ARTICLES
https://doi.org/10.1038/s41477-022-01226-7

nature plants

Check for updates

OPEN

Dynamic genome evolution in a model fern

D. Blaine Marchant[1,30] ✉, Guang Chen[2,3,30], Shengguan Cai[4,5,30], Fei Chen[6], Peter Schafran[7], Jerry Jenkins[8], Shengqiang Shu[9], Chris Plott[8], Jenell Webber[8], John T. Lovell[8,9], Guifen He[9], Laura Sandor[9], Melissa Williams[8], Shanmugam Rajasekar[10], Adam Healey[8], Kerrie Barry[9], Yinwen Zhang[11], Emily Sessa[12], Rijan R. Dhakal[13], Paul G. Wolf[13], Alex Harkess[8,14], Fay-Wei Li[7,15], Clemens Rössner[16], Annette Becker[16], Lydia Gramzow[17], Dawei Xue[6], Yuhuan Wu[6], Tao Tong[3], Yuanyuan Wang[18], Fei Dai[4], Shuijin Hua[19], Hua Wang[20], Shengchun Xu[4], Fei Xu[4], Honglang Duan[21], Günter Theißen[17], Michael R. McKain[22], Zheng Li[23], Michael T. W. McKibben[24], Michael S. Barker[24], Robert J. Schmitz[25], Dennis W. Stevenson[26], Cecilia Zumajo-Cardona[26], Barbara A. Ambrose[26], James H. Leebens-Mack[27] ✉, Jane Grimwood[8], Jeremy Schmutz[8,9], Pamela S. Soltis[28] ✉, Douglas E. Soltis[12,28] ✉ and Zhong-Hua Chen[5,29] ✉

The large size and complexity of most fern genomes have hampered efforts to elucidate fundamental aspects of fern biology and land plant evolution through genome-enabled research. Here we present a chromosomal genome assembly and associated methylome, transcriptome and metabolome analyses for the model fern species *Ceratopteris richardii*. The assembly reveals a history of remarkably dynamic genome evolution including rapid changes in genome content and structure following the most recent whole-genome duplication approximately 60 million years ago. These changes include massive gene loss, rampant tandem duplications and multiple horizontal gene transfers from bacteria, contributing to the diversification of defence-related gene families. The insertion of transposable elements into introns has led to the large size of the *Ceratopteris* genome and to exceptionally long genes relative to other plants. Gene family analyses indicate that genes directing seed development were co-opted from those controlling the development of fern sporangia, providing insights into seed plant evolution. Our findings and annotated genome assembly extend the utility of *Ceratopteris* as a model for investigating and teaching plant biology.

Ferns have shaped life on Earth since divergence from their common ancestor with seed plants over 360 million years ago (Ma)[1]. Ferns can be found across diverse ecosystems as colonizers[2], keystone species[3], invasives[4] and agricultural supplements[5], and with over 10,500 extant species, they are the second most species-rich clade of vascular plants behind angiosperms[6]. Accompanying enormous morphological and ecological diversity, ferns have evolved numerous adaptations for protection from environmental stresses[7]. Fern secondary metabolites and their associated genes provide valuable resources for bioremediation, agricultural applications and lifesaving drugs[8–10].

Ferns have notoriously immense genomes (average 1C, 12.3 billion bases (Gb); maximum 1C, 147 Gb) and very high chromosome numbers (average, 40.5; maximum, 720)[11], hypothesized to be the

[1]Department of Biology, Stanford University, Stanford, CA, USA. [2]Central Laboratory, State Key Laboratory for Managing Biotic and Chemical Threats to the Quality and Safety of Agro-products, Zhejiang Academy of Agricultural Sciences, Hangzhou, China. [3]College of Agriculture, Yangtze University, Jingzhou, China. [4]College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, China. [5]School of Science, Western Sydney University, Penrith, New South Wales, Australia. [6]College of Life and Environmental Sciences, Hangzhou Normal University, Hangzhou, China. [7]Boyce Thompson Institute, Ithaca, NY, USA. [8]Genome Sequencing Center, HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. [9]United States Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [10]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ, USA. [11]Institute of Bioinformatics, University of Georgia, Athens, GA, USA. [12]Department of Biology, University of Florida, Gainesville, FL, USA. [13]Department of Biological Sciences, University of Alabama in Huntsville, Huntsville, AL, USA. [14]Department of Crop, Soil, and Environmental Sciences, Auburn University, Auburn, AL, USA. [15]Plant Biology Section, Cornell University, Ithaca, NY, USA. [16]Justus-Liebig-University, Department of Biology and Chemistry, Institute of Botany, Gießen, Germany. [17]Matthias Schleiden Institute/Genetics, Friedrich Schiller University Jena, Jena, Germany. [18]Hubei Insect Resources Utilization and Sustainable Pest Management Key Laboratory, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China. [19]Institute of Crops and Nuclear Technology Utilization, Zhejiang Academy of Agricultural Sciences, Hangzhou, China. [20]State Key Laboratory for Managing Biotic and Chemical Threats to the Quality and Safety of Agro-products, Institute of Virology and Biotechnology, Zhejiang Academy of Agricultural Sciences, Hangzhou, China. [21]Institute for Forest Resources & Environment of Guizhou, Key Laboratory of Forest Cultivation in Plateau Mountain of Guizhou Province, College of Forestry, Guizhou University, Guiyang, China. [22]Department of Biological Sciences, University of Alabama, Tuscaloosa, AL, USA. [23]Department of Integrative Biology, University of Texas at Austin, Austin, TX, USA. [24]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA. [25]Department of Genetics, University of Georgia, Athens, GA, USA. [26]New York Botanical Garden, Bronx, NY, USA. [27]Department of Plant Biology, University of Georgia, Athens, GA, USA. [28]Florida Museum of Natural History, University of Florida, Gainesville, FL, USA. [29]Hawkesbury Institute for the Environment, Western Sydney University, Penrith, New South Wales, Australia. [30]These authors contributed equally: D. Blaine Marchant, Guang Chen, Shengguan Cai. ✉e-mail: danielm1@stanford.edu; jleebensmack@uga.edu; psoltis@flmnh.ufl.edu; dsoltis@ufl.edu; z.chen@westernsydney.edu.au

consequences of repeated rounds of whole-genome duplication (WGD)[12,13]. However, genetic and genomic signatures of rampant WGD in ferns have not been documented[14–16]. Unfurling the genetic complexities and processes that have shaped fern genomes will illuminate not only the evolutionary history of this phylogenetically pivotal plant clade, but also the evolution of genome features and gene function in seed plants.

Ceratopteris richardii (hereafter *Ceratopteris*) has long been a model for investigating and teaching plant biology (for example, C-Fern Curriculum)[17]. *Ceratopteris* is typical of most ferns in being homosporous (producing a single spore type with potentially bisexual gametophytes) and having a large genome with numerous chromosomes (1C = 9.6 Gb; $n = 39$) relative to most eukaryotes. By contrast, all seed plants (flowering plants and gymnosperms) are heterosporous (producing both male and female spores with unisexual gametophytes). *Ceratopteris* and other homosporous ferns as well as lycophytes also have independent, free-living haploid gametophytes and diploid sporophytes (Fig. 1a), unlike seed plants in which the gametophyte is dependent upon the dominant sporophyte. As such, plant research laboratories globally have incorporated *Ceratopteris* to investigate life history traits, reproductive biology, development, evolution, space biology and genome biology[18,19]. Heterosporous water ferns (Salviniales; <1% of all fern species), characterized by relatively small, compact genomes, are represented by two genome assemblies, *Azolla filiculoides* (1C = 0.75 Gb, $n = 22$) and *Salvinia cucullata* (1C = 0.26 Gb, $n = 9$). These two genomes serve as ideal heterosporous fern counterparts to *Ceratopteris*[20], but are not representative of the vast majority of ferns.

