nature plants

ARTICLES

https://doi.org/10.1038/s41477-022-01129-7

Check for updates

OPEN

The *Cycas* genome and the early evolution of seed plants

Yang Liu[1,2,34]✉, Sibo Wang[1,34], Linzhou Li[1,34], Ting Yang[1,34], Shanshan Dong[2,34], Tong Wei[1,34], Shengdan Wu[3,34], Yongbo Liu[4,34], Yiqing Gong[2], Xiuyan Feng[5], Jianchao Ma[6], Guanxiao Chang[6], Jinling Huang[5,6,7], Yong Yang[8], Hongli Wang[1,9], Min Liu[1], Yan Xu[1,9], Hongping Liang[1,9], Jin Yu[1,9], Yuqing Cai[1,9], Zhaowu Zhang[1,9], Yannan Fan[1], Weixue Mu[1], Sunil Kumar Sahu[1], Shuchun Liu[2], Xiaoan Lang[2,10], Leilei Yang[2], Na Li[2], Sadaf Habib[2,11], Yongqiong Yang[12], Anders J. Lindstrom[13], Pei Liang[14], Bernard Goffinet[15], Sumaira Zaman[15], Jill L. Wegrzyn[15], Dexiang Li[10], Jian Liu[5], Jie Cui[16], Eva C. Sonnenschein[17], Xiaobo Wang[18], Jue Ruan[18], Jia-Yu Xue[19], Zhu-Qing Shao[20], Chi Song[21], Guangyi Fan[1], Zhen Li[22], Liangsheng Zhang[23,24], Jianquan Liu[25], Zhong-Jian Liu[26], Yuannian Jiao[27], Xiao-Quan Wang[27], Hong Wu[28], Ertao Wang[29], Michael Lisby[30], Huanming Yang[1], Jian Wang[1], Xin Liu[1], Xun Xu[1], Nan Li[2], Pamela S. Soltis[31], Yves Van de Peer[19,22,32]✉, Douglas E. Soltis[31,33]✉, Xun Gong[5]✉, Huan Liu[1]✉ and Shouzhou Zhang[2]✉

Cycads represent one of the most ancient lineages of living seed plants. Identifying genomic features uniquely shared by cycads and other extant seed plants, but not non-seed-producing plants, may shed light on the origin of key innovations, as well as the early diversification of seed plants. Here, we report the 10.5-Gb reference genome of *Cycas panzhihuaensis*, complemented by the transcriptomes of 339 cycad species. Nuclear and plastid phylogenomic analyses strongly suggest that cycads and *Ginkgo* form a clade sister to all other living gymnosperms, in contrast to mitochondrial data, which place cycads alone in this position. We found evidence for an ancient whole-genome duplication in the common ancestor of extant gymnosperms. The *Cycas* genome contains four homologues of the *fitD* gene family that were likely acquired via horizontal gene transfer from fungi, and these genes confer herbivore resistance in cycads. The male-specific region of the Y chromosome of *C. panzhihuaensis* contains a MADS-box transcription factor expressed exclusively in male cones that is similar to a system reported in *Ginkgo*, suggesting that a sex determination mechanism controlled by MADS-box genes may have originated in the common ancestor of cycads and *Ginkgo*. The *C. panzhihuaensis* genome provides an important new resource of broad utility for biologists.

Cycads are often referred to as 'living fossils'; they originated in the mid-Permian and dominated terrestrial ecosystems during the Mesozoic, a period called the 'age of cycads and dinosaurs'[1]. Although the major cycad lineages are ancient, modern cycad species emerged from several relatively recent diversifications[2,3]. Cycads are long-lived woody plants that, unlike other extant gymnosperms, bear frond-like leaves clustered at the tip of the stem[4]. Extant cycads comprise 10 genera and approximately 360 species, two-thirds of which are on the International Union for Conservation of Nature Red List of threatened species[5]. All living cycad species are dioecious, with individual plants developing either male or female cones (except in *Cycas*, which produces a loose cluster of megasporophylls rather than a true female cone; Fig. 1a)[6]. Unlike other extant seed plants, cycads and *Ginkgo* retain flagellated sperm, an ancestral trait shared with bryophytes, lycophytes and ferns[7]. Cycads exhibit other special features, such as the accumulation of toxins that deter herbivores[8] in seeds and vegetative tissues. They also produce coralloid roots that host symbiotic cyanobacteria, making them the only gymnosperm associated with nitrogen-fixing symbionts[9]. The origin of the seed marked one of the most important events of plant evolution[10]. As one of the four extant gymnosperm groups (cycads, *Ginkgo*, conifers and gnetophytes), cycads hold an important evolutionary position for understanding the origin and early evolution of seed plants. We therefore generated a high-quality genome assembly for a species of *Cycas* to explore fundamental questions in seed plant evolution, including the phylogenetic position of cycads, the occurrence of ancient whole-genome duplications (WGDs), innovation in gene function and the evolution of sex determination.

A chromosome-scale genome assembly

Here, we report a high-quality, chromosome-level genome assembly of *Cycas panzhihuaensis* based on sequencing of the haploid megagametophyte using a combination of MGI-SEQ short-read, Oxford Nanopore long-read and Hi-C sequencing methods (Supplementary Note 2). The genome comprises 10.5 Gb assembled in 5,123 contigs (N50 = 12 Mb), with 95.3% of these contigs anchored to the largest 11 pseudomolecules, corresponding to the 11 chromosomes ($n = 11$) of the *C. panzhihuaensis* karyotype[11] (Supplementary Note 3 and Extended Data Fig. 1). The annotated genome describes

A full list of affiliations appears at the end of the paper.

32,353 protein-coding genes and is mostly composed of repetitive elements adding up to 7.8 Gb (Supplementary Note 4). Based on BUSCO[12] estimation, the gene space completeness of the *C. panzhihuaensis* genome assembly is 91.6% (Supplementary Note 4).

Compared with other gymnosperms, the size of the *Cycas* genome is similar to that of *Ginkgo* (10.6 Gb)[13,14] and intermediate between the relatively compact genome of *Gnetum* (4.1 Gb)[15] and the very large genomes of conifers (for example, ~20-Gb genomes of *Picea* and *Pinus*)[16–18]. As in other gymnosperm genomes, a large portion (76.14%) of the *C. panzhihuaensis* genome consists of ancient repetitive elements (Supplementary Note 4). In addition, the genome contains almost equal proportions of *copia* and *gypsy* long terminal repeat (LTR) elements, in contrast to other gymnosperm genomes, in which *gypsy* repeats are more frequent[14,15] (Supplementary Note 6). Among all sequenced plant genomes, *C. panzhihuaensis* has the longest average introns (~30.8 kb) and genes (~121.3 kb) (Extended Data Fig. 2a), surpassing those of *Ginkgo*[14]. In comparison with *Ginkgo*, in which LTRs dominate intron content, the introns of *C. panzhihuaensis* contain a large portion of unknown sequences (Extended Data Fig. 2b). The longest gene, *CYCAS_013063*, encoding a kinesin-like protein KIF3A, covers 2.1 Mb in the *C. panzhihuaensis* genome; the longest intron is approximately 1.5 Mb and was detected in *CYCAS_030563*, a gene that encodes a photosystem II CP43 reaction centre protein. Both genes are expressed, as evidenced by our long-read transcriptome data.

Phylogeny of cycads and seed plants

The *C. panzhihuaensis* genome provides an opportunity to revisit the long-standing debate on the evolutionary relationships among living seed plants. On the basis of molecular phylogenetic analyses, extant gymnosperms are resolved as a monophyletic group, but the branching order among their major lineages has remained controversial[19–23]. Our phylogenetic analyses of separate nuclear (Fig. 1b, Extended Data Fig. 3 and Supplementary Note 5) and plastid datasets strongly support cycads plus *Ginkgo* as sister to the remaining extant gymnosperms, in agreement with several other analyses[23,24], whereas mitochondrial data resolve cycads alone in that position (Fig. 1c). This conflict arising from the mitochondrial data cannot be explained by the presence of extensive RNA editing sites in the mitochondrial data (Fig. 1c), which in some cases has been reported to bias phylogenetic inferences[25,26], and instead may be best explained by incomplete lineage sorting, which is supported by our PhyloNet[27] and coalescent analyses of nuclear genes (Supplementary Note 5).

The extant diversity of cycads was previously considered to have arisen synchronously within the past 9–50 million years (Myr)[2,3]. Our inferences, based on 1,170 low-copy nuclear genes sampled for 339 cycad species and 6 fossil calibrations[3] corroborate recent broad analyses of gymnosperms indicating that extant species-rich cycad genera emerged from rapid radiations ranging from 11 to 20 Myr ago, which may have been a consequence of dramatic Miocene global temperature changes[24,28]. Notably, major temperate and tropical radiations in several major clades of flowering plants have been shown to be associated with Miocene cooling in the past 15 Myr (refs. [29–31]).

Cycas is an ancient polyploid

WGD is a major driving force in the evolution of land plants and has dramatically promoted the diversification of flowering plants[23,32]. Synonymous substitutions per synonymous site (K_S) analysis of duplicate genes[33] revealed a clear peak at similar K_S values (~0.85, range 0.5–1.2) for both *Cycas* and *Ginkgo*, suggestive of an ancient WGD possibly shared by these two lineages (Supplementary Note 7)[34]. However, the precise evolutionary position of this WGD event remains ambiguous. Our phylogenomic analyses based on 15 genomes and 1 transcriptome revealed 2,469 gymnosperm-wide duplications in 9,545 gene families and indicate that this WGD event dates to the most recent common ancestor (MRCA) of extant gymnosperms (Fig. 2a), supporting recent findings based on transcriptome data[24]. We also identified 69 ancient syntenic genomic segments that further support a gymnosperm-wide WGD (Extended Data Fig. 3, Supplementary Fig. 23 and Supplementary Tables 24 and 25). Furthermore, a mixed dataset with increased sampling—29 genomes and 61 transcriptomes—also yielded the same result (Fig. 2a and Extended Data Fig. 4). This gymnosperm-wide WGD, here named omega (ω), is independent of the WGD preceding the split between gymnosperms and angiosperms[35] and may have contributed to the subsequent evolution of gymnosperm-specific genes involved in plant hormone signal transduction, biosynthesis of secondary metabolites, plant–pathogen interaction and terpenoid biosynthesis (Supplementary Note 7).

Ancestral gene innovation in the origin of the seed

The origin of seed plants is marked by the emergence of key traits including the seed, pollen and secondary growth of xylem and phloem[36]. Reconstruction of the evolution of gene families across the seed plant tree of life revealed that 663 orthogroups were gained and 368 expanded in the MRCA of extant seed plants compared with non-seed plants (Fig. 2b, node 1). Among these, 106 of the new orthogroups and 55 of the expanded orthogroups are associated with seed development in *Arabidopsis*[37], including the regulation of development during early embryogenesis, seed dormancy and germination, and seed coat formation, as well as in immunity and stress response of the seed (Supplementary Note 6).

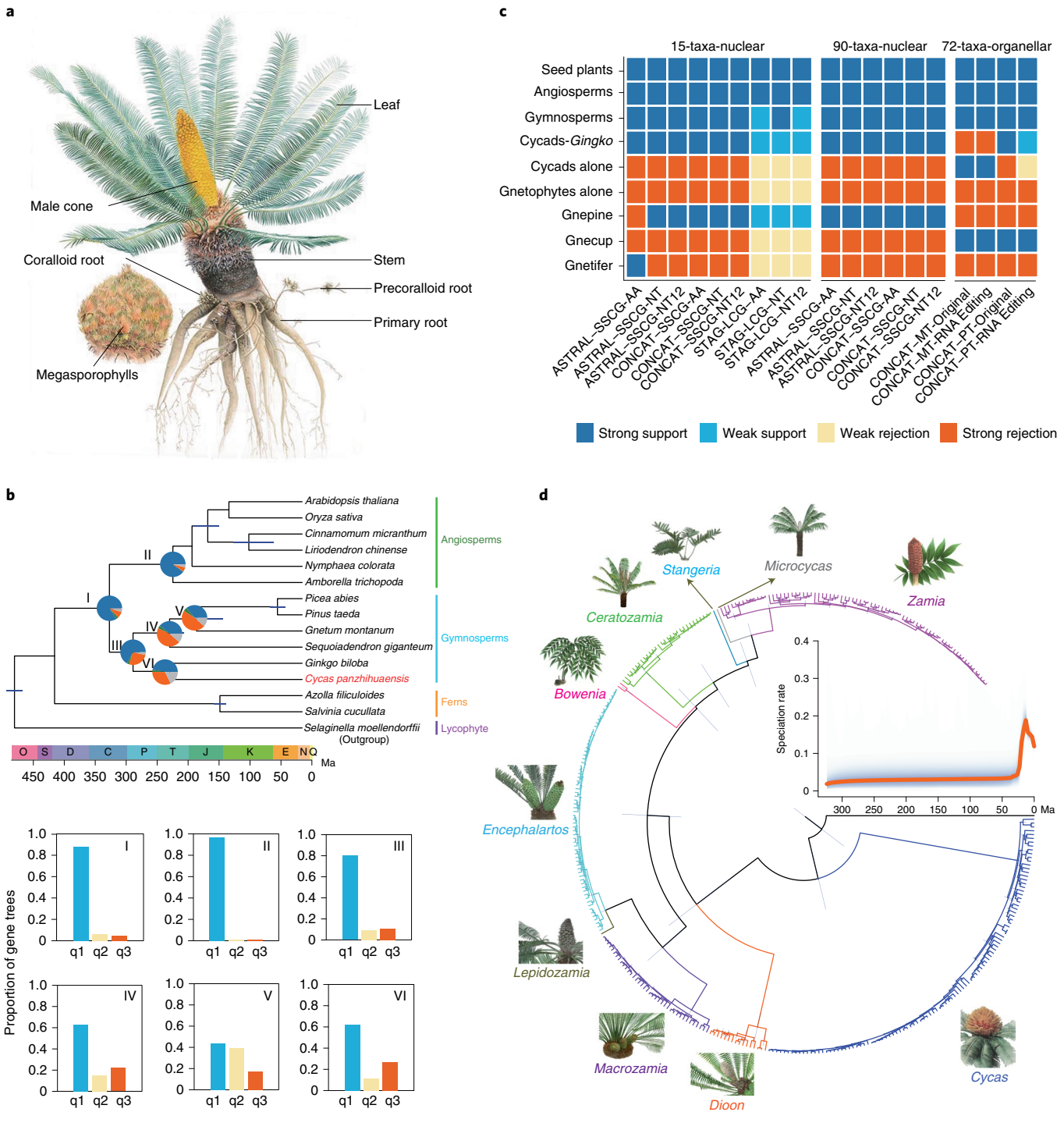
Genes of the LAFL family are well-known as core regulatory genes of seed development, including *LEAFY COTYLEDON1* (*LEC1*), *ABSCISIC ACID INSENSITIVE3* (*ABI3*), *LEAFY COTYLEDON2* (*LEC2*) and *FUSCA3* (*FUS3*), which encode master transcriptional regulators, interacting to form complexes that control embryo

Fig. 1 | Phylogenomic analyses of cycads and seed plants. a, Illustration of *Cycas panzhihuaensis*. **b**, Chronogram of seed plants on the basis of the SSCG-NT12 dataset inferred using MCMCTree. All branches are maximally supported by bootstrap values (ML) and posterior probabilities (ASTRAL). I, II, III, VI, V and VI indicate internal branches for which the pie charts depicting gene tree incongruence are complemented by histograms (lower panel) showing quartet support for the main topology (q1), the first alternative topology (q2) and the second alternative topology (q3). O, Ordovician; S, Silurian; D, Devonian; C, Carboniferous; P, Permian; T, Triassic; J, Jurassic; K, Cretaceous; Pg, Palaeogene; N, Neogene; Q, Quaternary; Ma, million years ago. **c**, DiscoVista species tree analysis: rows correspond to the nine hypothetical groups tested (see Supplementary Note 5 for details) and columns correspond to the results derived from the use of different datasets and methods. SSCG, single-copy genes; LCG, low-copy genes; MT, mitochondrial genes; PT, plastid genes; AA, amino acid sequences; NT, nucleotide sequences; NT12, codon 1st + 2nd positions; ASTRAL, coalescent tree inference method using ASTRAL; CONCAT, maximum likelihood tree inferred with IQ-TREE based on concatenated datasets; STAG, species tree inference using software STAG with low-copy genes (one to four copies); Original, original organellar nucleotide sequences; RNA Editing, organellar genes with RNA editing site modified. Strong support, the clade is reconstructed with a support value >95%. Weak support, the clade is reconstructed with support value <95%. Weak rejection, the clade is not recovered, but the alternative topology is not conflict if poorly supported branches (<85%) are collapsed. Strong rejection, the clade is not recovered, and the alternative topology is conflict even when poorly supported branches (<85%) are collapsed. **d**, Diversification of Cycadales. The chronogram of 339 cycad species was inferred with MCMCTree based on 100 nuclear single-copy genes with concordant evolutionary histories. All illustrations are specifically created for this study (a high-resolution version is available at https://db.cngb.org/codeplot/datasets/public_dataset?id=PwRftGHfPs5qG3gE).

development and maturation[38]. *LEC1* genes are found only in vascular plants, but *ABI3* is widely distributed in embryophytes (Supplementary Note 10.6). *Cycas* and *Ginkgo* each contain a small number of *LEC1* (two and three in each, respectively) and *ABI3* (one in each) genes, whereas *C. panzhihuaensis* encodes a burst of *FUS3* (ten) and *LEC2* (seven) genes in the form of tandem repeats. *FUS3* and *LEC2* are shared by all living seed plants; the *Cycas* and other gymnosperm genomes contain genes composing a new clade of B3 domain proteins, that is, the *FUS3*/*LEC2*-like clade, which is sister to the clade of *FUS3* and *LEC2* (Extended Data Fig. 5). The *FUS3*/*LEC2*-like families are unique to gymnosperms, show significant expression after pollination in *C. panzhihuaensis* (Extended Data Fig. 5c) and may play specific roles in initiating embryogenesis in gymnosperms.

