Small RNA Library Prep Set for Illumina® (NEB, USA) following manufacturer's recommendations. DNA fragments in the constructed library within the range of 140~160 bp were recovered and the library was assessed on an Agilent Bioanalyzer 2100 system and subsequently sequenced on an Illumina HiSeq 2500 platform. Raw reads of small RNA library were processed with Trimmomatic[68] (v0.36) to remove adapters and then aligned to the reference genome using Bowtie[96] with no mismatch allowed.

Whole genome bisulfite sequencing

About 100 ng high-quality genomic DNA spiked with 0.5 ng lambda DNA were sonicated with Covaris S220. The fragmented DNA (200-300 bp) were treated with bisulfite using EZ DNA Methylation-GoldTM Kit (Zymo Research), and the library was quality assessed and sequenced on the Illumina NovaSeq platform with the paired-end mode.

Raw reads were cleaned with the methods described above. To align the cleaned reads, both the reference genome and reads were transformed (C-to-T and G-to-A) and then aligned with Bismark[97] (version 0.16.3) with parameters "-X 700 –dovetail". Reads that produced a unique best alignment against both "Watson" and "Crick" strands of the genome were kept and the methylation state of all cytosine nucleotides were inferred. The sodium bisulfite conversion rate was estimated based on the read alignments to the lambda genome. Methylated sites were identified with a binomial test using the methylated counts (mC), total counts (mC+umC) and the conversion rate (r). Sites with FDR-corrected *P* value < 0.05 were considered as methylated sites. To calculate whole genome methylation level, we divided the genome into 10-kb bins, and the methylation level of each window was calculated as count(mC) / (count(mC) + count(umC)). Differentially methylated regions (DMRs) were identified using the DSS software[98] under the *P* value threshold of 0.05. DMRs were cataloged based on whether and how they overlapped with genes.

HGT identification

Potential HGTs were identified following a method described previously[99]. In brief, we created three customized databases, namely an out-group database comprising all protein sequences from archaea, bacteria and fungi, an in-group database including protein sequences from 10 published

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

gymnosperm species, and a mid-group database consisting of sequences from all published plants excluding gymnosperms. Protein sequences of *T. grandis* were blasted against the three customized databases separately with an e-value cutoff of 1E-5. For each query protein sequence, we preserved no more than 100 blast hits (one hit per species) for each database and calculated the average bit-score value (ABV) of the alignments. Query proteins with the ABV of out-group larger than that of mid-group were retained. We performed rigorous phylogenetic analyses for each of remaining query proteins and manually inspected the topology of the tree. *T. grandis* genes supported by both ABV and phylogeny were considered as potential horizontally transferred genes.

Fatty acid determination

About 0.5 g dried samples were mixed with 9 mL 10% H_2SO_4-CH_3OH solution at room temperature for 10 h. The fatty acid methyl esters were filtered and then extracted with 30 mL distilled water and 30 mL dichloromethane. The organic phase was dried with anhydrous sodium sulfate and concentrated to about 1 ml with a nitrogen blower. The concentrated extract was used for analysis by gas chromatography (Thermo Scientific TRACE-1300, Italy), using methyl fatty acid as an internal standard.

Subcellular Localization

The CDS of each gene, without stop codon, was cloned and fused to the N-terminus of the GFP gene of the pCAMBIA1300-GFP vector. The resultant plasmid was introduced into *Agrobacterium tumefaciens* GV3101. Positive clones were incubated to an OD_{600} of 0.6, and then centrifuged at 8000 rpm for 6 min. Collected cells were resuspended with infiltration buffer (10 mM $MgCl_2$, 0.2 mM acetosyringone, and 10 mM MES at pH 5.6), which was then injected into the leaves of *Nicotiana benthamiana*. After 3-d of culture, GFP fluorescence signal from the leaves was observed and captured using confocal laser scanning microscopy (LSM510: Karl Zeiss).