Here we present the chromosome-level genome assembly and associated genetic resources for *Ceratopteris*. We investigated the composition and evolution of the large genome typical of ferns, analysed DNA methylation, documented horizontal gene transfer (HGT) events, investigated the evolution of gene families essential to flower and seed development, and characterized genes of potential economic, medicinal and environmental importance. The reference genome assembly, annotation and associated datasets extend the utility of *Ceratopteris* as a foundational model species for integration of comparative genomics into plant science research and education.

Results and discussion

The impact of transposons on genome size and intron length. We sequenced and assembled 7.46 Gb of the *Ceratopteris richardii* genotype Hn-n genome (https://phytozome-next.jgi.doe.gov/info/Crichardii_v2_1) (Fig. 1b). The *k*-mer analyses yielded a genome size estimate of 9.6 Gb, 15% smaller than previous estimates by flow cytometry[15] but within the error range of such estimates for large genomes[21,22]. The assembly contains 10,785 contigs with a contig *N50* of 2.3 Mb and scaffold *N50* of 182 Mb, with 93.5% of the assembled sequence contained in the 39 *Ceratopteris* chromosomes (Table 1). This is one of the largest haploid genomes with a chromosomal assembly to date, surpassed only by the assembly of the giant sequoia (*Sequoiadendron giganteum*) genome, which totals 8.125 Gb in 11 chromosomes but with a contig *N50* of 348 kb and scaffold *N50* of 690 Mb[23].

Transposable elements vary wildly in number and proportion of the genome among major lineages of life, among related species and even among populations[24]. Long terminal repeat (LTR) retrotransposons (Class I RNA transposable elements), which copy and paste throughout the genome when active, are often the dominant group of repeat elements in plants[25]. Although the transposition of LTR retrotransposons into genes or regulatory regions can disrupt gene function, they can alternatively generate novelty at the genetic, regulatory, expression or isoform levels, providing genetic material for evolution and adaptation[26].

We found over seven million repetitive elements that account for 85.2% of the *Ceratopteris* assembly (Supplementary Table 1). LTR retrotransposons represented the majority (67.0%) of the genome assembly, with the Ty3 superfamily making up 23.8% of the genome and the Ty1 superfamily making up 28.2%. LTR retrotransposons within these superfamilies averaged 2,301 and 1,492 bp in length, respectively. The Class II DNA transposable elements composed only 6.9% of the genome, with the highest representation from the CMC-En/Spm family, whereas 6.3% of the assembled genome was made up of simple repeat elements (Fig. 1c).

Protein-coding regions were annotated using a combination of ab initio prediction and transcript evidence from isoform sequencing (Iso-Seq) and RNA sequencing (RNA-seq) derived from ten tissues and developmental stages of *Ceratopteris* (Fig. 1a). The annotation of the total genome assembly contains 36,857 protein-coding genes and 38,397 alternative transcripts with 33,567 (91%) protein-coding loci anchored to the 39 chromosomes (Supplementary Table 2). Of the 410 genes in the Viridiplantae (Odb10) Benchmarking Universal Single-Copy Orthologs (BUSCO; v.4.1.1) dataset, 94.8% were identified in the annotation of *Ceratopteris*[27]. The chromosome assembly has, on average, a gene density of 4.81 per Mb, gene length of 14,457 bp, exon length of 363 bp and intron length of 5,555 bp (Supplementary Table 2). Among the extreme outliers, we identified 706 genes in *Ceratopteris* over 100 kb in length. Remarkably, introns account for 30% of the *Ceratopteris* genome with 17,745 introns over 10 kb in length. Although exon length varied little among the major lineages of plants, intron length in *Ceratopteris* had the largest range, beyond even that of the 22-Gb genome of *Pinus taeda*[28] (Fig. 1d).

Analyses of compact flowering plant genomes, such as *Arabidopsis* and rice, with average intron lengths of 152 and 387 bp, respectively[29], document a positive correlation between intron length and gene expression[29,30]. However, we found no correlation between total gene length and expression in *Ceratopteris* ($r^2 = 0.00004$, $P > 0.05$; Fig. 1e). *Ceratopteris* may serve as a model for investigating functional aspects of intron length and content on gene expression and messenger RNA maturation.

WGD is masked by rapid genome evolution. Polyploidy has contributed to the complexity and gene content of all green plants[31,32]; however, its frequency in the evolutionary history of ferns has been contentious for decades[12,14,33,34]. The large size and numerous chromosomes of fern genomes have long been considered evidence that repeated polyploidy and subsequent gene loss/silencing have contributed to the diversification of molecular and ecological function across fern evolutionary history[34–37]. However, analyses of angiosperm genomes have demonstrated that chromosome number is a poor predictor of WGD frequency[31,38]. Surprisingly, the limited genetic and genomic data now available for ferns have pointed towards polyploidy being less frequent and genome content being less dynamic compared with angiosperms[14–16,39,40].

To clarify the impact of WGD on the evolutionary history of *Ceratopteris* and ferns more generally, we employed divergence-based, genomic and phylogenomic approaches. A single WGD event could be inferred from the paralogue synonymous substitution (K_s) distribution analysis of *Ceratopteris* with a K_s peak at 1.3 (Fig. 2a). Phylogenetic analyses using Multi-tAxon Paleopolyploidy Search (MAPS)[41] and NOTUNG[42,43] of more than 5,000 gene families, including protein sequences from *Ceratopteris* and other fern species, implicated two WGDs on the lineage leading to *Ceratopteris* within the last 300 million years (Myr) as inferred previously[31] (Fig. 2b). These analyses placed the most recent WGD (CERAα) after the divergence of *Ceratopteris* from its sister genus, *Acrostichum*, at just 62 Ma (ref. [44]; Fig. 2b and Extended Data Fig. 1), and queries of gene trees indicated that this was the only WGD event represented in the 1.3 peak observed in the K_s plot (Fig. 2a). Both phylogenetic analyses of fern gene sequences also supported a putative earlier WGD before divergence of the Polypodiales and Salviniales