Regulation of seed development in *Cycas*

To better understand the dynamic changes in gene regulation and regulatory programmes during ovule pollination and fertilization, we performed a weighted correlation network analysis (WGCNA) and identified 11 co-expression modules at different developmental stages of the *C. panzhihuaensis* ovule and seed (Fig. 3a). The modules are enriched in seed nutrition metabolic processes (M2, M6 and M8), membrane biosynthesis (M9, which may relate to the development of the integument) and genes synthesizing callose, a major component of the pollen tube (M4) (Supplementary Note 10). A survey of phytohormones showed that salicylic acid and jasmonic acid, which are both involved in pathogen resistance, were produced at higher levels in unpollinated ovules versus post-pollinated ovules (Fig. 3b), and genes involved in the biosynthesis of these two



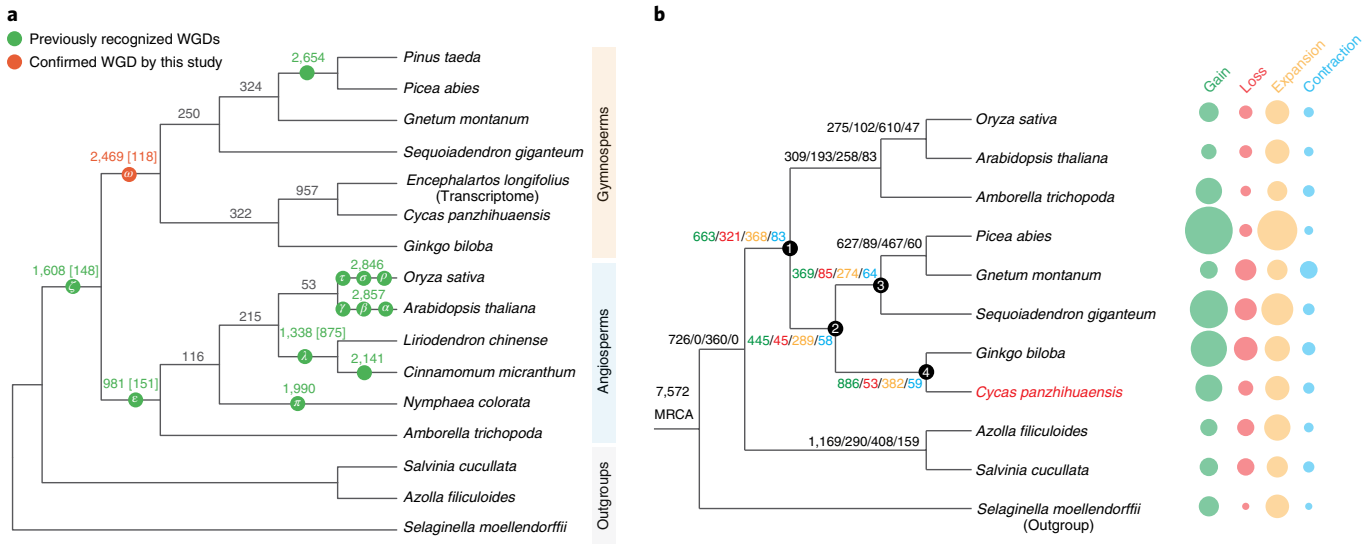


Fig. 2 | Ancient polyploidy events and evolution of gene families in seed plants. a, Inference of the number of gene families with duplicated genes surviving after WGD events mapped on a phylogenetic tree depicting the relationships among 16 vascular plants included in this study. The number of gene families with retained gene duplicates reconciled on a particular branch of the species tree are shown above the branch across the phylogeny (Methods). Numbers in square brackets denote the number of gene families with duplicated genes also supported by synteny evidence. **b**, Evolutionary analyses and phylogenetic profiles depicting the gains (light green), losses (light red), expansions (light yellow) and contractions (light blue) of orthogroups, according to the reconstruction of the ancestral gene content at key nodes and the dynamic changes of the lineage-specific gene characteristics.

phytohormones were also more highly expressed in unpollinated ovules, indicating the higher demand for these hormones as agents of pathogen resistance in the unpollinated ovule. Gibberellin, which is reported to regulate integument development in the ovules of flowering plants[39], accumulated in the late stage of the pollinated ovule in *Cycas*. We also found gene families related to integument development (for example, those involved in cutin, suberine and wax biosynthesis), with increased expression levels at the late stage of the pollinated ovule. Fertilized ovules accumulated a high level of abscisic acid and expressed the genes related to cell wall organization and biogenesis, indicating their activity in embryo development, seed coat formation, and seed maturation and dormancy[40] (Supplementary Note 10.1–10.5).

Among genes related to seed development, the most notable is the cupin protein family, expanded in *C. panzhihuaensis* compared with all other green plants. Phylogenetic analysis revealed that the cupin family can be subdivided into two groups: the germin-like and seed storage protein (SSP)-encoding genes. Surprisingly, we identified a new type of gene encoding vicilin-like storage proteins in *C. panzhihuaensis*; this type appears to be homologous to the vicilin-like antimicrobial peptides (v-AMP) and is organized as a tandem gene array in the *C. panzhihuaensis* genome (Fig. 3c). These v-AMP homologues are mostly expressed in *C. panzhihuaensis* at the late stage of pollinated ovules and fertilized ovules, with expression gradually decreasing during embryogenesis, suggesting the potentially important role of v-AMP genes in seed development (Fig. 3d and Supplementary Note 10.6).

Secondary growth and cell wall synthesis

Secondary growth is also a major innovation of seed plants[36], and it has been recognized from fossils of now-extinct progymnosperms, which predated the origin of seed plants[36,41]. Secondary phloem and xylem are produced by the activity of a bifacial vascular cambium (secondary meristem). We found that several genes that are known in angiosperms to regulate secondary growth in the positioning of the xylem, or in xylem/phloem patterning, underwent obvious expansions in the MRCA of extant seed plants compared with non-seed plants, including the MYB family member *ALTERED PHLOEM DEVELOPMENT* (*APL*), *WOL* and *BRASSINOSTEROID-INSENSITIVE LIKE 1* (*BRL1*) and *BRL3*. The *APL* gene is expressed in the phloem and cambium in vascular plants, and its encoded protein promotes phloem differentiation[42]. The expression of *APL* is regulated by *WOL* in the procambium[43]. The *BRL1* and *BRL3* genes encode brassinosteroid receptors that play major roles in xylem differentiation and phloem/xylem patterning in angiosperms[44]. Many copies of these genes were found to be highly expressed in cambium or apical meristem of *C. panzhihuaensis* (Supplementary Note 6).

Many gymnosperms are tall, woody plants with cell walls containing large quantities of cellulose, xyloglucan, glucomannan, homogalacturonans and rhamnogalacturonans[45]. In the *cellulose synthase* (*CESA*/*CSL*) superfamily[46], we discovered the existence of putative ancestral *cellulose synthase-like B*/*H* (*CSLB*/*H*) and *CSLE*/*G* that are specifically shared by gymnosperms, and both gene groups originated before the divergence of *CSLB* and *CSLH* in angiosperms (Extended Data Fig. 6). Cycads have manoxylic wood, with a large pith, large amounts of parenchyma and relatively few tracheids, in contrast to most other gymnosperms, which have pycnoxylic wood, with small amounts of pith, cortex and parenchyma, and a greater density of tracheids[4]. The *glutamyltransferase 77* (*GT77*) family, involved in the synthesis of rhamnogalacturonan II, which is essential for cell wall synthesis in rapidly growing tissues[47], is expanded in *C. panzhihuaensis* compared with other gymnosperms (Supplementary Note 11). In addition, gene families related to cell wall extension and loosening are uniquely expanded in *C. panzhihuaensis*, including those encoding hydroxyproline-rich glycoproteins, which are seven times more abundant in *Cycas* than in *Ginkgo*, and the fasciclin-like arabinogalactan proteins, which are twice as numerous in *Cycas* as in *Ginkgo*, *Sequoiadendron giganteum* and *Pseudotsuga menziesii*. How all these gene families related to wood features are regulated in cycads relative to other gymnosperms will be important for understanding the differences in wood density.

The evolution of pollen, pollen tube and sperm

Another major innovation during seed plant evolution is the production of pollen and the pollen tube[36]. We found that many genes

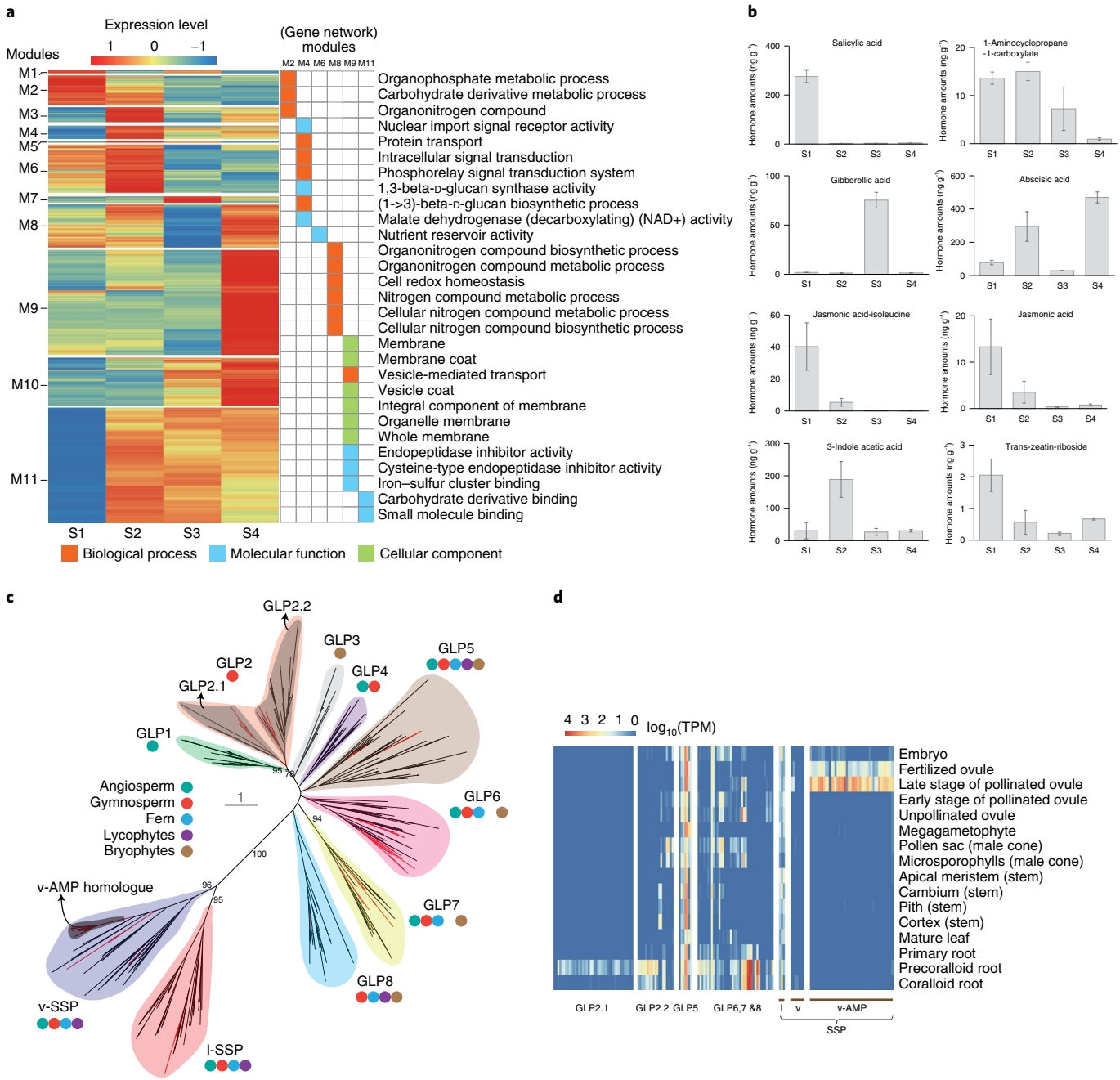


Fig. 3 | Gene expression and phytohormone synthesis at different developmental stages of the seed of *Cycas* and the evolution of seed storage proteins. **a**, Heatmap showing relative expression of genes in 11 co-expression modules by WGCNA across 4 developmental stages of the seed: S1, unpollinated ovule; S2, early stage of pollinated ovule; S3, late stage of pollinated ovule; and S4, fertilized ovule. **b**, Quantification of eight plant phytohormone amounts in the same four developmental stages of the *Cycas* seed as above. The grey histogram represents the amount of hormone ($n = 2$ biologically independent experiments) and the error bar represents the standard error. **c**, Phylogeny of SSPs in some representative species in land plants. The SSPs analysed include germin-like protein (GLP), legumin-like SSP (l-SSP), vicilin-like SSP (v-SSP) and v-AMP. A maximum likelihood tree with 500 bootstrap replicates was constructed using RAxML. Bootstrap values (≥50%) for each major clade (highlighted in colour) and the relationships among them are provided. The *Cycas* sequences are highlighted in red. **d**, Expression levels of SSP in different tissues of *C. panzhihuaensis*.

regulating pollen and pollen tube development (pollen maturation, pollen tube growth, pollen tube perception and prevention of multiple-pollen tube attraction) were gained (or the respective gene family expanded) in the MRCA of extant seed plants (Fig. 2b), as might be predicted for these features. For instance, those genes encoding egg cell-secreted proteins that prevent attraction of multiple pollen tubes[48] originated in the MRCA of living seed plants. The *Ole e 1*-like gene families, which encode proteins that accumulate in the pollen tube cell wall and play a role in pollen germination and pollen tube growth[49], are remarkably expanded in the MRCA of extant seed plants compared with non-seed plants (Supplementary Note 6). Such expansion also includes *polcalcin*, which is involved in calcium signalling to guide pollen tube growth[50] (Supplementary Note 11). Both the *COBRA* and *COBRA*-like protein gene families are expanded in *Cycas* and other seed plants compared with non-seed plants, and the