Quantitative real-time PCR analysis

Total RNA was extracted using the RNAprep pure Plant Kit (TIANGEN). First-strand cDNA was synthesized from 1 μg of total RNA using the PrimeScript^{TM} RT Master Mix Kit (Takara). SYBR Premix Ex Taq™ kit (Takara) was used to perform quantitative real-time PCR. Expression data of target genes were corrected with the expression of actin encoding gene. The reaction conditions

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

were 95 °C for 10 s, 55 °C for 10 s, 72 °C for 20 s, 45 cycles. The relative expression was calculated using $2^{-\Delta\Delta Ct}$ method.

Heterologous synthesis of SCA in Arabidopsis seeds and *N. benthamiana* leaves

Coding regions of *TgEOL1*, *TgDES1*, *AtSLD2* (*AT2G46210*) and two recombinant genes (*AtSLD2-Motif2* and *AtSLD2-Motif3*) were inserted into the downstream of the 35S promoter of the binary vector (pCAMBIA1300), respectively. Each of the resulting constructs was transformed into *Agrobactrium tumefaciens* strain GV3101, which was then grown at 28°C in LB medium supplemented with kanamycin (50 mg/L) and rifampicillin (50 mg/L) until OD_{600} reaching 0.6. For transient expression of *AtSLD2* and two recombinant genes in *N. benthamiana* leaves, cells were harvested and resuspended in 10 mM MES buffer (containing 10 mM $MgCl_2$ and 0.1 mM acetosyringone) to a final OD600 of 1.0. The cells of each strain were infiltrated into the young leaves of five-week-old *N. benthamiana* plants using a needleless syringe, which were harvested 5 days later for measurement of SCA content. For generation of *TgDES1* and *TgELO1* overexpressed Arabidopsis, the pCAMBIA1300-*TgELO1* and pCAMBIA1300-*TgELO1* constructs were transformed into Arabidopsis via *A. tumefaciens*-mediated floral dip method. Hygromycin-resistant T1 plants were planted for seed harvesting, and T2 seeds with a hygromycin resistance ratio of 3:1 were selected to collect T3 seeds. T3 seeds with 100% resistance to hygromycin were used for determination of SCA content.

Author contribution

J.W., X.S., H.L. and L.S. conceived and supervised the project. W.C., Y.G. and S.Z. collected samples and performed transgenic experiments. X.S. and X.L. performed bioinformatics analysis. X.S. and H.L wrote the manuscript. Z.F. and J.W. revised the manuscript.

Acknowledgements

This research was supported by grants from the National Natural Science Foundation of China (NSFC) to J.W. (grant no. U20A2049), L.S. (grant no. 31971699), X.S. (grant no. 32102318), the Key Research and Development Program of Zhejiang Province to H.L. (grant no. 2021C02001), the Scientific Research Startup Fund Project of Zhejiang A&F University to H.L. (grant no. 2018FR028), and the State Key Laboratory of Subtropical Silviculture grant to J.W. (grant no.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

ZY20180312 and ZY20180209). The authors thank Dr. Emily E.D. Coffey from Atlanta Botanical Garden (USA) and Prof. Mark W. Schwartz from University of California, Davis for providing plant samples.