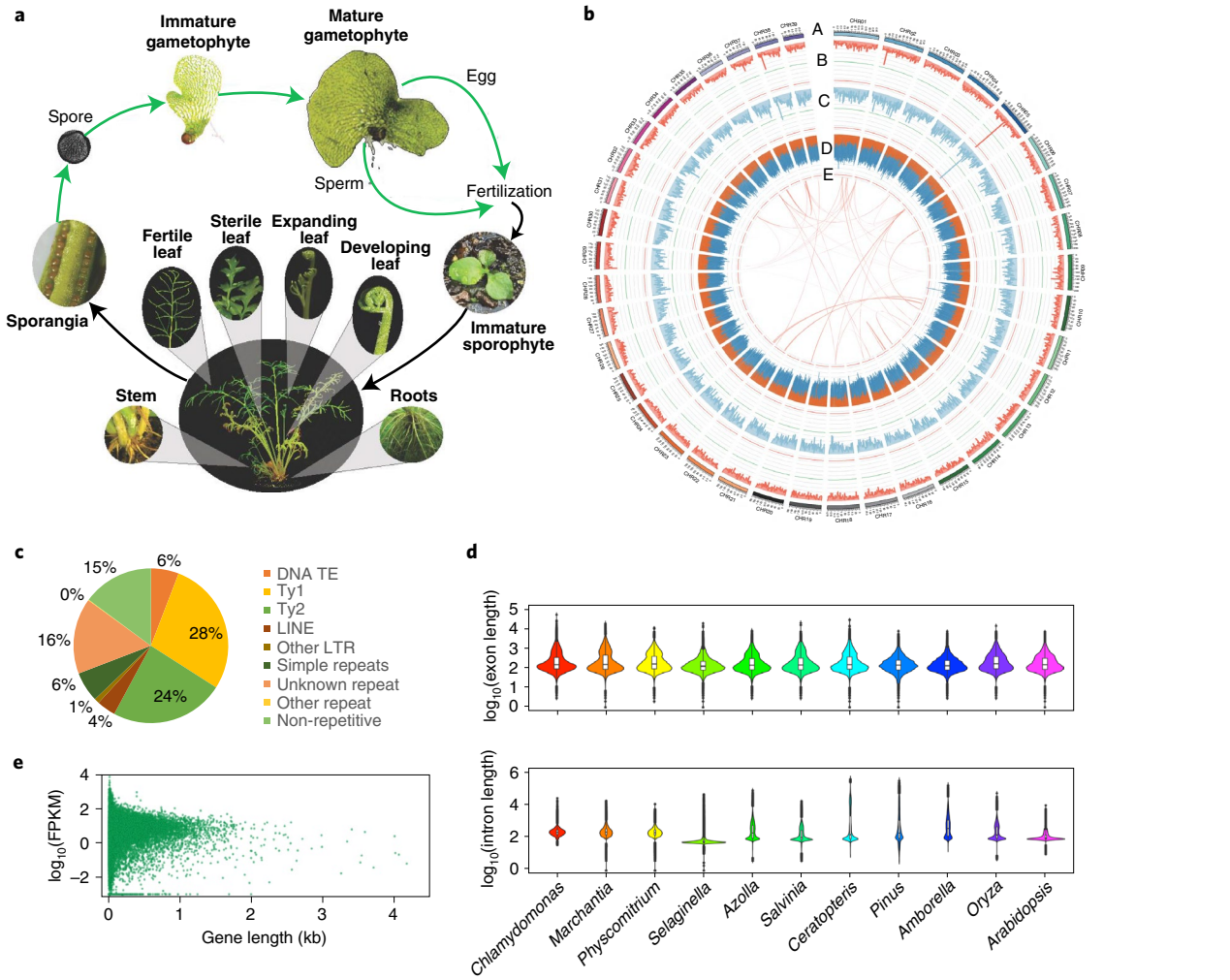


Fig. 1 | *Ceratopteris richardii* life cycle and genome assembly characteristics. a, Life cycle of *Ceratopteris* with tissues sampled for RNA-seq in bold. Images are not to scale. **b**, Genome assembly of *Ceratopteris* with: (A) chromosomes, (B) gene density in a 3-Mb sliding window, maximum value of 139; (C) mRNA expression density in a 3-Mb sliding window, maximum value of 170; (D) long terminal repeat retrotransposon density in a 3-Mb sliding window, orange and blue bands represent Ty3 and Ty1 LTRs, respectively, maximum value of 970; and (E) intragenomic syntenic regions of ten or more genes. Green horizontal lines represent the 5th percentile, red horizontal lines represent the 95th percentile. **c**, Genome composition of *Ceratopteris*. LTR, long terminal repeat; TE, transposable element. **d**, Intron and exon lengths from a green alga (n = 166,499 exons; 147,269 introns), liverwort (n = 137,019 exons; 112,345 introns), moss (n = 587,902 exons; 501,233 introns), lycophyte (n = 197,720 exons; 162,895 introns), two water ferns (*Azolla*: n = 127,875 exons; 107,674 introns; *Salvinia*: n = 122,980 exons; 103,200 introns), *Ceratopteris* (n = 437,785 exons; 400928 introns), gymnosperm (n = 268,745 exons; 187,344 introns), basal angiosperm (n = 111,241 exons; 83,928 introns), monocot (n = 196,916 exons; 152,273 introns) and eudicot (n = 313,952 exons; 237,746 introns). The widest part of the violin plot represents the highest point density, whereas the top and bottom are the maximum and minimum data respectively. Box plots are in the middle of violin plots, the top and bottom lines represent 25th and 75th percentiles, the centre line is the median and whiskers are the full data range. **e**, Correlation between fragments per kilobase million (FPKM) and gene length at genome-wide level of *Ceratopteris*.

Table 1 | Final summary statistics for chromosome-scale assembly

Genome assembly statistics	Number/size
Scaffold total	6,185
Contig total	10,785
Scaffold sequence total	7,463.3 Mb
Chromosome sequence	6,932.2 Mb
Contig sequence total	7,417.3 Mb (0.6% gap)
Scaffold *L/N50*	19/182.0 Mb
Contig *L/N50*	908/2.3 Mb

lineages 230 Ma (CYATγ), consistent with other fern WGD analyses (Fig. 2b)[45–47]. MAPS also inferred a third ancient WGD in the ancestry of the Polypodiales (PTERα)[31], but this was not observed in the NOTUNG analyses or another recent analysis using a different set of phylogenomic methods[47]. Additional genomes from the Polypodiales are needed to resolve the ultimate number and position of WGDs in this region of the fern phylogeny.

We expected to find numerous, large syntenic subgenome blocks of paralogous genes among the chromosomes of *Ceratopteris*, the typical genomic signature of WGD. However, only 45 syntenic blocks of ten or more genes, totalling 367 genes, were found within the 39 *Ceratopteris* chromosomes (Extended Data Fig. 2). For comparison, the paralogue K_s distribution for *Arabidopsis thaliana* (1C = 135 Mb, n = 5) exhibits median peak values of 0.7, 1.7