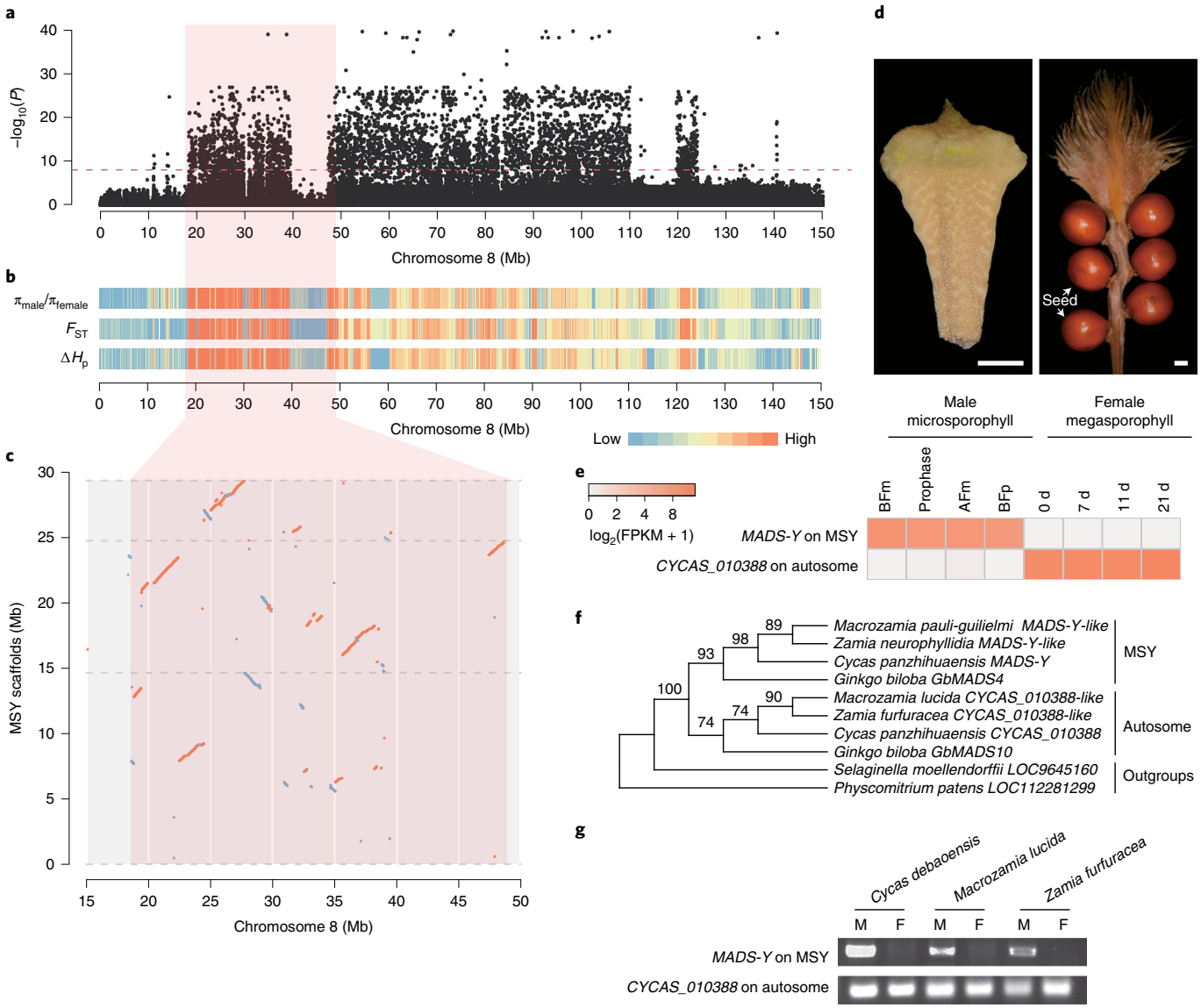


Fig. 4 | Identification of male-specific chromosomal region in *Cycas*. **a**, Manhattan plot of GWAS analysis of sex differentiation in 31 male and 31 female *Cycas* samples. The red horizontal dashed line represents the Bonferroni-corrected threshold for genome-wide significance ($\alpha = 0.05$). *P* values were calculated from a mixed linear model association of SNPs. Association analyses were performed once with a population of 31 male and 31 female individuals. **b**, Ratio of π, F_{ST} and difference of pooled heterozygosity (ΔH_p) within a 100-kb sliding window between the female and male sequences. Colour represents values from low (blue) to high (red). **c**, Genome alignment of the MSY scaffolds with the corresponding female-specific region on chromosome 8. Scaffolds are separated by grey dashed lines. Red lines represent alignments >5 kb on the forward strand, and blue lines represent those on the reverse strand. Pink boxes in **a**–**c** represent the most differentiated regions between the sex chromosomes. **d**, Photographs of microsporophyll and megasporophyll of *C. panzhihuaensis*. Bar, 1 cm. **e**, Sex-specific expression of *MADS-Y* (*CYCAS_034085*) and *CYCAS_010388* in male and female reproductive organs. Microsporophyll tissues were collected before meiosis (BFm), during prophase (Prophase), after meiosis (AFm) and before pollination (BFp); female tissues were collected at 0, 7, 11 and 21 days post-pollination. **f**, Phylogeny of *MADS-Y* homologues across land plants. Genes from MSY and autosomes are marked on the right, and those from *Selaginella* and *Physcomitrium* are used as outgroups. Numbers above branches represent bootstrap scores from IQ-TREE. **g**, Molecular genotyping of male and female cycad samples from *Cycas debaoensis*, *Macrozamia lucida* and *Zamia furfuracea* using primers specific to homologues of *MADS-Y* and *CYCAS_010388*.

COBRA-like protein localizes at the tip of the pollen tube membrane and plays an important role in pollen tube growth and guidance[51] (Supplementary Note 11).

All seed plants produce pollen and deliver their sperm through the growth of a pollen tube, whereas all non-seed land plants (that is, bryophytes, lycophytes and ferns) rely on free-swimming motile sperm for sexual reproduction, as do the ancestors of land plants[1,4] (Extended Data Fig. 7a,b). The exceptions among seed plants are cycads and *Ginkgo*, both of which have pollen grains that release motile spermatozoids that, following pollination, swim the remaining minute distance within the ovule to fertilize the egg[52] (Supplementary Video 1). Sperm motility is conferred by a flagellar apparatus, and most genes related to its assembly occur in the *C. panzhihuaensis* genome. *Ginkgo also* retains flagellar genes, although fewer, and most notably lacks those encoding radial spoke proteins (RSP) (that is, RSP2, RSP3, RSP9 and RSP11; Extended Data Fig. 7c). By contrast, *Gnetum*, conifers and angiosperms, which develop non-flagellated spermatozoa, lost many flagellar

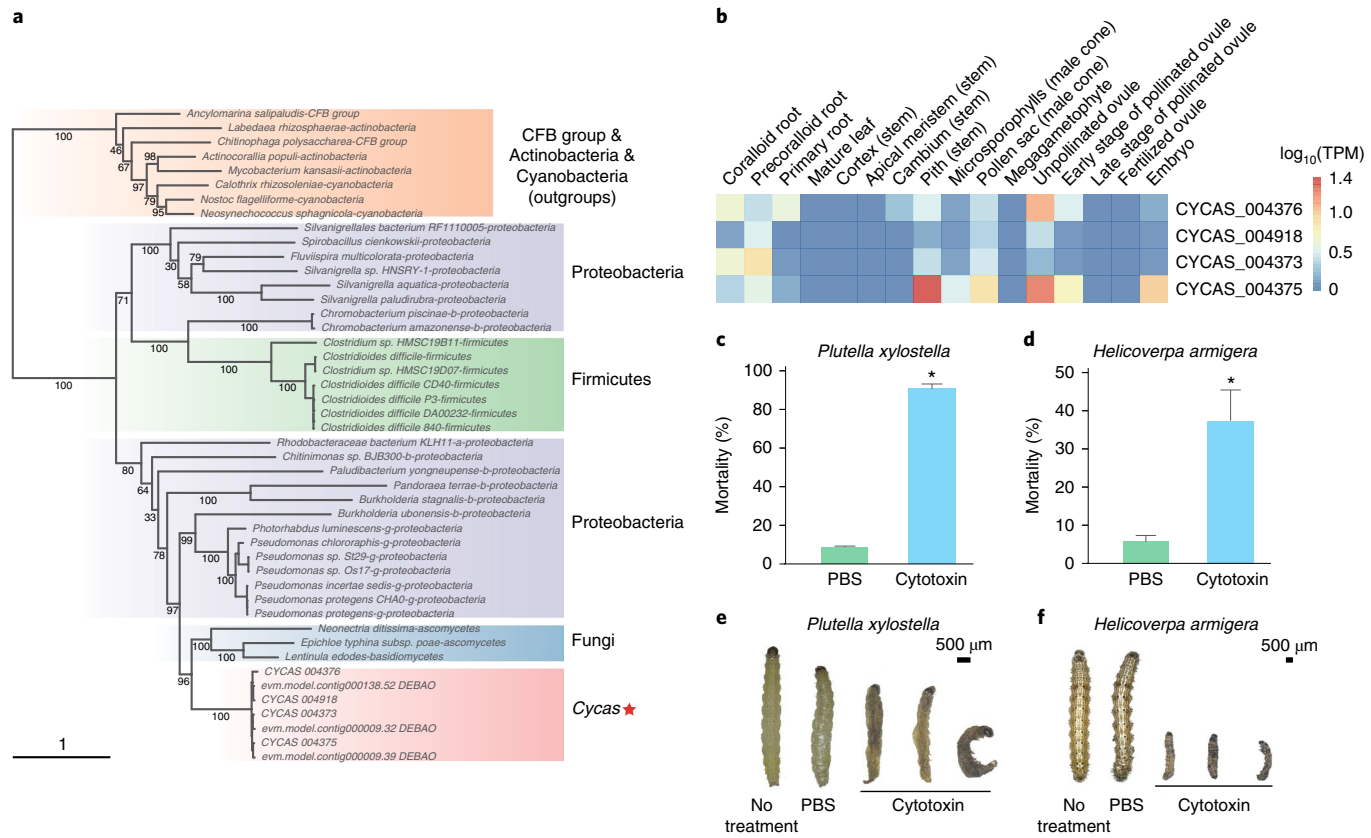


Fig. 5 | Origin of a *Cycas* insecticidal protein. a, Phylogenetic analysis of the TcdA/TcdB pore-forming domain containing proteins shows that the genes encoding four cytotoxin proteins of *Cycas* were likely acquired from fungi through an ancient horizontal gene transfer event. The maximum likelihood tree was generated by RAxML with the PROTCATGTR model and 1,000 bootstrap replicates. The numbers above the branches are bootstrap support values. **b**, The expression level of four cytotoxin proteins in different tissues of *C. panzhihuaensis*. The digital expression values were normalized using the TPM method. **c**,**d**, Mortalities of *Plutella xylostella* (**c**) and *Helicoverpa armigera* (**d**) after treatment with phosphate buffered saline (PBS) and cytotoxin. The asterisk indicates a significant difference (two-sided Student's *t*-test, $P < 0.05$, $n = 3$ biologically independent experiments), whereas the error bar represents the standard error. **e**,**f**, Morphologies of *Plutella xylostella* (**e**) and *Helicoverpa armigera* (**f**) after receiving PBS and cytotoxin treatments.

structural genes (Supplementary Note 12). Outer dense fibres are unique accessory structures that maintain the structural integrity of flagella and are vital for flagellar function[53]. Outer dense fibres exist in *C. panzhihuaensis* and *Gingko biloba*, as well as all non-seed land plants, but are absent in *Gnetum*, conifers and angiosperms, all of which have non-motile sperm (Extended Data Fig. 7c). The shift from swimming to non-motile sperm is a major innovation in land plant evolution, and *C. panzhihuaensis* and *G. biloba* exhibit an ancestral gene content that is part of the shift from producing flagellate to non-flagellate sperm cells.

Sex chromosomes and sex determination in *Cycas*

Heteromorphic chromosomes have been reported to be associated with sex determination in *Cycas*[54]. To reveal the underlying genetic mechanism of sex determination, we carried out genome-wide association studies (GWAS) analysis of sex as a binary phenotype for *C. panzhihuaensis* and identified the most significant association signals on chromosome 8, spanning the first 124 Mb on the reference female genome (Fig. 4a). This sex-associated region is also the most differentiated between male and female *Cycas* genomes, with the largest fixation index (F_{ST}; Supplementary Fig. 37) and the most differentiated nucleotide diversity (π) and heterozygosity ratios characterizing the window between 18 and 50 Mb on chromosome 8 (Fig. 4b and Supplementary Note 13). These results confirm that *Cycas* possesses an XY sex determination system positioned on chromosome 8.

Assembling the male-specific region of the Y chromosome (MSY) based on Nanopore long-read and Hi-C data resulted in 45.5 Mb of sequence distributed over 43 scaffolds, most of which aligned to the sex-differentiation region on chromosome 8 (Fig. 4c and Supplementary Fig. 38). The assembled MSY had an almost 80-Mb difference in length from the corresponding region on the X chromosome, which agrees with the heteromorphy of the *Cycas* sex chromosomes. We annotated 624 putative protein-coding genes within the MSY, 11 of which were highly expressed (transcripts per million (TPM) > 1) in the microsporophylls. The most highly expressed gene in the MSY and also the most differentially regulated gene between the two sexes is *CYCAS_034085* (Fig. 4d,e and Extended Data Fig. 8), which encodes a *GGM13*-like MADS-box transcription factor (TF), belonging to a lineage sister to the angiosperm *AP3/PI* clade that plays crucial roles in floral development. Its closest homologue, *CYCAS_010388*, was identified on autosomal chromosome 2. In contrast to *CYCAS_034085*, *CYCAS_010388* was much more highly expressed in the ovule than in the microsporophyll (Fig. 4e). A male-specific polymerase chain reaction (PCR) product of *CYCAS_034085* was amplified from all tested male cycad samples, but was not detected in female samples, whereas a *CYCAS_010388*-specific PCR product was amplified in both males and females (Fig. 4g and Supplementary Fig. 39b). Because of the presence in MSY and its exclusive expression pattern in males, we named *CYCAS_034085* as *MADS-Y*, a potential sex determination gene.

The reduced size of MSY compared with the X chromosome indicates that the Y chromosome of *Cycas*, unlike that reported for some angiosperms[55], underwent severe degeneration and gene loss. The most divergent 32-Mb region (between the 18 and 50 Mb locations) between the X and Y chromosomes probably represents an ancient evolutionary segment in the *Cycas* sex chromosomes. The broad association of the *MADS-Y* homologue with sex in cycads indicates a conserved sex determination system within this ancient lineage (Fig. 4f and Supplementary Fig. 39). Moreover, the presence of *GbMADS4*, a homologue of the *Cycas MADS-Y*, in *Ginkgo* male-specific contigs[56] suggests that the same mechanism for sex determination might have originated before the split of cycads and *Ginkgo*, thus representing an ancient system of sex determination in seed plants.

Evolution of disease and herbivore resistance genes

All three types of immune receptors—*CC-NBS-LRR* (*CNL*), *TIR-NBS-LRR* (*TNL*) and *RPW8-NBS-LRR* (*RNL*)—show patterns of expansion in *C. panzhihuaensis* and other gymnosperms, compared with non-seed plants (Supplementary Note 14). *CNLs* are expanded widely in both gymnosperms and angiosperms, whereas the *TNL* family tends to have been more expanded in gymnosperms than in most angiosperms, indicating different evolutionary patterns of plant resistance (*R*) genes in these two lineages. Our data suggest that *RNL* genes occur widely in gymnosperms. The *RNL* family plays a critical role in downstream resistance signal transduction in angiosperms, and the broad occurrence of the *RNL* family in gymnosperms suggests that this signalling pathway may have been established no later than the origin of seed plants. Gene families encoding resistance-related proteins are greatly expanded in *C. panzhihuaensis* and other gymnosperm genomes compared with non-seed plants (Supplementary Note 14). For example, genes encoding endochitinases and chitinases as defences against chitin-containing fungal pathogens are expanded as tandem repeats in the *C. panzhihuaensis* and most gymnosperm genomes compared with other land plants.

Cycads comprise many more living species[57] than *Ginkgo*, which was once diverse in the Mesozoic but includes only one extant species[58]. One possible explanation is that cycads may have acquired enhanced resistance to pathogens and herbivores through encoding diversified resistance-related genes and the biosynthesis of diversified secondary compounds[4,8]. Indeed, comparisons of the *Cycas* and *Ginkgo* genomes reveal many *Cycas*-specific orthogroups enriched in pathogen interaction pathways (Supplementary Note 14), and *C. panzhihuaensis* also shows remarkable expansions in plant immunity and stress response gene families compared with *Ginkgo*, including genes that encode programmed cell death, abiotic stress response, serine protease inhibitors against pests and ginkbilobin with antibacterial and antifungal activities (Supplementary Note 14).

Terpenoids are a diverse group of secondary metabolites encoded by terpene synthase (*TPS*) genes[59]. Several TPS subfamilies (*TPS-a* to *TPS-h*) are known in plants[60], among which the *TPS-d* family is unique to gymnosperms, and three of the four types of *TPS-d* were found in *C. panzhihuaensis*, with remarkable expansions of *TPS-d2* compared with *Ginkgo* and most other gymnosperms (Supplementary Note 15). In addition, we identified a novel *TPS* subfamily in *Cycas*, with three copies in *C. panzhihuaensis* and eight copies in *Cycas debaoensis* (Extended Data Fig. 9a). The gene expression levels of all *TPS* genes across different *C. panzhihuaensis* tissues (Extended Data Fig. 9b) reveal that many *TPS* genes are mainly expressed in the root (especially primary root and coralloid root), microsporophyll and pollen sac, late stage of the pollinated ovule and fertilized ovule. The three *Cycas*-specific *TPS* genes were mainly expressed in the root and male cone, but one of them (*CYCAS_009486*) is particularly highly expressed in the megagametophyte and in the post-pollination and fertilized ovule.