Competing interests

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Reference

1. Li, H.T. et al. Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants* **5**, 461–470 (2019).
2. Chen, X. & Jin, H. Review of cultivation and development of Chinese torreya in China. *For. Trees Livelihoods* **28**, 68–78 (2019).
3. Wang, H., Guo, T. & Ying, G.Q. Advances in studies on active constituents and their pharmacological activities for plants of Torreya Arn. *Chin. Tradit. Herb. Drugs* **38**, 1748–1750 (2007).
4. Dai, R. *et al.* The determination of the total flavonoids by UV and a flavone glycoside by HPLC in *Torreya grandis* Fort Leaves. *IEEE/ICME Int. Conf. Complex Med. Eng.* 1887–1891 (2007).
5. Ding, M. *et al.* Comparative transcriptome analysis of the genes involved in lipid biosynthesis pathway and regulation of oil body formation in *Torreya grandis* kernels. *Ind. Crops Prod.* **145**, 112051 (2020).
6. He, Z. *et al.* Chemical components of cold pressed kernel oils from different *Torreya grandis* cultivars. *Food Chem.* **209**, 196–202 (2016).
7. Hu, Y. *et al.* The interaction of temperature and relative humidity affects the main aromatic components in postharvest *Torreya grandis* nuts. *Food Chem.* **368**, 130836 (2022).
8. Lou, H. *et al.* Full-length transcriptome analysis of the genes involved in tocopherol biosynthesis in *Torreya grandis*. *J. Agric. Food Chem.* **67**, 1877–1888 (2019).
9. Suo, J. *et al.* Comparative transcriptome analysis reveals key genes in the regulation of squalene and β-sitosterol biosynthesis in *Torreya grandis*. *Ind. Crops Prod.* **131**, 182–193 (2019).
10. Wu, J. *et al. De novo* transcriptome sequencing of *Torreya grandis* reveals gene regulation in sciadonic acid biosynthesis pathway. *Ind. Crops Prod.* **120**, 47–60 (2018).
11. Berger, A. *et al.* Epidermal anti-inflammatory properties of 5,11,14 20:3: effects on mouse ear edema, PGE2 levels in cultured keratinocytes, and PPAR activation. *Lipids Health Dis.* **1**, 5 (2002).
12. Pédrono, F. *et al.* Sciadonic acid derived from pine nuts as a food component to reduce plasma triglycerides by inhibiting the rat hepatic Δ9-desaturase. *Sci. Rep.* **10**, 6223 (2020).
13. Endo, Y., Osada, Y., Kimura, F. & Fujimoto, K. Effects of Japanese torreya (*Torreya nucifera*) seed oil on lipid metabolism in rats. *Nutrition* **22**, 553–558 (2006).
14. Tanaka, T., Morishige, J.-i., Takimoto, T., Takai, Y. & Satouchi, K. Metabolic characterization of sciadonic acid (5c,11c,14c-eicosatrienoic acid) as an effective substitute for arachidonate of phosphatidylinositol. *Eur. J Biochem.* **268**, 4928–4939 (2001).
15. Song, L. *et al.* Advances on delta 5-unsaturated-polymethylene-interrupted fatty acids: resources, biosynthesis, and benefits. *Crit. Rev. Food Sci. Nutr.* 1–23 (2021).
16. Aitzetmüller, K. An unusual fatty acid pattern in *Eranthis* seed oil. *Lipids* **31**, 201–205 (1996).
17. Sun, Y., Shang, L., Zhu, Q.H., Fan, L. & Guo, L. Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci.* **27**, 391–401 (2022).
18. Rhie, A., Walenz, B.P., Koren, S. & Phillippy, A.M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

19. Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A. & Zdobnov, E.M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).

20. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).

21. Steinthorsdottir, M. *et al.* The Miocene: the future of the past. *Paleoceanogr. Paleoclimatol.* **36**, e2020PA004037 (2021).

22. Matzke, M.A. & Mosher, R.A. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15**, 394–408 (2014).

23. Niu, S. *et al.* The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* **185**, 204–217.e14 (2022).

24. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).

25. Leebens-Mack, J.H. *et al.* One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).

26. Liu, Y. *et al.* The *Cycas* genome and the early evolution of seed plants. *Nat. Plants* **8**, 389–401 (2022).

27. Xiong, X. *et al.* The *Taxus* genome provides insights into paclitaxel biosynthesis. *Nat. Plants* **7**, 1026–1036 (2021).

28. Cheng, J. *et al.* Chromosome-level genome of Himalayan yew provides insights into the origin and evolution of the paclitaxel biosynthetic pathway. *Mol. Plant* **14**, 1199–1209 (2021).

29. Wan, T. *et al.* The *Welwitschia* genome reveals a unique biology underpinning extreme longevity in deserts. *Nat. Commun.* **12**, 4247 (2021).

30. Sun, X. *et al.* Genome and evolution of the arbuscular mycorrhizal fungus *Diversispora epigaea* (formerly *Glomus versiforme*) and its bacterial endosymbionts. *New Phytol.* **221**, 1556–1573 (2019).

31. Delaux, P.-M. *et al.* Algal ancestor of land plants was preadapted for symbiosis. *Proc. Natl. Acad. Sci. USA* **112**, 13390–13395 (2015).