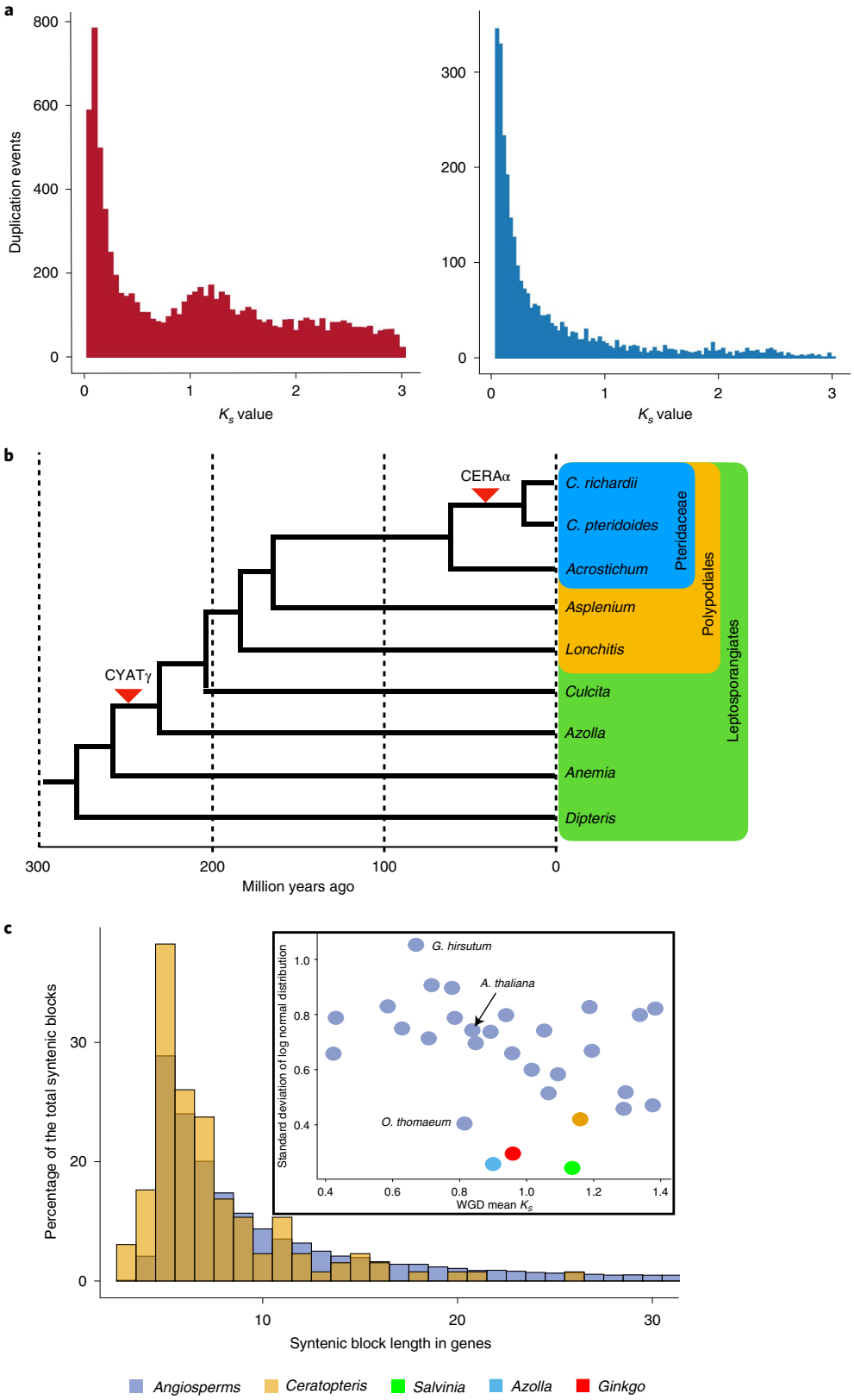


Fig. 2 | Evidence of polyploidy in the evolutionary history of *Ceratopteris*. **a**, K_s distributions of all paralogous genes (left) and tandemly duplicated paralogous genes (right). **b**, Placement of WGD events in the evolutionary history of *Ceratopteris* based on phylogenomic analyses. **c**, Proportion and length of syntenic regions in *Ceratopteris* (yellow) relative to the average of 27 flowering plant species (lilac). Inset shows the standard deviation of syntenic block length distribution relative to peak K_s value (WGD) for *Ceratopteris* (yellow), *Azolla* (blue), *Salvinia* (green), *Ginkgo* (red) and 27 flowering plant species (lilac).

and 2.7 for the *At-α*, *At-β* and *At-γ* paleopolyploidy events dated at 23.3 Ma (ref. [38]) to 50.1 Ma (ref. [48]), 61.2 Ma (ref. [48]) and >125 Ma (ref. [38]) respectively. Syntenic subgenome blocks are evident for all three events, although *At-γ* blocks are highly fragmented relative to syntenic segments of *Arabidopsis* subgenomes attributed to the *At-α* and *At-β* WGDs[49]. Similarly, three independent WGD events (ρ, σ and τ) can be discerned via phylogenetics and synteny in the lineage leading to rice (*Oryza sativa*; 1C = 430 Mb, $n = 12$)[31]. We further tested the scale of retained synteny across fern genomes by comparing the genome of *Ceratopteris* with that of the water fern *Salvinia cucullata*, which last shared a common ancestor 230 Ma (ref. [20]). Virtually no syntenic blocks were detected between the two fern species (Extended Data Fig. 3).

Diploidization, the process of returning a polyploid genome to a genetically diploid state via fractionation (loss) and silencing of WGD-derived genes, transposition events and genome rearrangements, can vary in rate among plant lineages[50–52]. To assess the degree of synteny in *Ceratopteris* relative to other land plant genomes, we compared syntenic block lengths in *Ceratopteris*, the two sequenced water ferns (*Azolla filiculoides* and *Salvinia cucullata*), a gymnosperm (*Ginkgo biloba*) and 27 angiosperms ranging in WGD history[53]. Only the grass *Oropetium thomaeum* had a smaller average length of syntenic blocks among the angiosperms sampled compared with *Ceratopteris* (Fig. 2c); however, it also has a small genome of 245 Mb and nine chromosomes. All four non-angiosperms had smaller syntenic blocks than the angiosperms other than *O. thomaeum*. These results suggest that although retention of synteny can be highly variable among land plants, the genomes of *Ceratopteris* and the other analysed non-angiosperms are much more fractionated relative to angiosperms. Tandem gene duplications have highly influenced the genome content and structure of *Ceratopteris* because recent tandem duplications account for a large proportion of the paralogue pairs included in the K_s distribution (>6,000 genes) (Fig. 2a). Taken together, we document rapid rates of genome evolution in *Ceratopteris* relative to those of angiosperms that serve to mask WGD events.

DNA methylation in *Ceratopteris*. Whole-genome bisulfite sequencing[54,55] enabled fine-scale resolution of DNA methylation in the *Ceratopteris* genome. CG and CHG methylation (H = A, C or T) were found throughout most genomic features; however, CHG methylation was especially enriched in repeats as well as the unusually large introns (Fig. 3a–e). Interestingly, CHH methylation initially appeared absent in the *Ceratopteris* genome because it could not be readily distinguished from background (Fig. 3a). In angiosperms, CHH methylation results from activities of the RNA-directed DNA methylation pathway and/or CHROMOMETHYLASE 2 (CMT2). In *Arabidopsis*, CMT2 has a preference for CWA (W = A or T) sites[56]. Closer examination of the CHH methylation results revealed CWA sites to be more highly methylated compared with other CHH contexts, albeit at low levels (Fig. 3d). Previous studies have shown that *CMT2* is not present in ferns because it evolved in the common ancestor of angiosperms[57]. However, there are two *CMT* genes present in *Ceratopteris*; at least one of these ancient CMTs presumably possesses the ability to methylate CWA sites. Our data suggest that CMT-associated CWA methylation is present in *Ceratopteris*, but the RNA-directed DNA methylation pathway is not active, consistent with its loss in certain fern species[58].

Curiously, gene body DNA methylation (gbM), which is common in angiosperms[59], is also present in *Ceratopteris* (Fig. 3e). gbM is associated with methylation of CG sites only and is present in genes that are often expressed constitutively, evolving slowly and possess 'housekeeping' functions[60]. Although it has been hypothesized that gbM could be present outside angiosperms[57,60], notably within certain gymnosperms and ferns, high coverage genome-wide data were lacking until now to confirm its presence. We therefore provide unambiguous documentation of gbM outside of angiosperms.

Gene family evolution across green plants. The identification of genes contributing to reproduction in homosporous ferns and their expression patterns can elucidate the evolution and potential origin of genes driving shifts in the reproductive biology of heterosporous ferns and seed plants[18]. Despite the considerable morphological and physiological differences between the *Ceratopteris* gametophyte and sporophyte (plus the respective haploidy versus diploidy of these alternating generations), only 273 and 1,397 genes were specifically expressed in the gametophyte and sporophyte, respectively (Fig. 4a). Similarly, 346 genes were solely expressed in meiotic tissues (fertile leaf and sporangia), whereas 1,270 genes were solely expressed in non-meiotic tissues and over 30,000 genes were expressed in both datasets (Fig. 4b). This low level of specificity supports recent work suggesting that leaf and seed developmental genes are co-opted from sporangia developmental networks[61–64].

To better understand the evolutionary transition from seedless plants to the production of seeds, flowers and fruits, we identified and analysed gene families in *Ceratopteris* known to be critical to flower induction in *Arabidopsis* and other angiosperms. The phosphatidyl-ethanolamine binding protein family is well conserved across green plants and animals controlling a variety of biological processes[65]. In angiosperms, the phosphatidyl-ethanolamine binding proteins FLOWERING LOCUS T (FT) and MOTHER OF FT (MFT) regulate flowering time and flower architecture[66]. We identified ten *FT* genes in *Ceratopteris*, compared with six in *Arabidopsis* and four in *Azolla filiculoides* (Fig. 4c). Of those ten, nine *Ceratopteris FT* homologues are present in subfamilies that are absent in flowering plants, whereas the one remaining, and most generally expressed, *Ceratopteris FT* gene was in the clade containing the *Arabidopsis* gene *AtMFT* (Fig. 4c). The three generally expressed *Ceratopteris FT* genes likely play major roles in the many phase changes of the fern life cycle. Interestingly, seven *Ceratopteris FT* homologues were highly expressed only in meiotic tissue (fertile leaf and sporangia), suggesting that these *FT* homologues may be associated with spore development in ferns, predating the function of regulating flowering in angiosperms (Fig. 4c).