Cycas obtained a cytotoxin defence gene via horizontal gene transfer

Genes of fungal or bacterial origin are rare in seed plants[61]. However, we identified a gene family in the *C. panzhihuaensis* genome that appears to have been acquired from a microbial organism and that codes for a *Pseudomonas* fluorescens insecticidal toxin (*fitD*). The acquired genes are flanked by vertically inherited plant sequences. We further confirmed that the relevant assembled regions were free of bacterial contamination. Transcriptomes and PCR amplification from genomic DNA indicated that these genes occur in many *Cycas* species (Supplementary Note 16). The *fitD* gene family comprises four gene copies in the *C. panzhihuaensis* genome and three copies in the *C. debaoensis* genome (Supplementary Table 51); each copy encodes a protein that is similar to the fit toxin and the 'makes caterpillars floppy' (mcf) toxin of the bacterium *Photorhabdus luminescens*, a lethal pathogen of insects. Both fit and mcf toxins are known for their insecticidal properties, and fit- or mcf-producing bacteria are often used in pest biocontrol[62–64]. Phylogenetic analyses suggest that the *fitD* genes might have been acquired from fungi and then expanded before the divergence of *C. panzhihuaensis* and *C. debaoensis* (Fig. 5a). The *fitD* family genes are mainly expressed in roots, reproductive tissues such as male cones, unpollinated or early stages of pollinated ovules and embryos (Fig. 5b). Injection of the synthesized *C. panzhihuaensis fitD* protein resulted in significantly higher mortality in larvae of both the diamondback moth (*Plutella xylostella*) and cotton bollworm (*Helicoverpa armigera*) (Fig. 5c,d). The acquisition of the *fitD* gene family may have provided an important defence for *Cycas* against insect pests.

Conclusions

The high-quality genome sequence for *Cycas*, the last major lineage of seed plants for which a high-quality genome assembly was lacking, closes an important gap in our understanding of genome structure and evolution in seed plants. This genome enables comparative genomics and phylogenomic analyses to unravel the genetic control of important traits in cycads and other gymnosperms, including a WGD shared by gymnosperms, a sex determination mechanism that appears to be shared by cycads and *Ginkgo*, and critical gene innovations including those that enable seed and pollen tube formation, as well as chemical defence.

Methods

Plant materials. Fresh megagametophytes of *Cycas panzhihuaensis*, cultivated in the garden of the Kunming Institute of Botany, Chinese Academy of Sciences, were collected for genome sequencing. The plant was originally transplanted from the Pudu River, Luquan county, Yunnan, China (25° 57′ 35.2584″ N, 102° 43′ 41.5848″ E) and the voucher specimen (collection number: PZHF03) has been deposited in the Herbarium of the Kunming Institute of Botany (KUN). For transcriptome sequencing, we sampled 12 different types of organs and tissues from *C. panzhihuaensis*, including megagametophyte, pollen sac, microsporophylls, apical meristem of stem, cortex of stem, pith of stem, cambium of stem, mature leaf, young leaf, primary root, precoralloid roots and coralloid roots (Supplementary Table 2). Ovule material was collected from two artificially pollinated individuals, and we divided the development stages into four: unpollinated ovule (before the artificial pollination), early stage of pollinated ovule (21 d after the artificial pollination), late stage of pollinated ovule (88 d after the artificial pollination) and fertilized ovule or seed (119 d after the artificial pollination) (Supplementary Tables 2 and 19). In addition, stem and root tissues of *C. panzhihuaensis* were used to generate full-length transcriptomes (Supplementary Table 2). For phylogenomic analyses, we newly generated transcriptomes of 47 gymnosperms (Supplementary Tables 2 and 13). We also sequenced transcriptomes of 339 cycad species (Supplementary Tables 2 and 14). For population resequencing, fresh leaf samples were collected for 31 male and 31 female plants that were randomly sampled in the *Cycas panzhihuaensis* National Natural Reserve in Sichuan, China, where there is a population of approximately 38,000 *C. panzhihuaensis* individuals (Supplementary Table 4).

DNA and RNA sequencing. For genome sequencing, the genomic DNA was extracted by the QIAGEN Genomic kit followed the manufacturer's instructions[65]. Nanodrop and Qubit (Invitrogen) were used to quantify the DNA. Nanopore libraries were prepared by SQK-LSK108 and sequenced using

a Nanopore PromethION sequencer. The rest of the DNA was used to generate short-read sequences using an MGI-SEQ platform, with 150-bp read length and 300–500 DNA-fragment insert size. Hi-C libraries were created from fresh megagametophyte, following a previously published method[66]. Briefly, the tissue was fixed in formaldehyde, lysed and the cross-linked DNA was digested overnight with HindIII. Sticky ends were biotinylated and proximity-ligated to generate chimeric junctions, which were subsequently physically sheared to 500–700 bp in size. The initial cross-linked long-distance physical interactions were then represented by chimeric fragments, which were processed into paired-end sequencing libraries. Paired-end reads were produced on both the MGI-SEQ and Illumina HiSeq X platforms. See Supplementary Note 3 for details on transcriptome, organelle genome and small RNA sequencing.

Genome assembly. About 1,010 Gb (~100×) Nanopore long-read data were used for genome assembly using NextDenovo (https://github.com/Nextomics/NextDenovo) with default parameters (read_cutoff = 1k, seed_cutoff = 12k, minimap2_options_cns = -x ava-ont -k17 -w17). To further enhance assembly contiguity, about 456 Gb of Hi-C data were used to execute Hi-C chromosome conformation in conjunction with 3D-dna algorithm[67]. The accuracy of Hi-C based chromosomal assembly was assessed using Juicerbox's chromatin contact matrix.

Repeat annotation. We identified tandem repeats and transposable elements throughout the genome. Tandem repeats were predicted using Tandem Repeat Finder (v.4.07)[68] with the following parameters: 'Match = 2, Mismatch = 7, Delta = 7, PM = 80, PI = 10, Minscore = 50 and MaxPeriod = 2,000'. To maximize the opportunity of identifying transposable elements, a combination of de novo and homology-based approaches was performed following the Repeat Library Construction-Advanced pipeline (http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction-Advanced). RepeatMasker[69] and RepeatProteinMask[69] were used to search for known repeat sequences; MITE-hunter[70], LTR_retriever[71], LTR_FINDER (v.1.0.6)[72] and RepeatModeler[73] were then used to search the repeats de novo. The MITE, LTR and consensus repetitive libraries generated by RepeatModeler were combined and further used as the input data for RepeatMasker.

LTR identification and estimation of LTR insertion times. All the candidate LTR elements were first identified using LTR_FINDER and LTR_retriever. LTR_STRUC[74] was then used to extract the complete 5′- and 3′-ends of the LTR elements. RepeatClassifier was then used to classify the candidate LTR. Distmat from the EMBOSS (v.6.5.7.0) package was then used to calculate the K value of the retrotransposons' 5′- and 3′-LTR sequences. Finally, the insertion time (T) of LTRs was calculated using the formula $T = K/2r$, where r is the average substitution rate of 2.2×10^{-9} substitutions per year per synonymous site.

Gene annotation and functional annotation. Three types of evidence were used to predict protein-coding genes in the *C. panzhihuaensis* genome. For protein evidence, Genewise[75] was used to predict gene models based on *Cycas* proteins downloaded from the UniProt protein database and other proteins collected from representative plant species. Next, Hisat[76] was used to map the transcriptome to the genome, and then StringTie[77] was used to predict transcriptome-based gene models. Next, a custom training hint parameter was used to predict ab initio-based gene models in AUGUSTUS[78]. All the evidence was finally combined and integrated by EVidenceModeler[79]. To maximize the opportunity of identifying high-confidence genes, we further filtered the genes that were not expressed in the full-length transcriptome or did not match to functional annotation results. For functional annotation, the gene models were blasted against the UniProt, TrEMBL, KEGG, KOG and NR databases. The domain and gene ontology of the gene models was identified by InterProScan[80] (using data from Pfam, PRINTS, SMART, ProDom and PROSITE).

Identification of key candidate functional genes. Based on the following criteria, all candidate genes were screened: first, candidate gene sequences were detected by BLAST searches with an e value cut-off of $1 \times 10^{-5}$to the collected query gene sequences gathered from previous studies or public databases; and second, features of candidate genes should be similar to the online functional annotation or UniProt functional annotation as the query genes. With regard to the identification of flagellar genes, 58 flagellar-related genes were collected from previous studies[81]. The Reciprocal Best Blast hit method was employed to identify flagella-related genes. For seed-related genes, we searched the genes against both the known seed database (seedgenes.org/) and previous studies. We firstly used an e value ($<1 \times 10^{-20}$) as a cut-off to filter candidates and then filtered the candidates with functional annotation. Regarding the identification of TFs, we used the HMMER search method. HMMER domain structure models were downloaded from the Pfam website (https://pfam.xfam.org/), for each TF as present in the TAPscan v.2 database for TFs (https://plantcode.online.uni-marburg.de/tapscan/). Preliminary TF candidate genes were collected for each species ($<1 \times 10^{-5}$) by searching the Hidden Markov Model profile. Parts of genes were then filtered if they were not the homologues according to their functional annotation of SwissProt ($<1 \times 10^{-5}$). In the end, we filtered genes containing a wrong domain under the TAPscan v.2 transcription factor database domain rules. Phylogenetic tree analysis was used to verify the majority of TFs and transcriptional regulators. Details about phylogenetic tree reconstruction for each TF can be found in the figure captions.

Phylogenetic reconstruction and divergence-time estimation. *Nuclear phylogenetic reconstruction.* The downloaded genome sequences and the newly generated genome sequences of *C. panzhihuaensis* were used to construct the orthogroups using OrthoFinder[82] with default settings. The software KinFin[83] was used to select single-copy gene families for phylogenetic reconstruction with default parameters. TranslatorX[84] was used to build gene alignments for codon (nt), codon 1st + 2nd (nt12) and amino acid (aa) sequences (command: perl translatorx_vLocal.pl -i gene.fa -o gene.out -p F -t F -w 1 -c 1 -g "-b1="$b1" -b2="$b1" -b3=8 -b4=5 -b5=h -b6=y"). IQ-TREE 2 (ref. [85]) was used to infer the maximum likelihood trees with an initial partition scheme of codon positions combing ModelFinder, tree search, and ultrafast bootstrap. ASTRAL[86] was used to summarize the coalescent species tree and the quartet supports with default settings (-t 8). ASTRAL uses the quartet trees of the maximum likelihood phylogenies of each gene to produce the topology of the species tree while quartet supports (bar charts) show the percentage of quartets that agree with a specific branch in the species tree. STAG (https://github.com/davidemms/STAG) was also used to construct the species tree with default settings using low-copy genes (one to four copies). The software PHYPARTS[87] was used to infer and visualize the gene tree conflicts on the species tree topology with default settings. The software DISCOVISTA[88] was used to summarize the conflicts among different analytical methods and datasets, regarding several focal phylogenetic relationships.

Molecular dating and diversification analysis. The transcriptome sequencing reads from 339 cycad species were generated in the current study. Clean reads were assembled with TRINITY[89], and the longest transcripts were selected and translated with TRANSDECODER (https://github.com/TransDecoder). OrthoFinder[82] was then used to construct orthogroups for all the cycad species using *Ginkgo* as the outgroup. The software KinFin[83] was used to select the mostly single-copy genes for phylogenetic reconstruction with default settings. TranslatorX[84], IQ-TREE 2 (ref. [85]) and ASTRAL[86] were used to align the sequences and to infer the species tree for cycads as aforementioned. The software SORTADATE[90] was used to select genes with mostly concordant evolutionary histories for dating analyses using MCMCTREE within the software PAML 4 (ref. [91]). Rate priors and time priors were set following the method of Morris et al.[92]. A total of 27 fossils were used to calibrate the chronogram of seed plants, and six fossils for the chronogram of cycads. The diversification pattern for cycads were analysed with Bayesian analysis of macroevolutionary mixture (www.bamm-project.org) following Condamine et al.[93]

See Supplementary Note 5 for details on organellar phylogenetic reconstruction, evaluation of the impact of RNA editing and investigation of cyto-nuclear incongruences.

Identification of whole-genome duplication. An integrated phylogenomic approach and a method to analyse synteny as described previously[35,94,95] were used to identify the WGD events in seed plant evolution. The protein-coding sequences of 15 completely sequenced genomes and 1 transcriptome, representing seven gymnosperms (*C. panzhihuaensis*, *Encephalatos longifolius, G. biloba*, *Gnetum montanum*, *Picea abies*, *Pinus taeda* and *Sequoiadendron giganteum*), six angiosperms (*Arabidopsis thaliana*, *Amborella trichopoda*, *Cinnamomum micranthum*, *Liriodendron chinense*, *Nymphaea colorata* and *Oryza sativa*) and three other vascular plant outgroups (*Azolla filiculoides*, *Salvinia cucullate* and *Selaginella moellendorffii*), were classified into putative gene families/subfamilies by OrthoFinder[82], and then scored for gene duplications across global gene families. For the phylogenetic analysis of gene families, amino acid sequences of each gene family were first aligned with MAFFT[96], the program PAL2NAL[97] was then used to construct their corresponding nucleotide sequence alignments. We used trimAl[98] to remove poorly aligned portions of alignments using the 'automated1' option, which implements a heuristic algorithm to optimize the process for trimming the alignment. Finally, maximum likelihood trees were calculated using RAxML[99] with the GTRGAMMA model and bootstrap support was estimated based on 100 replicates. Following Wu et al.[95], we applied two basic requirements for the determination of a reliable duplication event: (1) at least one common species' genes are present in two child branches; and (2) the bootstrap values of the parental node and one of the child nodes are both ≥50%. After scoring gene duplications in a large-scale analysis on gene families, we were able to confidently identify the nodes with concentrated gene duplications across the phylogeny, which possibly support the WGD events. Furthermore, because syntenic information is the most solid evidence for WGD, and the legacy of syntenic blocks may be found if the concentrated gene duplications are indeed derived from WGD events, we also looked into whether such syntenic blocks exist. The intra- and intergenomic syntenic analyses were conducted using MCscanX[100], with the default settings.

In addition, the Nei–Gojobori method[101] as implemented in the PAML package's yn00 program[91] was used to estimate synonymous substitutions per synonymous site (K_S) for pairwise comparisons of paralogous genes located on syntenic blocks. To search for genome-wide duplications, we used DupGen_finder

(https://github.com/qiao-xin/DupGen_finder) to identify duplicated genes that were classified into five different categories: WGD duplicates, tandem duplicates, proximal duplicates, transposed duplicates and dispersed duplicates.

Identification of the sex-differentiation region. To identify the sex-differentiation region in the *Cycas* genome, a GWAS approach was adopted on sequence variations from 31 male and 31 female individuals with sex treated as a binary phenotype. Briefly, raw reads were filtered by Trimmomatic (v.0.38) (ILLUMINACLIP:adapter.fa:2:30:10 HEADCROP:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:140), and read alignment and single-nucleotide polymorphism (SNP) calling were performed using the Sentieon pipeline[102]. SNPs were filtered using the following criteria: (1) SNPs were filtered by GATK VariantFiltrations with 'QD < 2.0 || FS > 60.0 || MQ < 40.0 || SOR > 3.0 || MQRankSum < −12.5 || ReadPosRankSum < −8.0', and indels with 'QD < 2.0 || FS > 200.0 || SOR > 10.0 || MQRankSum < −12.5 || ReadPosRankSum < −8.0'; (2) total depth <80 or >1,300; (3) variants with more than two alleles; (4) variants with a missing rate >10% or minor allele frequencies <0.1 were removed; and (5) a linkage disequilibrium pruning with PLINK (v.1.9) using a window size of 10 kb with a step size of one SNP and r^2 threshold of 0.5, resulting a 4.65-million pruned SNP set for association analysis of sex differentiation. GWAS analysis of sex differentiation was performed on the linkage disequilibrium-pruned SNP set using the EMMAX program[103] (beta-07Mar2010 version). The BN kinship matrix and the first five components calculated from the principal component analysis[104] (v.1.91.4beta3) were included as random effects. Genetic differentiation (F_{ST}) and nucleotide diversity (π) were calculated within a non-overlapping 100-kb window using VCFtools[105] (v.0.1.13). See Supplementary Note 13 for details on assembly of *Cycas* male-specific regions, phylogenetic analysis of MADS-Y and CYCAS_010388 homologues, and genotyping of cycad male and female samples.

Analysis of the differentially expressed genes. Transcriptome sequencing reads were trimmed using Trimmomatic[106] program (ILLUMINACLIP:adapter.fa:2:30:10 HEADCROP:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:140) and mapped against *C. panzhihuaensis* annotated gene models using bowtie2 (with sensitive mode and default alignment parameters) by retaining the best alignments. TPM were calculated using the eXpress program, which was incorporated in the Trinity[89] package. Furthermore, differentially expressed genes with a differential expression level of false discovery rate ≤0.01 and at least a twofold expression change were identified using DESeq2 (ref. [107]). To identify the co-expressed genes during the seed development, we used the R package WGCNA[108] on the basis of the TPM data of the genes whose expression showed a coefficient of variation >0.5 across the four stages. To better visualize the expression levels, we normalized the expression results. For each gene, the TPM value normalized by the maximum TPM value of all stages is shown. Fisher's exact test was used to examine whether the functional categories were over-represented. The resulting *P* values were adjusted to *Q* values by the false discovery rate correction.