32. Jiao, C. *et al.* The *Penium margaritaceum* genome: hallmarks of the origins of land plants. *Cell* **181**, 1097–1111.e12 (2020).

33. Bowman, J.L. *et al.* Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell* **171**, 287–304.e15 (2017).

34. Bouché, F., Lobet, G., Tocquin, P. & Périlleux, C. FLOR-ID: an interactive database of flowering-time gene networks in *Arabidopsis thaliana*. *Nucleic Acids Res.* **44**, D1167–D1171 (2016).

35. Li, L.X. *et al. Arabidopsis thaliana* NOP10 is required for gametophyte formation. *J. Integr. Plant Biol.* **60**, 723–736 (2018).

36. Schoof, H. *et al.* The stem cell population of *Arabidopsis* shoot meristems in maintained by a regulatory loop between the *CLAVATA* and *WUSCHEL* genes. *Cell* **100**, 635–644 (2000).

37. Soltis, D.E., Chanderbali, A.S., Kim, S., Buzgo, M. & Soltis, P.S. The ABC model and its applicability to basal angiosperms. *Ann. Bot.* **100**, 155–163 (2007).

38. Dreni, L. & Zhang, D. Flower development: the evolutionary history and functions of the AGL6 subfamily MADS-box genes. *J. Exp. Bot.* **67**, 1625–1638 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

39. Kong, X. *et al.* The wheat AGL6-like MADS-box gene is a master regulator for floral organ identity and a target for spikelet meristem development manipulation. *Plant Biotechnol. J.* **20**, 75–88 (2022).
40. Zhang, L. *et al.* The water lily genome and the early evolution of flowering plants. *Nature* **577**, 79–84 (2020).
41. Chanderbali, A.S. *et al.* Conservation and canalization of gene expression during angiosperm diversification accompany the origin and evolution of the flower. *Proc. Natl. Acad. Sci. USA* **107**, 22570–22575 (2010).
42. Souza, P.F.N. The forgotten 2S albumin proteins: importance, structure, and biotechnological application in agriculture and human health. *Int. J. Biol. Macromol.* **164**, 4638–4649 (2020).
43. Tandang-Silvas, M.R. *et al.* Conservation and divergence on plant seed 11S globulins based on crystal structures. *Biochim. Biophys. Acta.* **1804**, 1432–1442 (2010).
44. Cheng, S. *et al.* Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. *Cell* **179**, 1057–1067.e14 (2019).
45. Weng, J.K. & Chapple, C. The origin and evolution of lignin biosynthesis. *New Phytol.* **187**, 273–285 (2010).
46. De La Torre, A.R. *et al.* Functional and morphological evolution in gymnosperms: a portrait of implicated gene families. *Evol. Appl.* **13**, 210–227 (2020).
47. Yamaguchi, M. *et al.* VASCULAR-RELATED NAC-DOMAIN6 and VASCULAR-RELATED NAC-DOMAIN7 effectively induce transdifferentiation into xylem vessel elements under control of an induction system. *Plant Physiol.* **153**, 906–914 (2010).
48. Zhong, R., Richardson, E.A. & Ye, Z.H. Two NAC domain transcription factors, SND1 and NST1, function redundantly in regulation of secondary wall synthesis in fibers of Arabidopsis. *Planta* **225**, 1603–1611 (2007).
49. Meesapyodsuk, D. & Qiu, X. The front-end desaturase: structure, function, evolution and biotechnological use. *Lipids* **47**, 227–237 (2012).
50. Sayanova, O., Haslam, R., Venegas Caleron, M. & Napier, J.A. Cloning and characterization of unusual fatty acid desaturases from *Anemone leveillei*: identification of an acyl-coenzyme A C_{20} Δ^5-desaturase responsible for the synthesis of sciadonic acid. *Plant Physiol.* **144**, 455–467 (2007).
51. Xue, J.A. *et al.* Expression of yeast acyl-CoA-Δ^9 desaturase leads to accumulation of unusual monounsaturated fatty acids in soybean seeds. *Biotechnol. Lett.* **35**, 951–959 (2013).
52. Lim, Z.L., Senger, T. & Vrinten, P. Four amino acid residues influence the substrate chain-length and regioselectivity of *Siganus canaliculatus* Δ4 and Δ5/6 desaturases. *Lipids* **49**, 357–367 (2014).
53. Buček, A., Vazdar, M., Tupec, M., Svatoš, A. & Pichová, I. Desaturase specificity is controlled by the physicochemical properties of a single amino acid residue in the substrate binding tunnel. *Comput. Struct. Biotechnol. J.* **18**, 1202–1209 (2020).
54. Linkies, A., Graeber, K., Knight, C. & Leubner-Metzger, G. The evolution of seeds. *New Phytol.* **186**, 817–831 (2010).
55. Chen, M. *et al.* Seed genome hypomethylated regions are enriched in transcription factor genes. *Proc. Natl. Acad. Sci. USA* **115**, E8315–E8322 (2018).
56. Niederhuth, C.E. *et al.* Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* **17**, 194 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