The evolution of plant architecture. MADS-box genes have been identified in almost all eukaryotes, but have expanded most in green plants, where they are well known for their roles in numerous aspects of plant architecture and development[67]. More than 20 years ago, the first MADS-box genes were identified in *Ceratopteris*[68–70], but owing due to the lack of genomic data, the entire complement of MADS-box genes in a homosporous fern genome has been unclear until now. We identified 35 MADS-box genes in the *Ceratopteris* genome, classified into 8 Type I and 27 Type II MADS-box genes based on phylogeny reconstruction. Type II genes were further subdivided into MIKC^C- and MIKC*-group genes based on a separate phylogenetic analysis of Type II genes (Fig. 4d). MIKC^C-group genes are of special interest owing to their crucial importance for flower development and evolution[71]. Studies on *Ceratopteris* in the pre-genomics era had already identified three clades of fern-specific genes (*CRM1*-, *CRM3*- and *CRM6/CRM7*-like genes), with each clade containing several paralogues[68,72]. Twenty-one Type II genes belong to the clade of MIKC^C-group genes, and six are MIKC*-group genes. Surprisingly, recent analyses based on comparative transcriptomics additionally identified a large 'orphan' clade of previously unknown fern-specific MIKC^C-group genes for which no *Ceratopteris* representative was previously known[31]. Analysis of the *Ceratopteris* genome corroborates the view that there are no orphan clade members in this species. Because representatives exist in all major groups of ferns[31], these genes must have been established early in fern evolution and lost relatively recently in the lineage that led to *Ceratopteris*. Interestingly, even though MIKC*-group genes have more exons (9–12) than MIKC^C-group genes (6–8 exons), all of the genomic loci of MIKC*-group genes were smaller (<40 kb) than

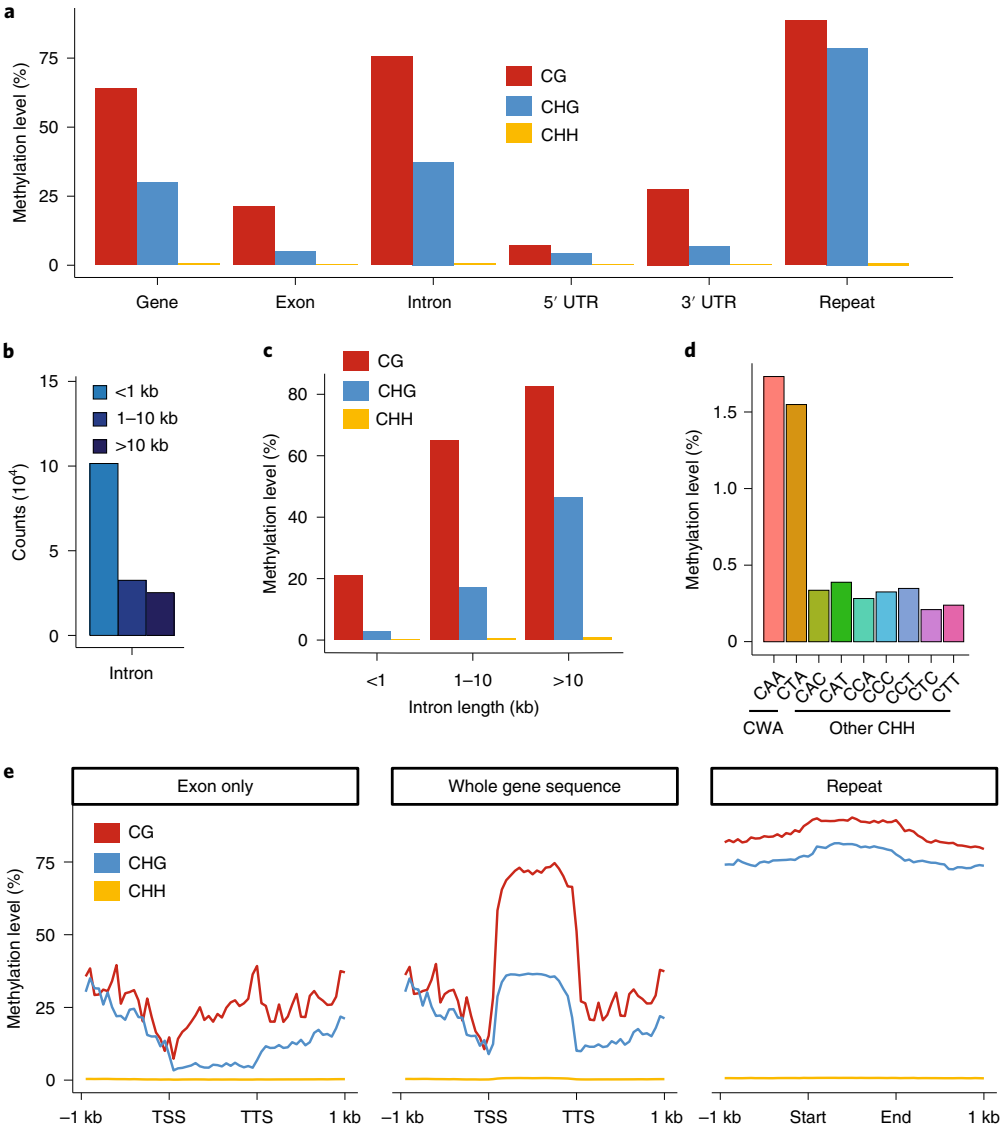


Fig. 3 | Whole-genome methylation profiling of *Ceratopteris*. **a**, CG, CHG and CHH methylation level in all genes (whole gene sequence) (n = 36,857), exons (n = 159,326), introns (n = 122,469), untranslated region (UTR) (n = 32,307) and randomly selected repeats (n = 1,000). **b**, Introns were classified into three groups based on length: shorter than 1 kb (n = 101,510); 1–10 kb (n = 32,547); and longer than 10 kb (n = 25,269). **c**, CG, CHG and CHH methylation level for introns of different lengths. **d**, Methylation level of nine CHH contexts in repeats. **e**, Metaplots show distribution of DNA methylation level across gene and repeat body (right). For genes, methylation level was calculated for C-contexts in exons (left) and in whole gene sequences, including exons and introns (middle). TSS, transcription start site. TTS, transcription termination site.

all of the genomic loci of MIKC^C-group genes. Nine MIKC^C-type genes are encoded by genomic loci spanning 100,000 to 216,247 bp. The first intron, known to include regulatory elements in other plants[73–76], is often the largest intron of *Ceratopteris* MADS-box genes. Notably, we also found two MIKC^C-group MADS-box genes each encompassing two alternative MADS boxes. For one of these genes, we found mRNAs including one or the other, but not both MADS boxes. For the other gene, unfortunately, there is not enough mRNA data to judge which mRNAs are formed. The generation of mRNAs from these loci with multiple MADS boxes potentially involves alternative promoters and differential splicing. A similar phenomenon has so far only been described for a number of MADS-box genes in Norway spruce, *Picea abies*[77].

HGT and the evolution of defence genes. Novel biopesticides have been discovered in a number of fern species and have benefited sustainable agriculture and food security[10,78]. For example, a gene encoding a novel insecticidal protein, *Tma12*, was identified in the fern *Tectaria macrodonta* and cloned into cotton to battle phytophagous whiteflies[10]. *Tma12* was also identified in the genome of *Salvinia cucullata* and in the transcriptomes of other ferns[20]. Phylogenetic placement of the fern *Tma12* genes among bacterial sequences suggests that the fern genes originated from HGT from bacteria to ferns[20] (Extended Data Fig. 4a). We identified two homologues of *Tma12* in *Ceratopteris*, expression of which differed dramatically between tissues and developmental stages, with sporangia showing over 3,000 times greater expression compared with the rest of the RNA-seq samples (Extended Data Fig. 4b). Similar to increased defensive metabolite production in unripe fruits[79], *Tma12* expression in *Ceratopteris* may be an adaptation that protects the sporangia from insect attack before spore dispersal.