Identification of the horizontally transferred cytotoxin genes in *C. panzhihuaensis*. The cytotoxin protein sequences of *Cycas* were used as query to perform BLASTP searches against the NCBI nr protein sequence database using the cut-off e value = 1×10^{-5} and max_target_seqs = 20,000. We also performed additional BLAST searches against the OneKP database and many other available genomes. See Supplementary Note 16 for details on verification and phylogenetic analysis of the cytotoxin gene.

Assessing the effectiveness of cytotoxin. To improve the expression efficiency of cytotoxin in the prokaryotic system, the full-length coding sequence of the *C. panzhihuaensis* cytotoxin protein was optimized for its codons. *C. panzhihuaensis*, the optimized sequence was synthesized and ligated to the pET-28a vector. The pET-28a-CR toxin plasmid was transformed into *Escherichia coli* BL 21 (DE3) pLysS cells, the resulting strain was used for expression and purification of recombinant proteins under the control of isopropyl-β-D-thiogalactoside-inducible T7 promoter. Overnight-grown cultures were diluted 100-fold with 200 ml of fresh LB medium and further grown at 37 °C and 220 r.p.m. rotation until the optical density at 600nm reached 0.5. The culture was induced by adding a 0.01 mM final concentration of isopropyl-β-D-thiogalactoside and incubated at 28 °C for 6 h. Cells were then harvested and suspended with 20 ml 50of mM Tris–HCl buffer with pH 8 at 4 °C, containing 200 mM NaCl, then disrupted by sonication at 4 °C. In an RC5 plus centrifuge, the cell lysate was spun at 13,800*g* for 40 min at 4 °C. The preceding step's supernatant was put onto a Ni-NTA agarose column that had been pre-equilibrated with Tris–NaCl buffer at 4 °C. Tris–NaCl buffer containing 20 mM imidazole was used to thoroughly wash the column, and the 6× His-tagged protein was eluted with Tris–NaCl buffer containing 250 mM imidazole. The elution product containing pure protein were washed three times with Tris–NaCl buffer and concentrated using centricon (Millipore PM10). Using an horseradish peroxidase-conjugated monoclonal antibody and a western blot assay, the purified His-tagged protein was identified (HRP-66005). See Supplementary Note 16 for further details on experimental verification of the function of *Cycas* cytotoxin.

Detection of metabolites and phytohormones. The plant tissues were collected and stored in liquid nitrogen, then transferred to freezer at −80 °C. For detection of metabolites, tissue samples were preliminarily disposed using 2-chlorophenylalanine (4 ppm) methanol. Samples and glass beads were then put into a tissue grinder for 90 s at 55 Hz, followed by centrifugation at 13,780*g* at 4 °C for 10 min, taking the supernatant and filtering through a 0.22-μm membrane, and transferring the filtrate into the detection bottle before liquid chromatography mass spectrometry analysis. The sample extracts were the analysed using the ultra high-performance liquid chromatography system Vanquish (ThermoFisher Scientific) and Q Exactive HF-X (ThermoFisher Scientific). For the quantitative detection of phytohormones (auxin, cytokinins, ethylene, abscisic acid, jasmonic acid, gibberellin, salicylic acid and brassinolide), tissue samples of primary root, precoralloid roots and coralloid roots, unpollinated ovule, early stage of pollinated ovule, late stage of pollinated ovule, fertilized ovule and mature embryo were collected. Vanquish (ThermoFisher Scientific) and the Q Exactive HF-X (ThermoFisher Scientific) were used for the detection of various phytohormones. The qualitative study was carried out using a self-constructed database that was built using the reference standards. To accomplish quantitative analysis, different concentrations of standard were utilized.

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

Data availability

The genome and transcriptome data, genome assemblies and annotations can be found at https://db.cngb.org/codeplot/datasets/public_dataset?id=PwRftGHfPs5qG3gE. The raw genomic, transcriptomic and Hi-C data generated in this study were deposited in the NCBI Sequence Read Archive (SRA, BioProject PRJNA734434) and the CNGB data center (https://db.cngb.org/) under project number CNP0001756. Source data are provided with this paper.

Received: 3 September 2021; Accepted: 10 March 2022;
Published online: 18 April 2022

References

1. Raven, P. H., Evert, R. F. & Eichhorn, S. E. *Biology of Plants* 7th edn (Macmillan, 2005).
2. Nagalingum, N. S. et al. Recent synchronous radiation of a living fossil. *Science* **334**, 796–799 (2011).
3. Condamine, F. L., Nagalingum, N. S., Marshall, C. R. & Morlon, H. Origin and diversification of living cycads: a cautionary tale on the impact of the branching process prior in Bayesian molecular dating. *BMC Evol. Biol.* **15**, 65 (2015).
4. Norstog, T. J. & Nicholls, K. J. *The Biology of the Cycads* (Cornell Univ. Press, 1997).
5. Calonje, M., Stevenson, D. W. & Osborne, R. *The World List of Cycads* http://www.cycadlist.org (2013–2021).
6. Sultana, M., Mukherjee, K. K. & Gangopadhyay, G. in *Reproductive Biology of Plants* (eds Johri, B. M. & Srivastava, P. S.) 118–132 (Springer Science & Business Media, 2014).
7. Paolillo, D. J. Jr The swimming sperms of land plants. *BioScience* **31**, 367–373 (1981).
8. Brenner, E. D., Stevenson, D. W. & Twigg, R. W. Cycads: evolutionary innovations and the role of plant-derived neurotoxins. *Trends Plant Sci.* **8**, 446–452 (2003).
9. Costa, J.-L. & Lindblad, P. in *Cyanobacteria in Symbiosis* (eds Rai, A. N. et al.) 195–205 (Springer, 2002).
10. Pettitt, J. Heterospory and the origin of the seed habit. *Biol. Rev.* **45**, 401–415 (1970).
11. Yang, D.-Q. & Zhu, X.-F. Karyotype analysis of *Cycas panzhihuaensis* L. Zhou et S. Y. Yang. *J. Syst. Evol.* **23**, 352–354 (1985).
12. 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).
13. Guan, R. et al. Draft genome of the living fossil *Ginkgo biloba*. *GigaScience* **5**, 49 (2016).
14. Liu, H. et al. The nearly complete genome of *Ginkgo biloba* illuminates gymnosperm evolution. *Nat. Plants* **7**, 748–756 (2021).
15. Wan, T. et al. A genome for gnetophytes and early evolution of seed plants. *Nat. Plants* **4**, 82–89 (2018).
16. Nystedt, B. et al. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**, 579–584 (2013).
17. Stevens, K. A. et al. Sequence of the sugar pine megagenome. *Genetics* **204**, 1613–1626 (2016).
18. Niu, S. et al. The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* **185**, 204–217 (2021).

19. Ran, J.-H., Shen, T.-T., Wang, M.-M. & Wang, X.-Q. Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. *Proc. Biol. Sci.* **285**, 20181012 (2018).
20. Li, Z. et al. Single-copy genes as molecular markers for phylogenomic studies in seed plants. *Genome Biol. Evol.* **9**, 1130–1147 (2017).
21. Xi, Z., Rest, J. S. & Davis, C. C. Phylogenomics and coalescent analyses resolve extant seed plant relationships. *PLoS ONE* **8**, e80870 (2013).
22. Soltis, D. et al. *Phylogeny and Evolution of the Angiosperms: Revised and Updated Edition* (Univ. of Chicago Press, 2018).
23. Leebens-Mack, J. H. et al. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
24. Stull, G. W. et al. Gene duplications and phylogenomic conflict underlie major pulses of phenotypic evolution in gymnosperms. *Nat. Plants* **7**, 1015–1025 (2021).
25. Dong, S., Li, H., Goffinet, B. & Liu, Y. Exploring the impact of RNA editing on mitochondrial phylogenetic analyses in liverworts, an early land plant lineage. *J. Syst. Evol.* **60**, 16–22 (2021).
26. Du, X.-Y., Lu, J.-M. & Li, D.-Z. Extreme plastid RNA editing may confound phylogenetic reconstruction: A case study of *Selaginella* (lycophytes). *Plant Divers.* **42**, 356–361 (2020).
27. Wen, D., Yu, Y., Zhu, J. & Nakhleh, L. Inferring phylogenetic networks using PhyloNet. *Syst. Biol.* **67**, 735–740 (2018).
28. Zachos, J., Pagani, M., Sloan, L., Thomas, E. & Billups, K. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**, 686–693 (2001).
29. Folk, R. A. et al. Rates of niche and phenotype evolution lag behind diversification in a temperate radiation. *Proc. Natl Acad. Sci. USA* **116**, 10874–10882 (2019).
30. Sun, M. et al. Recent accelerated diversification in rosids occurred outside the tropics. *Nat. Commun.* **11**, 1–12 (2020).
31. Soltis, P. S., Folk, R. A. & Soltis, D. E. Darwin review: angiosperm phylogeny and evolutionary radiations. *Proc. Biol. Sci.* **286**, 20190099 (2019).
32. Van de Peer, Y., Ashman, T.-L., Soltis, P. S. & Soltis, D. E. Polyploidy: an evolutionary and ecological force in stressful times. *Plant Cell* **33**, 11–26 (2021).
33. Vanneste, K., Van de Peer, Y. & Maere, S. Inference of genome duplications from age distributions revisited. *Mol. Biol. Evol.* **30**, 177–190 (2013).
34. Roodt, D. et al. Evidence for an ancient whole genome duplication in the cycad lineage. *PLoS ONE* **12**, e0184454 (2017).
35. Jiao, Y. et al. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100 (2011).
36. Doyle, J. A. Phylogenetic analyses and morphological innovations in land plants. *Annu. Plant Rev.* **45**, 1–50 (2018).
37. Tzafrir, I. et al. The *Arabidopsis* SeedGenes Project. *Nucleic Acids Res.* **31**, 90–93 (2003).
38. Lepiniec, L. et al. Molecular and epigenetic regulations and functions of the LAFL transcriptional regulators that control seed development. *Plant Reprod.* **31**, 291–307 (2018).
39. Gomez, M. D., Ventimilla, D., Sacristan, R. & Perez-Amador, M. A. Gibberellins regulate ovule integument development by interfering with the transcription factor ATS. *Plant Physiol.* **172**, 2403–2415 (2016).
40. Staszak, A. M., Rewers, M., Sliwinska, E., Klupczyńska, E. A. & Pawłowski, T. A. DNA synthesis pattern, proteome, and ABA and GA signalling in developing seeds of Norway maple (*Acer platanoides*). *Funct. Plant Biol.* **46**, 152–164 (2019).
41. Spicer, R. & Groover, A. Evolution of development of vascular cambia and secondary growth. *New Phytol.* **186**, 577–592 (2010).
42. Baucher, M., El Jaziri, M. & Vandeputte, O. From primary to secondary growth: origin and development of the vascular system. *J. Exp. Bot.* **58**, 3485–3501 (2007).
43. Mähönen, A. P. et al. A novel two-component hybrid molecule regulates vascular morphogenesis of the *Arabidopsis* root. *Genes Dev.* **14**, 2938–2943 (2000).
44. Caño-Delgado, A. et al. BRL1 and BRL3 are novel brassinosteroid receptors that function in vascular differentiation in *Arabidopsis*. *Development* **131**, 5341–5351 (2004).
45. Harris, P. J. in *Plant Diversity and Evolution: Genotypic and Phenotypic Variation in Higher Plants* (ed. Henry, R. J.) 201–227 (CAB International, 2005).
46. Yin, Y., Huang, J. & Xu, Y. The cellulose synthase superfamily in fully sequenced plants and algae. *BMC Plant Biol.* **9**, 99 (2009).
47. Dumont, M. et al. The cell wall pectic polymer rhamnogalacturonan-II is required for proper pollen tube elongation: implications of a putative sialyltransferase-like protein. *Ann. Bot.* **114**, 1177–1188 (2014).
48. Sprunck, S. et al. Egg cell-secreted EC1 triggers sperm cell activation during double fertilization. *Science* **338**, 1093–1097 (2012).
49. Prado, N. et al. Nanovesicles are secreted during pollen germination and pollen tube growth: a possible role in fertilization. *Mol. Plant* **7**, 573–577 (2014).
50. Neudecker, P. et al. Solution structure, dynamics, and hydrodynamics of the calcium-bound cross-reactive birch pollen allergen Bet v 4 reveal a canonical monomeric two EF-hand assembly with a regulatory function. *J. Mol. Biol.* **336**, 1141–1157 (2004).
51. Higashiyama, T. & Takeuchi, H. The mechanism and key molecules involved in pollen tube guidance. *Annu. Rev. Plant Biol.* **66**, 393–413 (2015).
52. Bold, H. C., Alexopoulos, C. J. & Delevoryas, T. *Morphology of Plants and Fungi* 5th edn (Harper and Row, 1987).
53. Zhao, W. et al. Outer dense fibers stabilize the axoneme to maintain sperm motility. *J. Cell. Mol. Med.* **22**, 1755–1768 (2018).
54. Abraham, A. & Mathew, P. M. Cytological studies in the cycads: sex chromosomes in *Cycas*. *Ann. Bot.* **26**, 261–266 (1962).
55. Ming, R., Bendahmane, A. & Renner, S. S. Sex chromosomes in land plants. *Annu. Rev. Plant Biol.* **62**, 485–514 (2011).
56. Liao, Q. et al. The genomic architecture of the sex-determining region and sex-related metabolic variation in *Ginkgo biloba*. *Plant J.* **104**, 1399–1409 (2020).
57. Jones, D. L. *Cycads of the World: Ancient Plants in Today's Landscape* 2nd edn (Smithsonian Institution Press, 2002).
58. Crane, P. R. An evolutionary and cultural biography of ginkgo. *Plants People Planet* **1**, 32–37 (2019).
59. Zhou, F. & Pichersky, E. More is better: the diversity of terpene metabolism in plants. *Curr. Opin. Plant Biol.* **55**, 1–10 (2020).
60. 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).
61. Chen, R. et al. Adaptive innovation of green plants by horizontal gene transfer. *Biotechnol. Adv.* **46**, 107671 (2020).
62. Ruffner, B. et al. Oral insecticidal activity of plant-associated pseudomonads. *Environ. Microbiol.* **15**, 751–763 (2013).
63. Daborn, P. J., Waterfield, N., Silva, C. P., Au, C. P. Y. & Sharma, S. A single Photorhabdus gene, makes caterpillars floppy (mcf), allows *Escherichia coli* to persist within and kill insects. *Proc. Natl Acad. Sci. USA* **99**, 10742–10747 (2002).
64. Péchy-Tarr, M. et al. Molecular analysis of a novel gene cluster encoding an insect toxin in plant-associated strains of *Pseudomonas fluorescens*. *Environ. Microbiol.* **10**, 2368–2386 (2008).
65. Sahu, S. K., Thangaraj, M. & Kathiresan, K. DNA extraction protocol for plants with high levels of secondary metabolites and polysaccharides without using liquid nitrogen and phenol. *ISRN Mol. Biol.* **2012**, 205049 (2012).
66. Xie, T. et al. De novo plant genome assembly based on chromatin interactions: a case study of *Arabidopsis thaliana*. *Mol. Plant* **8**, 489–492 (2015).
67. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
68. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
69. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **5**, 4.10.11–14.10.14 (2004).
70. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199 (2010).
71. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
72. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
73. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
74. McCarthy, E. M. & McDonald, J. F. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **19**, 362–367 (2003).
75. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome Res.* **14**, 988–995 (2004).
76. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
77. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
78. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).

79. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
80. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
81. Li, L. et al. The genome of *Prasinoderma coloniale* unveils the existence of a third phylum within green plants. *Nat. Ecol. Evol.* **4**, 1220–1231 (2020).
82. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
83. Laetsch, D. R. & Blaxter, M. L. KinFin: software for taxon-aware analysis of clustered protein sequences. *G3 (Bethesda)* **7**, 3349–3357 (2017).
84. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38**, W7–W13 (2010).
85. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
86. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinform.* **19**, 15–30 (2018).
87. Smith, S. A., Moore, M. J., Brown, J. W. & Yang, Y. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* **15**, 150 (2015).
88. Sayyari, E., Whitfield, J. B. & Mirarab, S. DiscoVista: Interpretable visualizations of gene tree discordance. *Mol. Phylogenet. Evol.* **122**, 110–115 (2018).
89. 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).
90. Smith, S. A., Brown, J. W. & Walker, J. F. So many genes, so little time: A practical approach to divergence-time estimation in the genomic era. *PLoS ONE* **13**, e0197433 (2018).
91. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
92. Morris, J. L. et al. The timescale of early land plant evolution. *Proc. Natl Acad. Sci. USA* **115**, E2274–E2283 (2018).
93. Condamine, F. L., Rolland, J., Höhna, S., Sperling, F. A. & Sanmartín, I. Testing the role of the Red Queen and Court Jester as drivers of the macroevolution of Apollo butterflies. *Syst. Biol.* **67**, 940–964 (2018).
94. Jiao, Y., Li, J., Tang, H. & Paterson, A. H. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell* **26**, 2792–2802 (2014).
95. Wu, S., Han, B. & Jiao, Y. Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Mol. Plant* **13**, 59–71 (2020).
96. Katoh, K., Kuma, K.-i, Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).
97. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
98. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
99. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
100. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
101. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
102. Kendig, K. I. et al. Sentieon DNASeq variant calling workflow demonstrates strong computational performance and accuracy. *Front. Genet.* **10**, 736 (2019).
103. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
104. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
105. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
106. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
107. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
108. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559 (2008).