57. Ausin, I. *et al.* DNA methylome of the 20-gigabase Norway spruce genome. *Proc. Natl. Acad. Sci. USA* **113**, E8106–E8113 (2016).
58. Bewick, A.J. & Schmitz, R.J. Gene body DNA methylation in plants. *Curr. Opin. Plant Biol.* **36**, 103–110 (2017).
59. Lin, J.Y. *et al.* Similarity between soybean and Arabidopsis seed methylomes and loss of non-CG methylation does not affect seed development. *Proc. Natl. Acad. Sci. USA* **114**, E9730–E9739 (2017).
60. Chen, F., Tobimatsu, Y., Havkin-Frenkel, D., Dixon, R.A. & Ralph, J. A polymer of caffeyl alcohol in plant seeds. *Proc. Natl. Acad. Sci. USA* **109**, 1772–1777 (2012).
61. Hiraide, H. *et al.* Localised laccase activity modulates distribution of lignin polymers in gymnosperm compression wood. *New Phytol.* **230**, 2186–2199 (2021).
62. Cosgrove, D.J. Loosening of plant cell walls by expansins. *Nature* **407**, 321–326 (2000).
63. Vigeolas, H., van Dongen, J.T., Waldeck, P., Huhn, D. & Geigenberger, P. Lipid storage metabolism is limited by the prevailing low oxygen concentrations within developing seeds of oilseed rape. *Plant Physiol.* **133**, 2048–2060 (2003).
64. Zhu, W. *et al.* Altered chromatin compaction and histone methylation drive non-additive gene expression in an interspecific Arabidopsis hybrid. *Genome Biol.* **18**, 157 (2017).
65. Wang, C. *et al.* Genome-wide analysis of local chromatin packing in *Arabidopsis thaliana*. *Genome Res.* **25**, 246–256 (2015).
66. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
67. Roach, M.J., Schmidt, S. & Borneman, A.R. Purge Haplotigs: synteny reduction for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
68. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
69. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
70. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
71. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
72. Flynn, J.M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
73. Bao, W., Kojima, K.K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
74. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **25**, Chapter 4:Unit 4.10 (2009).
75. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
76. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
77. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
78. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.28.514327; this version posted October 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

80. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).

81. Alioto, T. Blanco, E. Parra, G. & Guigó, R. Using geneid to identify genes. *Curr. Protoc. Bioinformatics* **64,** e56 (2018).

82. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).

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

84. She, R., Chu, J.S., Wang, K., Pei, J. & Chen, N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res.* **19**, 143–149 (2009).

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

86. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).

87. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**, 2325–2329 (2011).

88. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).

89. Morishima, K., Tanabe, M., Furumichi, M., Kanehisa, M. & Sato, Y. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, 353–361 (2016).

90. Mitchell, A. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

91. Emms, D.M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

92. Nguyen, L.T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

93. 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).

94. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, 49 (2012).

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

96. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

97. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).

98. Park, Y. & Wu, H. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics* **32**, 1446–1453 (2016).

99. Li, Y. *et al.* HGT is widespread in insects and contributes to male courtship in lepidopterans. *Cell* **185**, 2975–2987.e10 (2022).