We discovered a block of 36 recently tandemly duplicated aerolysin-like protein-coding genes on chromosome 9 of *Ceratopteris* (Fig. 5a). These genes are well-studied in bacteria

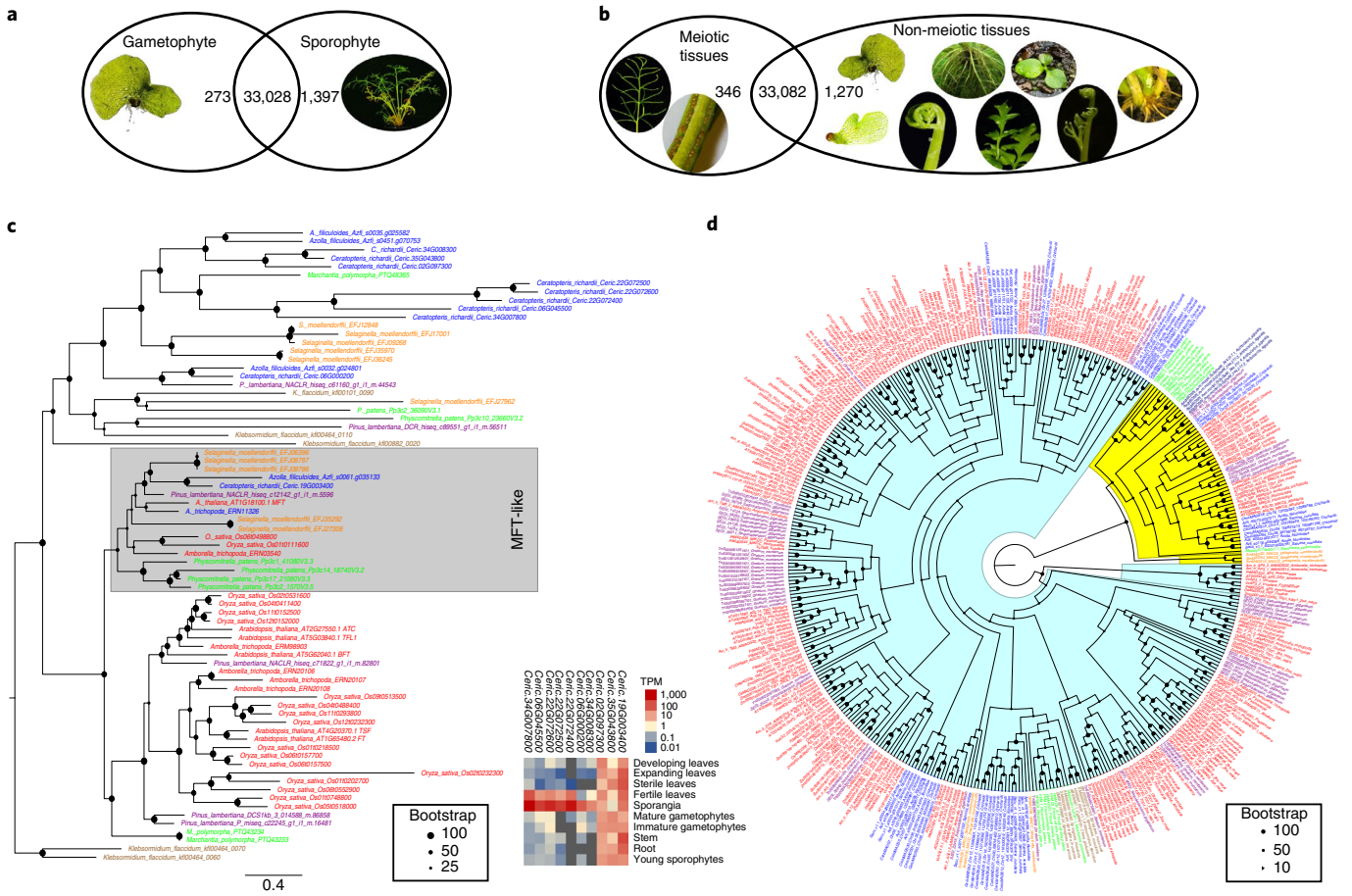


Fig. 4 | Transcriptome profiling and evolution of gene families for plant reproduction and architecture. a, Venn diagram of gametophyte- and sporophyte-specific genes and associated expression heatmaps with $\log_2$(transcripts per million). **b**, Venn diagram of meiotic- and non-meiotic-specific genes and associated expression heatmaps with $\log_2$(transcripts per million). **c**, Phylogeny of *FT* genes and their expression in *Ceratopteris*. The MFT-like gene clade is highlighted in grey. **d**, Phylogeny of Type II MADS-box genes across green plants. MIKC*-group genes are highlighted in yellow, MIKC^C-group genes are highlighted in light blue. For **c** and **d**, angiosperm genes are red, gymnosperm genes are purple, fern genes are blue, lycophyte genes are orange, bryophyte genes are green and algal genes are brown. TPM, transcripts per million.

because they encode a pore-forming cytolytic toxin that forms channels in plasma membranes and has been widely incorporated into biological nanopore research[80,81]. Aerolysin-like genes are hypothesized to have recurrently undergone HGT among different kingdoms[82]. Searches of plant transcriptomic datasets[31,83] found aerolysin-like transcripts in the transcriptomes of each major lineage of land plants with the exception of seed plants (Extended Data Fig. 5). However, when we confined our search to gene models from reference genome assemblies, aerolysin-like genes were found only in *Ceratopteris* and the lycophyte *Selaginella moellendorffii*, as well as diverse algae, fungi, bacteria and fish. HiC data anchored the large array of aerolysin-like genes on chromosome 9 (Extended Data Fig. 4c), refuting the possibility of contamination in *Ceratopteris*. Four and three additional copies of aerolysin-like genes were found on chromosomes 2 and 34, respectively. We found subfunctionalization of the aerolysin-like genes among the different tissues of *Ceratopteris*; 34 of the aerolysin-like genes on chromosome 9 were highly expressed in the stem and roots, and all three of the aerolysin-like genes on chromosome 34 were highly expressed in sterile leaves. One aerolysin-like gene on chromosome 2 was highly expressed in stem, roots and young sporophytes, and the remaining aerolysin-like genes were generally expressed across all tissues (Extended Data Fig. 4d). Similar to *Tma12*, these aerolysin-like genes are likely the result of HGT, possibly recurrent, from bacteria to early land plants. We expect future non-angiosperm reference genomes will provide support for the presence of these genes in plants beyond *Ceratopteris* and *Selaginella* and clarify their evolutionary history.