Acknowledgements

This study was supported by the Scientific Foundation of Urban Management Bureau of Shenzhen (No. 201916 to Yang Liu, No. 202019 to Shouzhou Zhang and No. 202105 to Y.G.), the National Key R&D Program of China (No. 2019YFC1711000 to Huan Liu), the Biodiversity Survey and Assessment Project of the Ministry of Ecology and Environment, China (No. 2019HJ2096001006 to Shouzhou Zhang and Yongbo Liu), the Major Science and Technology Projects of Yunnan Province (Digitalization, development and application of biotic resource, No. 860 202002AA100007 to Huan Liu) and Shenzhen Municipal Government of China (No. JCYJ20151015162041454 to Huan Liu). Y.V.d.P. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 833522) and from Ghent University (Methusalem funding, BOF.MET.2021.0005.01). Plant illustrations were drawn by S. Li, Z. Li, D. Cui and X. Zeng. We are grateful to the Orchid Conservation and Research Centre of Shenzhen for allowing us to access their computing resources. We also acknowledge T. Wan (Fairy Lake Botanical Garden) and D. Stevenson (New York Botanical Garden), who kindly commented on an earlier draft of the manuscript, and T. Takaso (University of the Ryukyus), who provided the video for swimming sperm of *Cycas*. The study was supported by the National Cycad Conservation Center at Fairy Lake Botanical Garden. This work is part of the 10KP project (https://db.cngb.org/10kp/) and was also supported by China National GeneBank (CNGB; https://www.cngb.org/).

Author contributions

S.Z., H.L., X.G. and Y.L. led and managed the project. S.Z., H.L. and Yang Liu conceived the study. Yang Liu, S.W., L.L., S.D., T.W., J.M. and S. Wu wrote the manuscript. S.D., Y.G., X.F., A.J.L., Y.Y., X.G., D.L., N.L., H.W. and L.Y. prepared materials. S.W., L.L., T.Y., Yang Liu, J.R., J.W., S. Zaman, J.-Y.X., L.Z., J.C., Z.-Q.S., C.S., S.H., Na Li, M.L., G.F., H. Wang, J.Y., M. Lisby, S.K.S., W.M., Y.F., Y.C. and Z.Z. performed bioinformatics analysis. J.H., J.M., G.C. and P.L. performed horizontal gene transfer analysis. T.W., S.L., X.W. and X.L. performed SDR analysis. S.D., Yang Liu, Y.G., J.L., Y.Y. and Jianquan Liu performed gene family clustering and comparative phylogenomics. S. Wu, Y.V.d.P., Y.J., Z.-J.L. and Z.L. performed WGD analysis. P.S.S., Y.V.d.P., D.E.S., B.G., X.-Q.W., J.H., E.C.S., E.W. and M. Lisby contributed substantially to revisions. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41477-022-01129-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-022-01129-7.

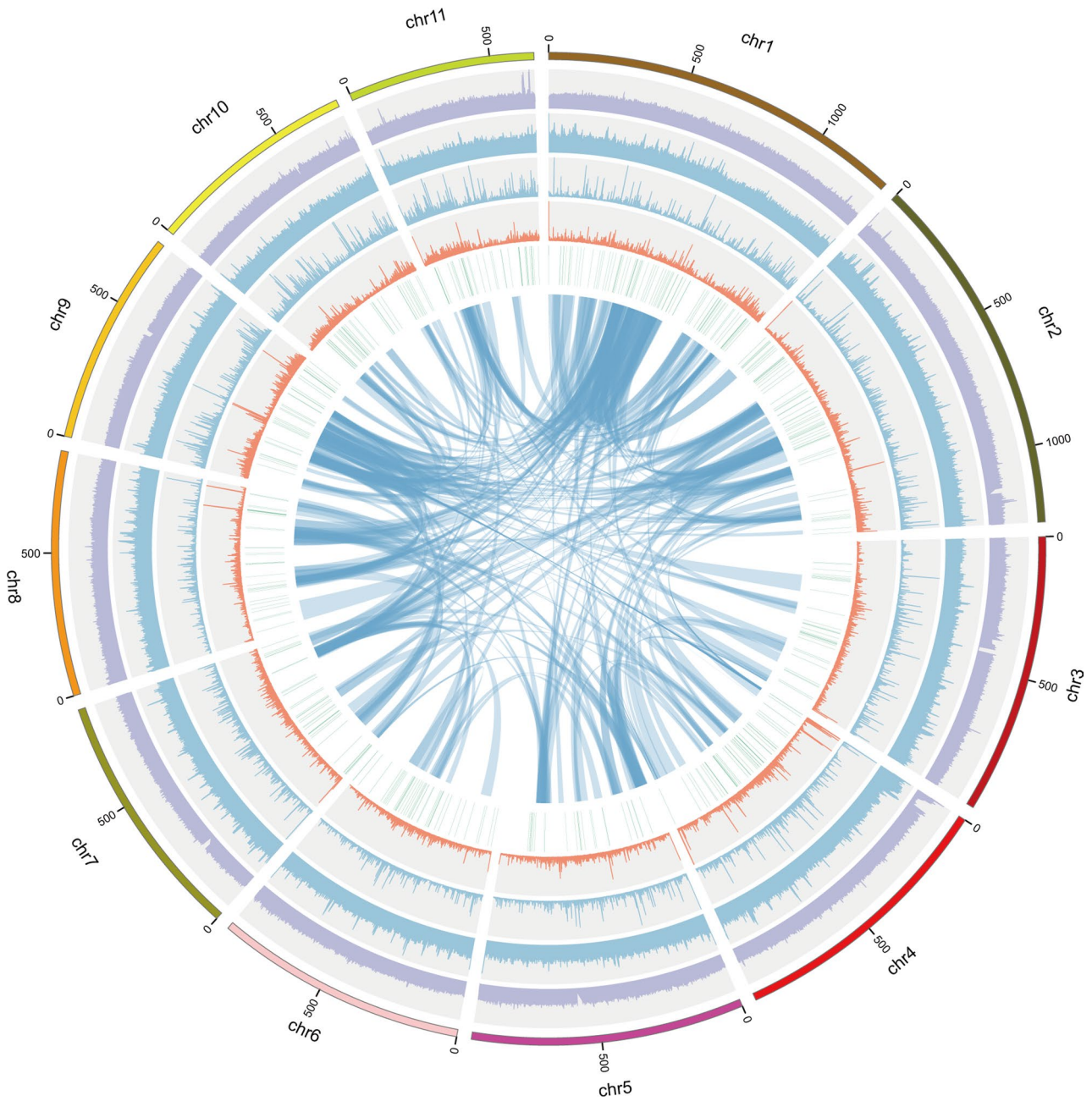
Correspondence and requests for materials should be addressed to Yang Liu, Yves Van de Peer, Douglas E. Soltis, Xun Gong, Huan Liu or Shouzhou Zhang.

Peer review information *Nature Plants* thanks James Clugston and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

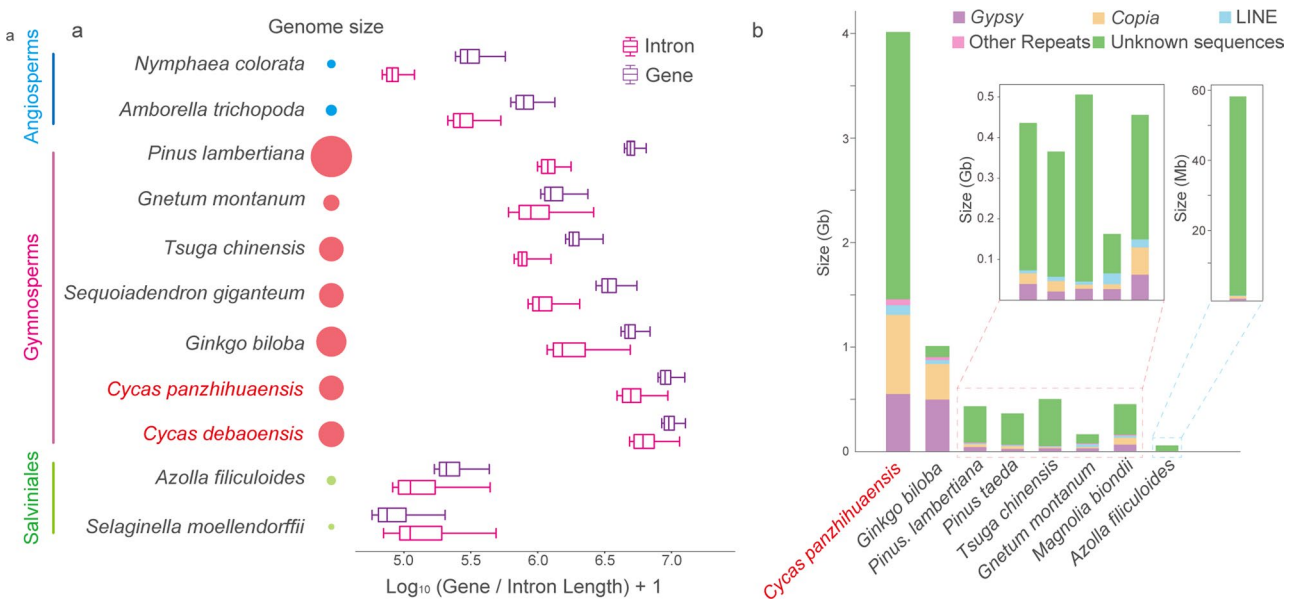
Reprints and permissions information is available at www.nature.com/reprints.

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

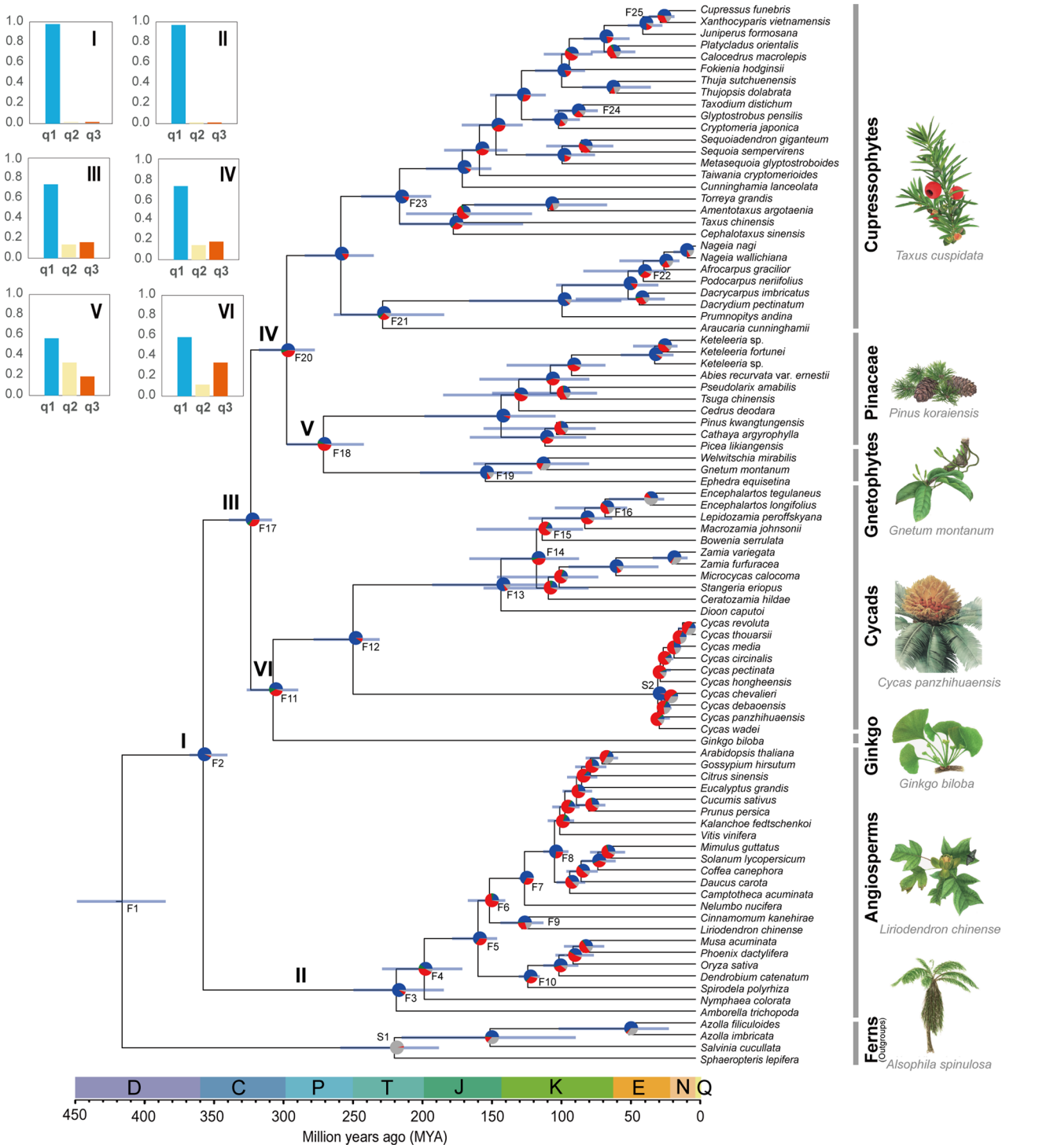
[1]State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen, China. [2]Key Laboratory of Southern Subtropical Plant Diversity, Fairy Lake Botanical Garden, Shenzhen & Chinese Academy of Sciences, Shenzhen, China. [3]State Key Laboratory of Grassland Agro-Ecosystems, College of Ecology, Lanzhou University, Lanzhou, China. [4]State Environmental Protection Key Laboratory of Regional Eco-process and Function Assessment, Chinese Research Academy of Environmental Sciences, Beijing, China. [5]Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China. [6]Key Laboratory of Plant Stress Biology, State Key Laboratory of Crop Stress Adaptation and Improvement, Henan University, Kaifeng, China. [7]Department of Biology, East Carolina University, Greenville, NC, USA. [8]College of Biology and Environment, Nanjing Forestry University, Nanjing, China. [9]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China. [10]Nanning Botanical Garden, Nanning, China. [11]School of Life Sciences, Sun Yat-sen University, Guangzhou, China. [12]Sichuan Cycas panzhihuaensis National Nature Reserve, Panzhihua, China. [13]Global Biodiversity Conservancy, Chonburi, Thailand. [14]Department of Entomology, China Agricultural University, Beijing, China. [15]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA. [16]Guangdong Provincial Key Laboratory for Plant Epigenetics, Longhua Institute of Innovative Biotechnology, College of Life Sciences and Oceanography, Shenzhen University, Shenzhen, China. [17]Department of Biotechnology and Biomedicine, Technical University of Denmark, Lyngby, Denmark. [18]Shenzhen Agricultural Genome Research Institute, Chinese Academy of Agricultural Sciences, Shenzhen, China. [19]College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing, China. [20]State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing, China. [21]Chengdu University of Traditional Chinese Medicine, Chengdu, China. [22]Department of Plant Biotechnology and Bioinformatics, Ghent University, VIB UGent Center for Plant Systems Biology, Gent, Belgium. [23]College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, China. [24]Hainan Institute of Zhejiang University, Sanya, China. [25]The College of Life Sciences, Sichuan University, Chengdu, China. [26]Key Laboratory of Orchid Conservation and Utilization of National Forestry and Grassland Administration at College of Landscape Architecture, Fujian Agriculture and Forestry University, Fuzhou, China. [27]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China. [28]College of Life Sciences, South China Agricultural University, Guangzhou, China. [29]National Key Laboratory of Plant Molecular Genetics, Chinese Academy of Sciences Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China. [30]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [31]Florida Museum of Natural History, University of Florida, Gainesville, FL, USA. [32]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa. [33]Department of Biology, University of Florida, Gainesville, FL, USA. [34]These authors contributed equally: Yang Liu, Sibo Wang, Linzhou Li, Ting Yang, Shanshan Dong, Tong Wei, Shengdan Wu, Yongbo Liu. ✉e-mail: yang.liu0508@gmail.com; yvpee@psb.vib-ugent.be; dsoltis@ufl.edu; gongxun@mail.kib.ac.cn; liuhuan@genomics.cn; shouzhouz@szbg.ac.cn



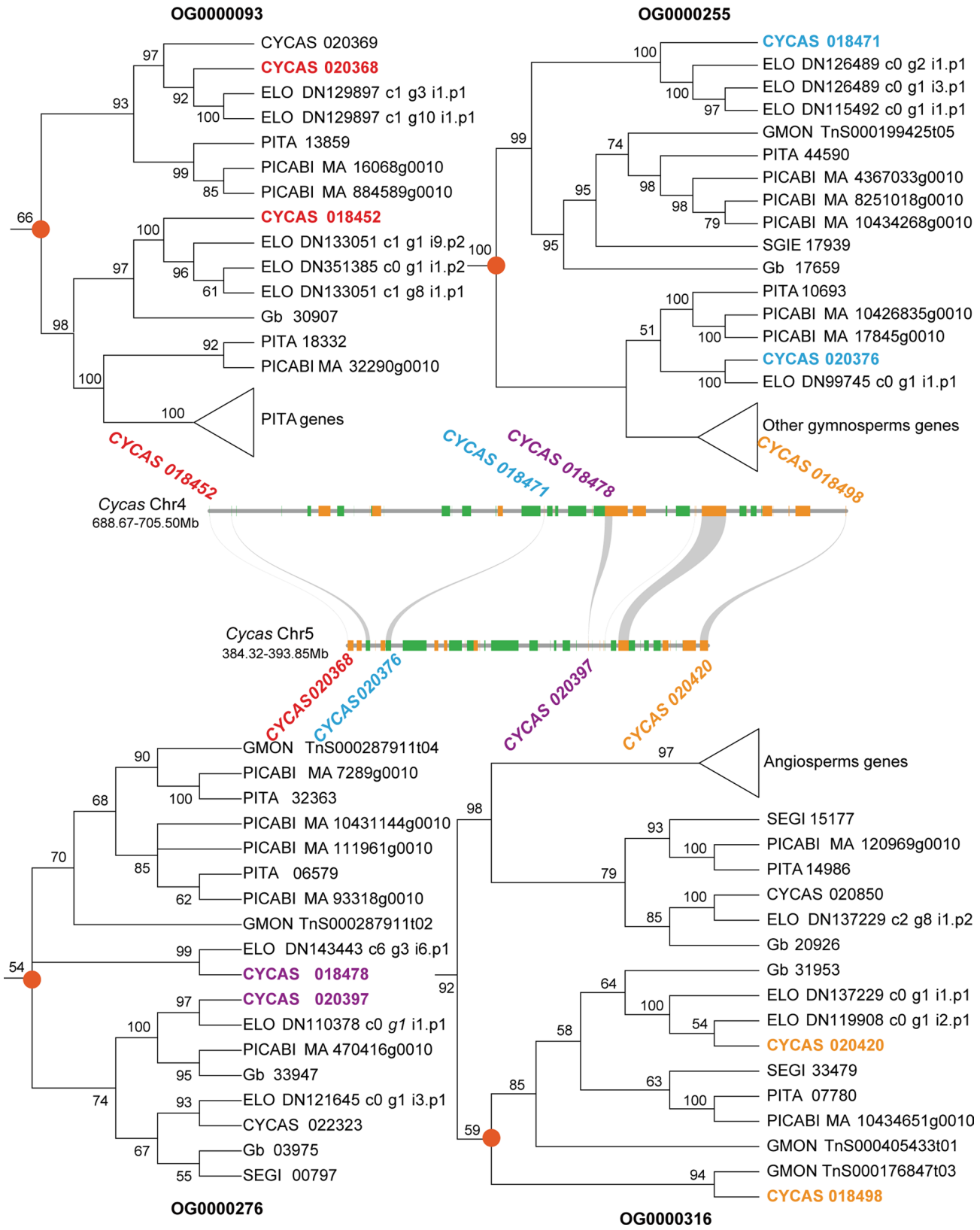
Extended Data Fig. 1 | Genome features of *C. panzhihuaensis*. Outer ring: The 11 chromosomes are labeled from Chr1 to Chr11. Inner rings 1-4 (from outside to inside): Repeat elements number shown in light purple. GC content colored indicated in light blue (y-axis min-max: 0.27–0.48). Expressed base percentage colored in light blue (y-axis min-max: 0–0.20). Gene numbers colored in light orange (y-axis min-max: 0-30). The sliding window of the inner rings 1-4 is 1 Mb. The inner ring 5 indicates the miRNA location over the genome. The blue lines inside represent the syntenic regions in *Cycas*.



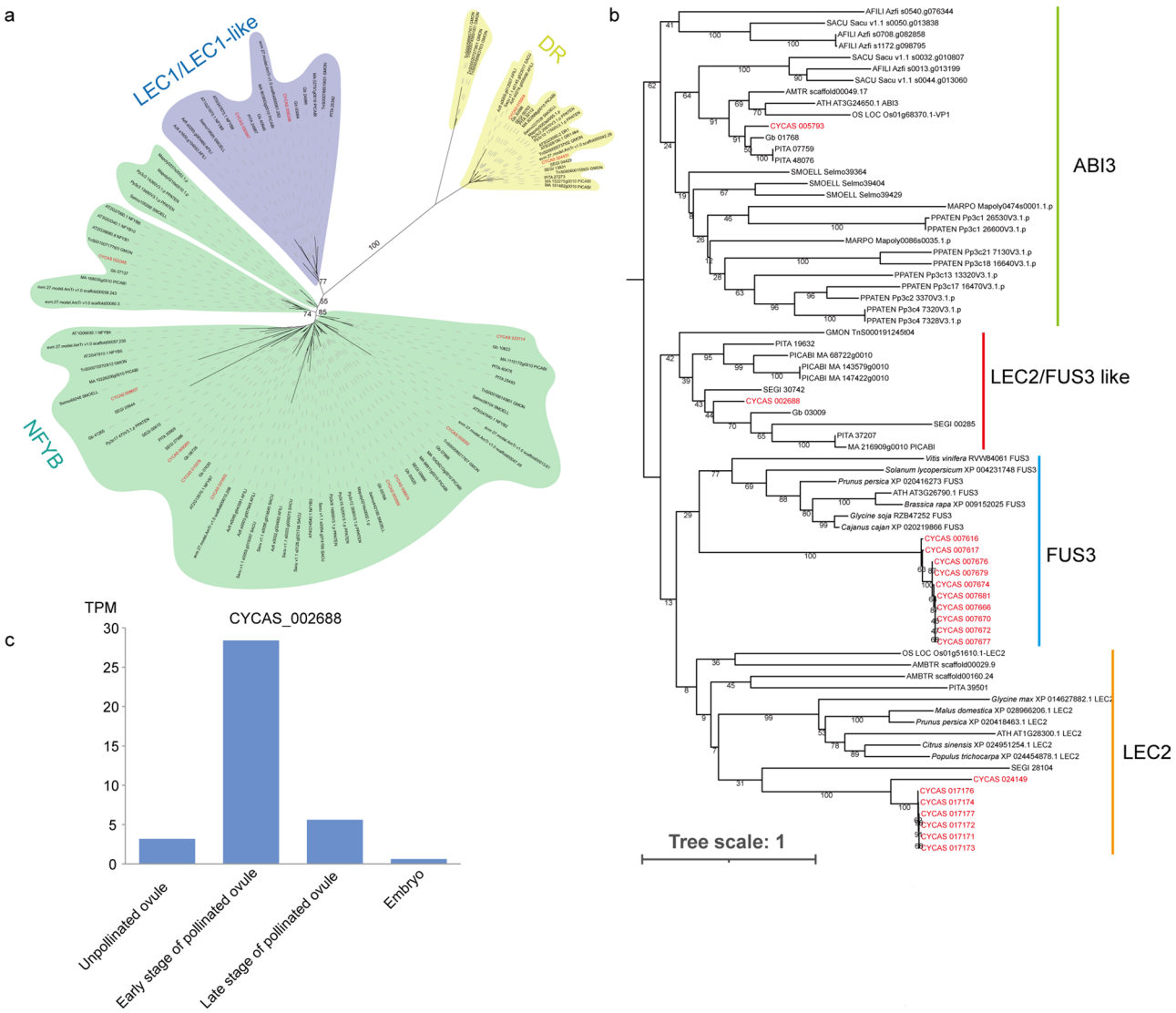
Extended Data Fig. 2 | Comparative analysis of *C. panzhihuaensis*. Extended Data Fig. 2. Comparative analysis of C. panzhihuaensis. (**a**) Comparison of the longest 10% of introns and gene in the representative land plants. The minimum, first quartile (Q1), median, third quartile (Q3), and maximum value was indicated in the box-plot by order after excluding the outliers. (**b**) Comparison of components of intron across the selected plants.



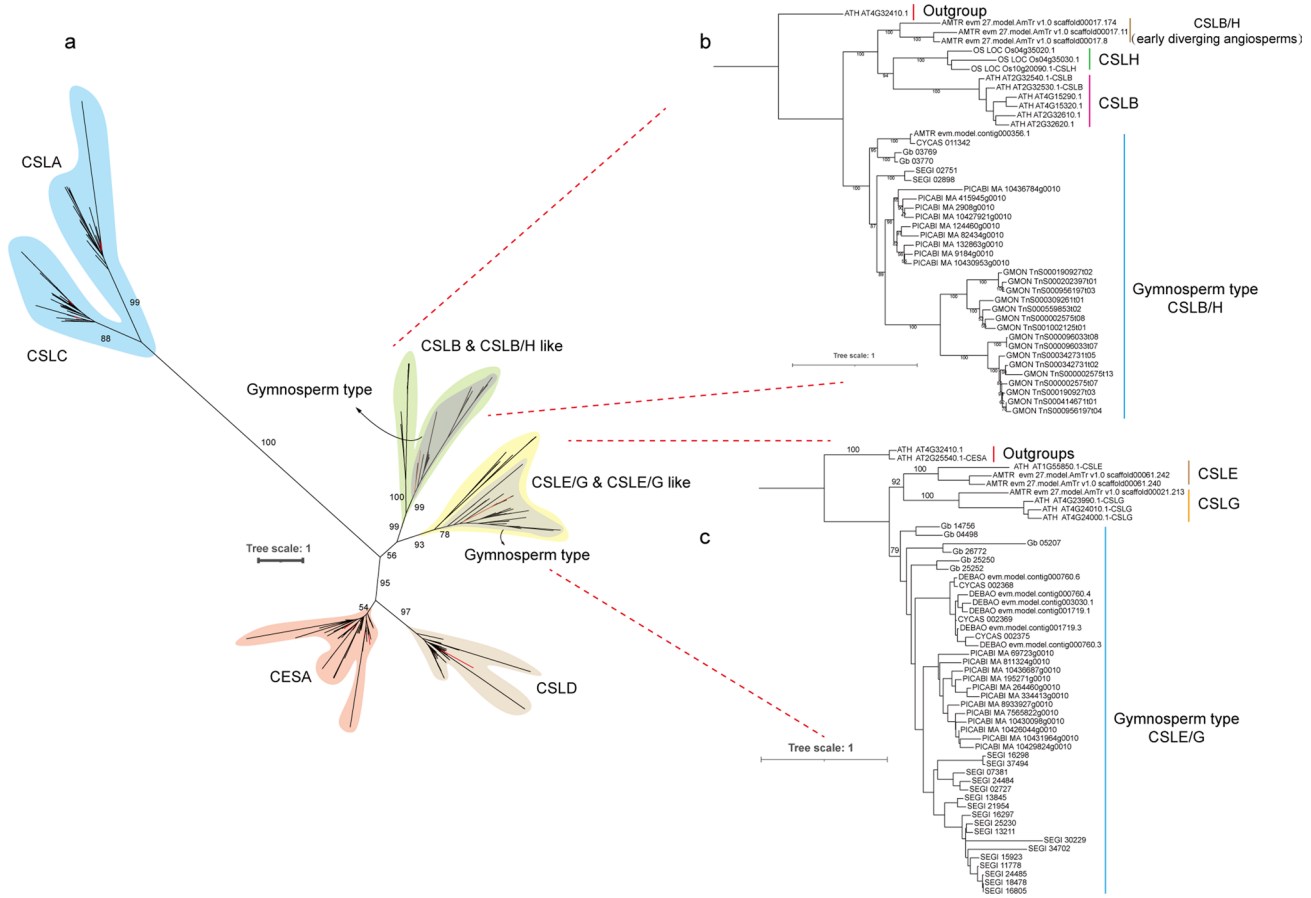
Extended Data Fig. 3 | The chronogram of 90 vascular plant species inferred with MCMCTree based on 100. nuclear single copy genes with concordant evolutionary histories. 25 fossil calibrations and 2 secondary calibrations were used. Individual gene trees (1,569 NT tree) were mapped on the nuclear coalescent tree with Phyparts. The pie charts at each node show the proportion of genes in concordance (blue), conflict (green = a single dominant alternative; red = all other conflicting trees), and without enough information (gray). Quartet support for six internal branches I, II, III, IV, V, VI were indicated on the left panel as barcharts. Image courtesy of Zanqian Li and Xiaolian Zeng.



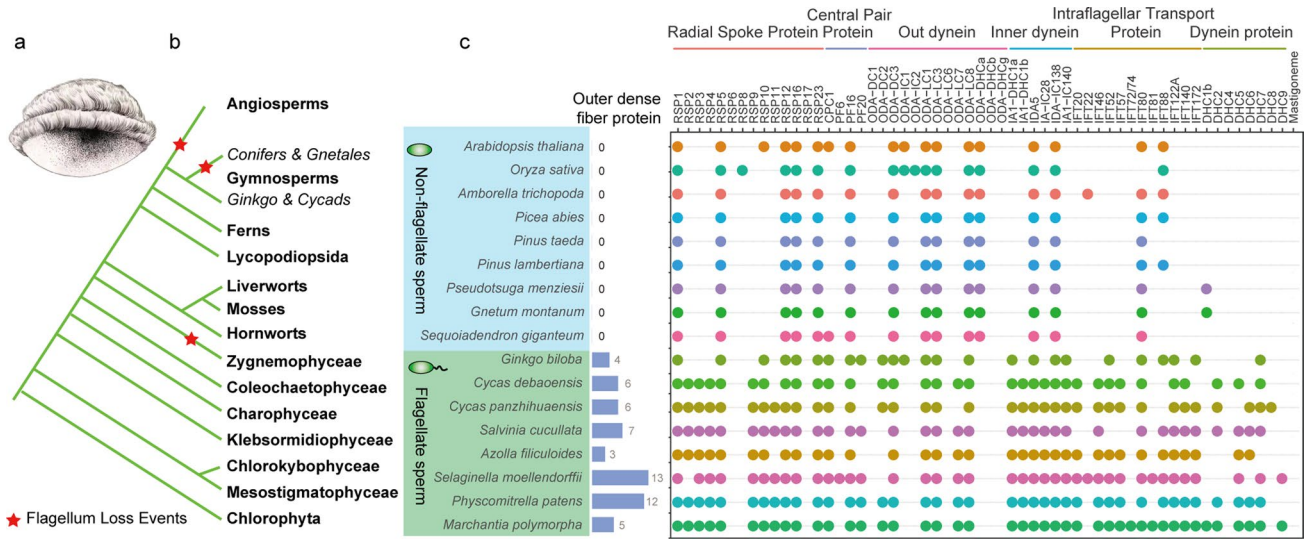
Extended Data Fig. 4 | Ancestral polyploidy events in extant gymnosperms. Example showing both the phylogenomic and syntenic evidence supporting an ancestral polyploidy event in extant gymnosperms. Four pairs of paralogous genes in OG0000093, OG0000255, OG00000276 and OG0000316 were duplicated before the divergence of gymnosperms and after the split of angiosperms and gymnosperms based on phylogenetic trees. These pairs of duplicated genes are located on the same syntenic block identified in the *C. panzhihuaensis* genome. The abbreviated name given before the protein ID represents species name: CYCAS: *Cycas panzhihuaensis*, Gb: *Ginkgo biloba*, ELO: *Encephalartos longifolius*, SEGI: *Sequoiadendron giganteum*, GMON: *Gnetum montanum*, PICABI: *Picea abies*, PITA: *Pinus taeda*.