A second block of tandemly duplicated genes was found on chromosome 11. These genes were annotated as phenolic acid decarboxylases (PADs), which are of particular interest because they catalyse the non-oxidative decarboxylation of potentially toxic phenolic acids to their *p*-vinyl derivatives[84]. PADs are of distinct interest to bioengineering because they have been proposed as biocatalysts given that these vinyl derivative compounds can be used as polymer precursors and flavour/fragrance additives in the food-processing industry[84]. To date, *PAD* genes have only been documented in bacteria; however, we discovered 26 *PAD* genes in *Ceratopteris*, 21 of which originated from tandem duplications on chromosome 11 (43–45 cM). We further searched plant transcriptomic datasets for *PADs* and found them solely in leptosporangiate ferns with bacteria as the sister clade (Fig. 5b). Similar to the aerolysin-like genes, the *PADs* have subfunctionalized because 20 genes were highly expressed in fertile leaves, sporangia and gametophytes, whereas the remaining six were generally expressed across all tissues and developmental stages (Fig. 5c). HGT coupled with rapid diversification via tandem duplications and subfunctionalization of these genes in *Ceratopteris* together provide

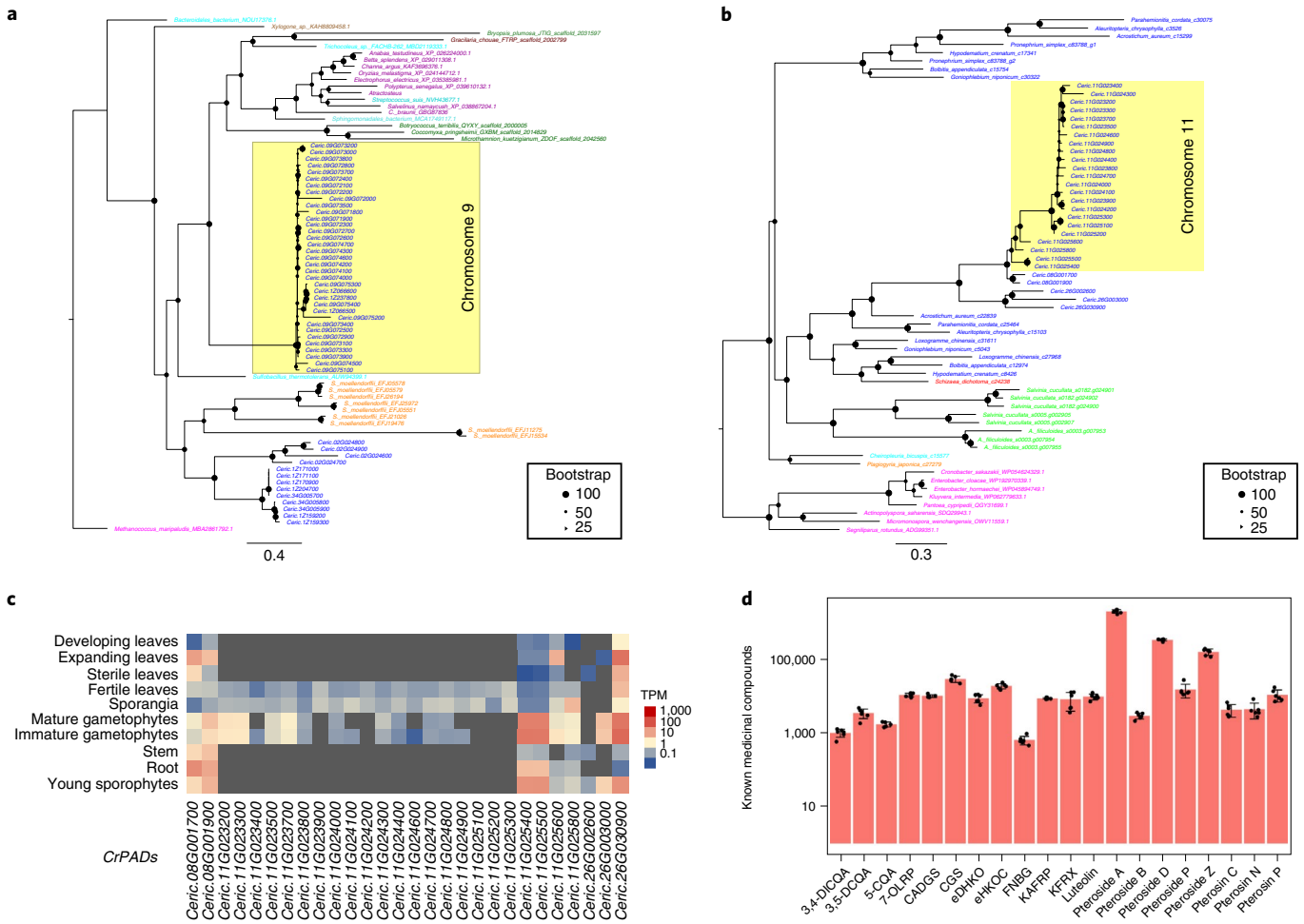


Fig. 5 | HGTs and medicinal compounds in *Ceratopteris*. **a**, Phylogeny of aerolysin-like genes suggests an HGT from bacteria to early vascular plants and a second HGT specifically to ferns, followed by tandem duplications on chromosome 9 in *Ceratopteris* (highlighted). Fern genes are blue, lycophyte genes are orange, fish are purple, bacteria are cyan, fungi are brown, archaea are magenta, chlorophytic algae are dark green and red algae are rust. **b**, Phylogeny of *PAD* genes in *Ceratopteris* and leptosporangiate ferns suggests an HGT from bacteria to ferns followed by rampant tandem duplication across chromosome 11 (highlighted). Polypodiales are blue, Schizaeales are red, Salviniales are light green, Gleicheniales are cyan, Cyatheales are orange and bacteria are pink. **c**, Expression of *Ceratopteris* PAD genes (*CrPADS*) across tissues/life stages. **d**, Metabolic profile of previously identified medicinal compounds in *Ceratopteris*. CADGS, casuarine 6-alpha-D-glucoside; CGS, casuarine 3-glucoside; 5-CQA, *cis*-5-caffeoylquinic acid; 3,5-DCQA, 3,5-di-*O*-caffeoylquinic acid; 3,4-DICQA, 4,5-di-*O*-caffeoylquinic acid; eDHKO, ent-7alpha,12beta-dihydroxy-16-kauren-19,6beta-olide; eHKOC, ent-17-hydroxy-15-kauren-19-oic acid; FNBG, flavanone_7-*O*-beta-D-glucoside; KAFRP, kaempferol 3-arabinofuranoside 7-rhamnofuranoside; KFRX, kaempferol 3-rhamnoside 7-xyloside; luteolin, 2,4′,5,7-tetrahydroxyflavanone; 7-OLRP, 7-*O*-alpha-L-rhamnopyranoside. Data are presented as means ± s.e.m. ($n = 6$).

unique insight into land plant evolution and the integration of novel genes.

The medicinal potential of fern genomics. Ferns have long been used in traditional medicine worldwide and more recently have been a source for bioprospecting medicinal compounds to treat cancer, diabetes and osteoarthritis[9,85–87]. However, the lack of a high-quality reference genome of a homosporous fern species has hindered the research and development of novel fern compounds for medicinal application. Here we leveraged the genomic resources from *Ceratopteris* and performed metabolite profiling of sporophytic tissue to investigate potentially medicinal compounds produced by *Ceratopteris* and the genes underlying their production.

Metabolite profiling of *Ceratopteris* fertile leaf tissue identified several known medicinal compounds, including eight pterosides, seven flavonoids, three caffeic acids and two kauranes (Fig. 5d and Supplementary Table 3). We identified 906 high-confidence metabolites in *Ceratopteris*, compared with 644 metabolites in wheat and rice[88] (Supplementary Table 3). Of those 906 metabolites, 57 were unique compounds only detected in *Ceratopteris* and 131 were novel because they could not be annotated using known metabolome databases.

High-confidence compounds related to the treatment of human diseases were identified in *Ceratopteris*, with enrichment in pathways such as flavonoid biosynthesis, endocrine and metabolic disease, and antimicrobial function (Extended Data Fig. 6). Flavonoids are vital for human nutrition, healthcare and medicine[89]. The flavonoid biosynthesis pathway is specific to land plants and arose when plants colonized land over 470 Ma (ref. [90,91]). In angiosperms, flavonoids play crucial roles in abiotic stress tolerance and are also important for pollination and seed dispersal signalling[91]. Our combined analyses identified the genes, expression patterns and metabolites that compose the flavonoid biosynthesis pathway in *Ceratopteris* (Extended Data Fig. 6c). Alongside such metabolic analyses, our high-quality *Ceratopteris* genome will readily advance our understanding of the molecular origins and functions of these

known and novel compounds to benefit drug discovery in ferns to improve human health.