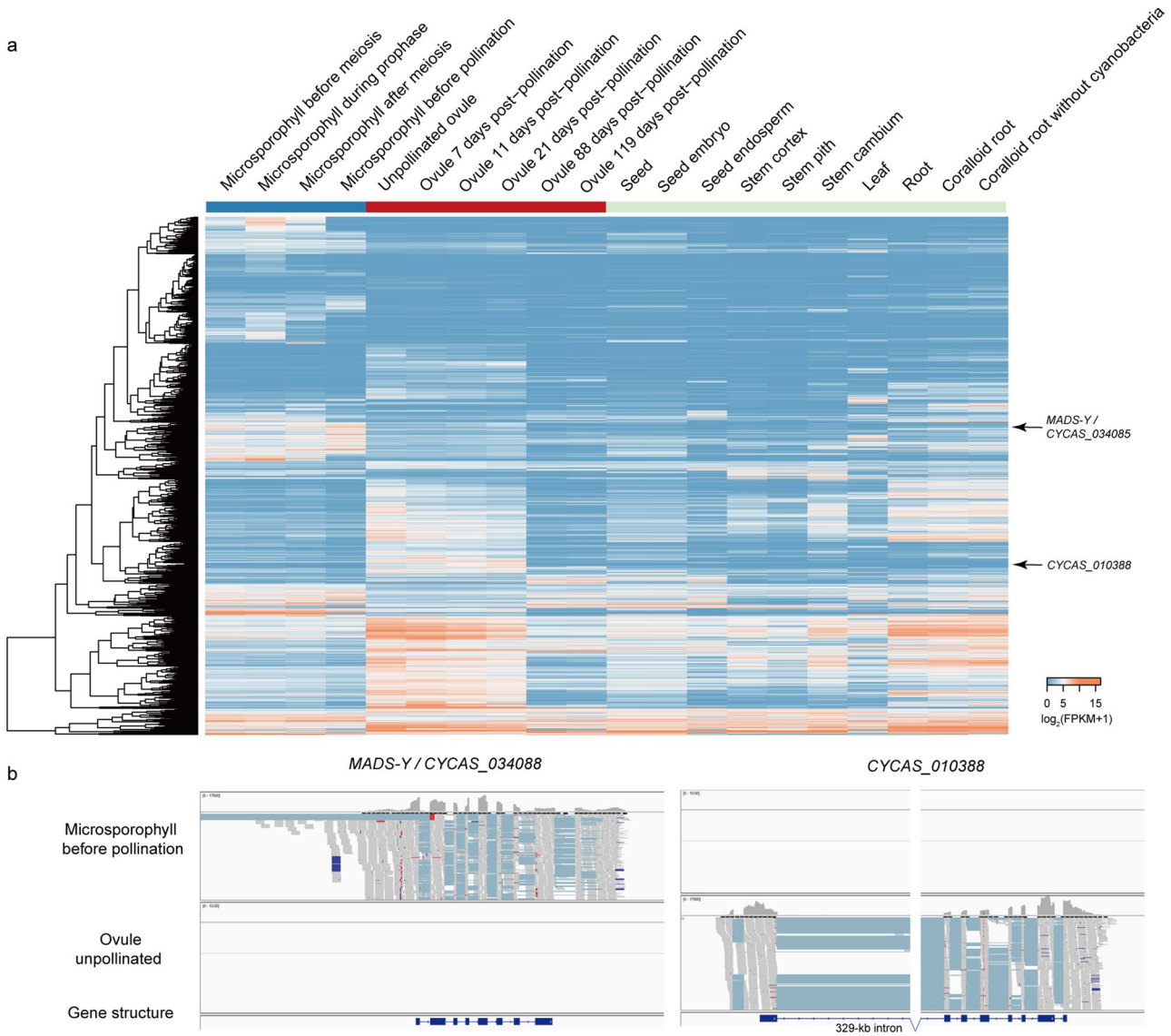
Extended Data Fig. 5 | The phylogeny of LAFL(NF-YB, ABI3, FUS3, and LEC2) transcriptional regulators. (**a**) Phylogenetic tree of the NF-YB. The tree was constructed using the maximum likelihood method with 500 bootstrap replicates. The bootstrap values are shown on the branches. (**b**) Phylogenetic tree of the B3 domain containing the gene family of *C. panzhihuaensis*. Bootstrap values are shown on the braches. (**c**) Transcript expression level is indicated by TPM during seed development. The phylogenetic trees were built using RAxML (estimating branch support values by bootstrap iterations with 500 replicates) with PROTGAMMAGTRX amino acid substitution model. The abbreviated name given before the protein ID represents species name: CYCAS: *Cycas panzhihuaensis*, Gb: *Ginkgo biloba*, SEGI: *Sequoiadendron giganteum*, GMON: *Gnetum montanum*, PICABI: *Picea abies*, PITA: *Pinus taeda*, ATH, *Arabidopsis thaliana*, DEBAO: *Cycas debaoensis*, AMTR: *Amborella trichopoda*, OS: *Oryza sativa*, AFILI: *Azolla filiculoides*, SACU: *Salvinia cucullata*, SELMO: *Selaginella moellendorffii*, PPATEH: *Physcomitrella patens*, MARPO: *Marchantia polymorpha*.



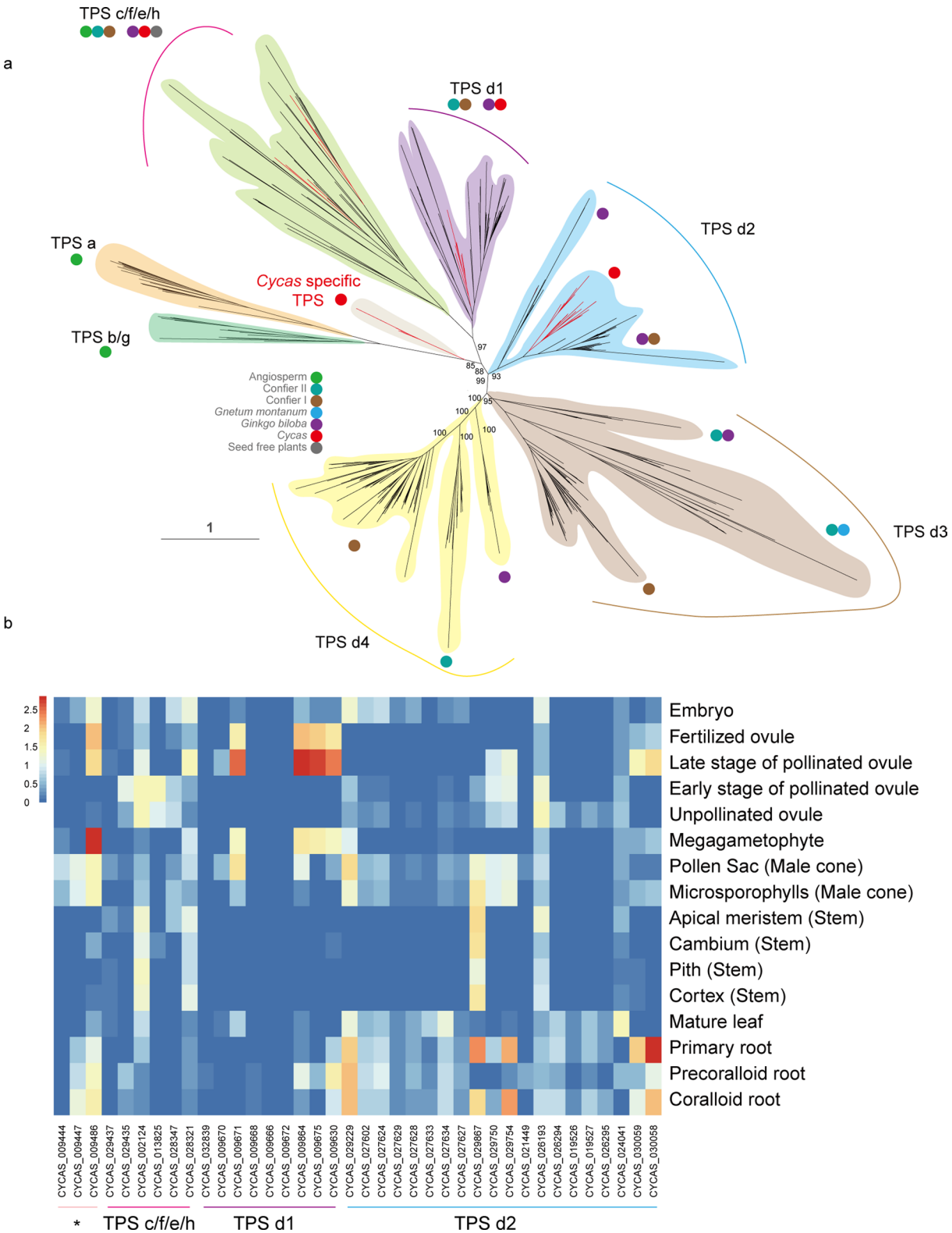
Extended Data Fig. 6 | Phylogenetic tree of CESA/CSL gene families. (**a**) Phylogenetic trees of CESA and CSL gene families. (**b**) Phylogenetic tree of CSLB and CSLH genes. (**c**) The phylogenetic tree of CSLE and CSLG genes. The CSLE/G from gymnosperm are the ancestral form of the angiosperm CSLE and CSLG. The phylogenetic trees were generated using RAxML with PROTCATGTR model and 500 bootstrap replicates. Bootstrap values ≥ 50% are shown. The abbreviated name given before the protein ID represents species name: CYCAS: *Cycas panzhihuaensis*, Gb: *Ginkgo biloba*, SEGI: *Sequoiadendron giganteum*, GMON: *Gnetum montanum*, PICABI: *Picea abies*, PITA: *Pinus taeda*, ATH, *Arabidopsis thaliana*, DEBAO: *Cycas debaoensis*, AMTR: *Amborella trichopoda*, OS: *Oryza sativa*.



Extended Data Fig. 7 | The Evolution of flagella related genes in embrophyta. (**a**) Sketch of the Cycas sperm. (**b**) Schematic diagram of flagellum loss events in green linage. (c) Distribution of outer dense fiber protein and other key flagellar proteins across representative embrophyta.



Extended Data Fig. 8 | The phylogeny and expression level of TPS. (**a**) Phylogenetic tree of the TPS gene family. The tree was constructed using RAxML (the maximum-likelihood method) with PROTCATGTR amino acid substitution model and 500 bootstrap replicates. The bootstrap values ≥ 50% are shown in the central branches. The red colors in the tree represent the cycas genes. (**b**) Heatmap of TPS gene family in different tissues of *C. panzhihuaensis*. The * denotes the *C. panzhihuaensis* specific TPS genes.



Extended Data Fig. 9 | Two MADS-box transcription factor genes differentially expressed in reproductive organs of *C. panzhihuaensis*. (**a**) Heatmap of 1,971 genes differentially expressed in males and females' organs. Arrows indicate CYCAS_034085 on the MSY and CYCAS_010388 on chromosome 2. (**b**) Expression of CYCAS_034085 on MSY and CYCAS_010388 on chromosome 2 in male microsporophyll and in the ovule.

nature portfolio

Corresponding author(s): Shouzhou Zhang

Last updated by author(s): Feb 22, 2022

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the 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
☒	☐ A statement on 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 statistical parameters 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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The long read data were generated by the Nanopore PromethION sequencer, and the short read data were sequenced by the MGI-SEQ or Illumina HiSeq X platform, with 150-bp read length and 300-500 insert size.
Data analysis	The softwares used in this study are listed as follows: wtdbg2 version2.5 (https://github.com/ruanjue/wtdbg2) miniasm version 0.3 (https://github.com/lh3/miniasm) NextDenovo version 2.3.0 (https://github.com/Nextomics/NextDenovo) NextPolish version 1.2.4 (https://github.com/Nextomics/NextPolish) 3d-dna version 201008 (https://github.com/aidenlab/3d-dna) Juicebox version 1.13.01 (https://github.com/aidenlab/Juicebox) MITE-hunter version 11-2011 (http://target.iplantcollaborative.org/mite_hunter.html) LTRharvest version 1.5.10 (http://genometools.org/) BUSCO version 4.0.2 (https://busco.ezlab.org/) RepeatModeler version 2.0.1 (http://repeatmasker.org/RepeatModeler/) RepeatMasker version 4.0.7 (https://www.repeatmasker.org/) LTRdigest version 1.5.10 (http://genometools.org/) Genewise version 2 (https://www.ebi.ac.uk/Tools/psa/genewise/) Hisat version 2.2.1 (http://www.ccb.jhu.edu/software/hisat/index.shtml) Stringtie version 1.3.3b (https://ccb.jhu.edu/software/stringtie/) AUGUSTUS version 3.2.3 (https://bioinf.uni-greifswald.de/augustus/) EVidenceModeler version 1.1.1 (https://evidencemodeler.github.io/) InterProScan version 5.30-69.0 (https://www.ebi.ac.uk/interpro/search/sequence/) miRDeep2 version 0.1.3 (https://github.com/rajewsky-lab/mirdeep2)

TargetFinder2 (https://github.com/carringtonlab/TargetFinder)
IQ-TREE2 version 2.0.6 (http://www.iqtree.org/)
ASTRAL version 5.7.3 (https://github.com/smirarab/ASTRAL)
OrthoFinder version 2.3.11 (https://github.com/davidemms/OrthoFinder)
KinFin version 1.0.3 (https://github.com/DRL/kinfin)
TranslatorX version local (https://translatorx.org/)
STAG version 1.0.0 (https://github.com/bbenligiray/stag)
PHYPARTS version 0.0.1 (https://bitbucket.org/blackrim/phyparts/src/master/)
DISCOVISTA version 1.0 (https://github.com/esayyari/DiscoVista)
HybPiper version 1.3.1 (https://github.com/mossmatters/HybPiper)
NOVOPlasty version 4.3.1 (https://github.com/ndierckx/NOVOPlasty)
PHYLONET version 2.4 (https://bioinfocs.rice.edu/phylonet/index.html)
PHYBASE version 2.0 (https://github.com/lliu1871/phybase)
TWISST version 0.2 (https://github.com/simonhmartin/twisst)
TRINITY version 2.13.1 (https://github.com/trinityrnaseq/trinityrnaseq)
TRANSDECODER version 3.0.0 (https://github.com/TransDecoder/TransDecoder)
SORTADATE version 2018 (https://github.com/FePhyFoFum/SortaDate)
PAML version 4.9 (http://abacus.gene.ucl.ac.uk/software/paml.html)
MCMCTREE in PAML version 4.9 (http://abacus.gene.ucl.ac.uk/software/paml.html)
MCscanX (https://github.com/wyp1125/MCScanX)
MAFFT version 7 (https://mafft.cbrc.jp/alignment/software/linuxportable.html)
PAL2NAL version 14 (http://www.bork.embl.de/pal2nal/)
Partitionfinder version 2 (https://github.com/brettc/partitionfinder)
trimAl version 1.2 (https://github.com/inab/trimal)
RAxML version 8 (https://cme.h-its.org/exelixis/web/software/raxml/)
DupGen_finder version (https://github.com/qiao-xin/DupGen_finder)
SSK_finder version (https://github.com/BGI-Qingdao/SSK_finder)
Trimmonmatic version 0.38 (http://www.usadellab.org/cms/index.php?page=trimmomatic)
GVCFtyper version 201911 (https://www.sentieon.com/products/#dnaseq)
PLINK version 1.9 (https://github.com/chrchang/plink-ng/tree/master/2.0)
EMMAX version beta-07Mar2010 (https://github.com/topics/emmax)
GCTA version 1.91.4beta3 (https://gump.qimr.edu.au/gcta/)
VCFtools version 0.1.13 (https://vcftools.github.io/man_latest.html)
eXpress version 1.5.3 (https://github.com/adarob/eXpress)
DESeq2 version 1.34.0 (https://bioconductor.org/packages/release/bioc/html/DESeq2.html)
WGCNA (https://github.com/cran/WGCNA)
PHYLIP version 3.696 (https://csbf.stanford.edu/phylip/)
CLUMPP version 1.1.2 (http://rosenberglab.bioinformatics.med.umich.edu/clumpp.html)
FigTree version 1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/)
ADMIXTURE version 1.3.0 (https://dalexander.github.io/admixture/)

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The genome and transcriptome data, genome assemblies, and annotations can be found at https://db.cngb.org/codeplot/datasets/public_dataset?id=PwRftGHfPs5qG3gE. The raw genomic, transcriptomic and HiC data generated in this study were deposited in the NCBI Sequence Read Archive (SRA, BioProject PRJNA734434), and the CNGB data center (https://db.cngb.org/) under project number CNP0001756.
Some known functional protein databases we used: Uniprot database (version 2021_01), KEGG (version 93.0), NCBI NR (version 20201015), KOG (version 20090331).

Field-specific reporting

Please select the one below that is 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/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	For genome sequencing, for each experiment, megagametophytes from a individual of Cycas panzhihuaensis was used. For genome and transcriptome short-read sequencing for phylogeny studies, one sample from a same individual for each taxon was used.
Data exclusions	The sequence reads with low quality are more likely derived from sequencing errors, and were thus excluded. To reduce the effect of sequencing error on assembly, we performed the quality control of raw data using Trimmonmatic (v. 0.38).
Replication	Since this is a genome sequencing project, no replication was applied for our genome sequencing experiment. For RNA-seq in gene expression studies and metabolite measurements, three pr two biological replicates were applied.
Randomization	Since this is a genome sequencing project. The data were generated from a single individual, no randomizations were required.
Blinding	Since this is a genome sequencing project. The data were generated from a single individual, no blinding experiment was performed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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