Conclusions

Homosporous ferns have been the last frontier in green plant genomics owing to their notoriously large genomes and numerous chromosomes; the mechanisms driving the evolution of these genomes have been debated for decades. Despite longstanding hypotheses of rampant WGD in ferns, the *Ceratopteris* genome assembly revealed evidence for at least two WGD events distributed over 300 Myr of fern evolution. By contrast, the *Arabidopsis* and rice genomes each exhibit evidence for three independent sets of WGD events over roughly 125 Myr of flowering plant evolution[31]. Surprisingly, syntenic genomic segments were not evident for even the most recent WGD in *Ceratopteris* owing to frequent tandem duplications, high rates of fractionation and genome rearrangements. Defence-related gene families expanded via extensive tandem duplications and probably originated from separate HGTs from bacteria. In addition, we document CMT-associated CWA methylation and provide unambiguous evidence of gbM in a fern. Importantly, we traced the evolution of genes involved in flower and seed development and overall plant architecture to homologues in fern genes. Ferns have been underutilized as sources of novel genetic material for applications in ecological remediation, medicine and bioprospecting; we demonstrate the potential of these resources. Beyond these scientific discoveries, *Ceratopteris* has long been utilized in biology classrooms as a model for teaching the alternation of generations in green plants[17,92]. With the genetic, genomic and metabolomic resources provided in this study, *Ceratopteris* can become the primary plant system for teaching next-generation plant biology. In conclusion, the *Ceratopteris* genome data provide critical resources for future investigations of gene function in this fern model, and support research in plant biology, genome evolution, biotechnology and medicine, as well as advancing plant biology curricula.

Methods

Genome sequencing. We sequenced *Ceratopteris richardii* genotype Hn-n using a range of sequencing technologies, including single-molecule real-time long-read sequencing from Pacific Biosciences (PacBio), chromosome conformation capture using HiC sequencing, Illumina short-read sequencing, bisulfite sequencing, PacBio Iso-Seq and RNA-seq (Supplementary Table 4). High molecular mass DNA was isolated from fresh leaf tissue at the Arizona Genomics Institute. Sequencing reads were collected using Illumina and PacBio platforms. Illumina and PacBio reads were sequenced at the Department of Energy Joint Genome Institute in Walnut Creek, CA, USA and the HudsonAlpha Institute for Biotechnology in Huntsville, AL, USA. Illumina reads were sequenced using the Illumina NovaSeq platform, and the PacBio reads were sequenced using the SEQUEL platform. Before assembly, Illumina fragment reads were screened for PhiX contamination. Reads composed of >95% simple sequences were removed. Illumina reads <50 bp after trimming for adaptor and quality ($q < 20$) were removed. The final read set consists of 2,438,428,350 reads for a total of 47.27× of high-quality Illumina bases. For the PacBio sequencing, a total of 93 PB chemistry 2.1 chips (10-h movie time) was sequenced with a sequence yield of 777.1 Gb, with a total coverage of 69.02× (Supplementary Table 5).

RNA-seq. Total RNA was isolated from ten tissues and developmental stages of *Ceratopteris richardii* genotype Hn-n (Fig. 1a) using the RNeasy Plant Mini kit (Qiagen). Plate-based RNA sample preparation was performed on the PerkinElmer Sciclone NGS robotic liquid handling system using Illumina's TruSeq Stranded mRNA HT sample prep kit utilizing poly-A selection of mRNA following the protocol outlined by Illumina in their user guide (https://support.illumina.com/sequencing/sequencing_kits/truseq-stranded-mrna.html) with the following conditions: total RNA starting material was 1 µg per sample and eight cycles of PCR were used for library amplification. There are four biological replicates for each tissue and developmental stage of the RNA-seq experiment.

Iso-Seq. With 1 µg of total RNA as input, full-length complementary DNA was synthesized using template switching technology with the SMARTer PCR cDNA Synthesis kit (Clontech). The first-strand cDNA was amplified with PrimeSTAR GL DNA polymerase (Clontech) using template switching oligos to make double-stranded cDNA. Double-stranded cDNA was purified with non-size selected AMPure PB beads (PacBio). The amplified cDNA was end-repaired and ligated with blunt-end PacBio sequencing adaptors using SMRTbell Template Prep Kit 1.0. The ligated products were treated by exonuclease to remove unligated products and purified by AMPure PB beads (PacBio).

Genome assembly and construction of pseudomolecule chromosomes. The v.2.0 assembly was generated by error correcting the 52,299,716 PacBio reads (69.02× sequence coverage) using MECAT assembler v.1.1 (ref. [93]), followed by assembly with the Canu assembler v.1.8 (ref. [94]) and subsequent polishing using ARROW[95]. This produced an initial assembly of 35,249 contigs, with a contig *N50* of 1.5 Mb, 11,081 scaffolds larger than 100 kb and a total assembled size of 9,204.7 Mb (Supplementary Table 6).

Misjoins in the assembly were identified using HiC data as part of the JUICER pipeline[96]. A total of 44 misjoins was identified in the polished assembly. The contigs were then oriented, ordered and joined together using HiC data. In all, 5,031 joins were applied to the broken assembly to form the final assembly of 39 chromosomes. Each chromosome join is padded with 10,000 Ns (placeholders representing any base). A substantial amount of telomeric sequence was identified using the $(TTAGGG)_n$ repeat, and care was taken to make sure that contigs terminating in telomeres were properly oriented in the production assembly. The remaining scaffolds were screened against bacterial proteins, organelle sequences and GenBank nr (non-redundant proteins) and removed if found to be a contaminant. Additional scaffolds were classified as repetitive (>95% masked with 24mers that occur more than four times in the genome) (21 scaffolds, 784.5 kb), prokaryote (158 scaffolds, 8.4 Mb), low quality (composed of <70% PACBIO polished bases) (5 scaffolds, 26.3 kb), redundant (409 scaffolds, 15.2 Mb) and contaminant (361 scaffolds, 4.8 Mb).

Chromosomal scaffolding and validation were enabled with HiC sequencing of the Hn-n genotype and genome skimming of the 58 double haploid lines derived from a HαPQ45 (paraquat-tolerant mutant of Hn-n) × ΦN8 cross[15]. Homozygous single nucleotide polymorphisms (SNPs) and insertions/deletions (INDELs) were corrected in the release consensus sequence using 42× of Illumina reads (2×150, 400 bp insert) by aligning the reads using bwa mem[97] and identifying homozygous SNPs and INDELs with GATK's UnifiedGenotyper tool[98]. A total of 676 homozygous SNPs and 13,913 homozygous INDELs were corrected in the release. The final v.2.0 release contains 7,417.3 Mb of sequence, consisting of 10,785 contigs with a contig *N50* of 2.3 Mb and a total of 93.5% of assembled bases in the 39 chromosomes.

Over 371,000 transcript assemblies were made from 1.5 billion pairs of 2×150 stranded paired-end Illumina RNA-seq reads using PERTRAN, which conducts genome-guided transcriptome short-read assembly via GSNAP[99] and builds splice alignment graphs after alignment validation, realignment and correction on transcript assemblies from PASA[100]. About 15 million PacBio Iso-Seq circular consensus sequences were corrected and collapsed by a genome-guided correction pipeline to obtain >819,000 putative full-length transcripts. Loci were determined by transcript assembly alignments and/or EXONERATE alignments of proteins from *Arabidopsis thaliana*, *Glycine max*, *Sorghum bicolor, Oryza sativa*, *Setaria viridis*, *Solanum lycopersicum*, *Aquilegia coerulea*, *Vitis vinifera*, *Marchantia polymorpha*, *Spaghnum magellanicum*, *Ceratodon purpureus*, *Selaginella moellendorffii*, *Physcomitrium patens, Nymphaea colorata*, *Amborella trichopoda*, *Papaver somniferum*, *Azolla filiculoides, Salvinia cucullata*, and Swiss-Prot proteomes to repeat-soft-mask the *Ceratopteris richardii* genome using RepeatMasker[101] with up to 2,000 bp extension on both ends unless extending into another locus on the same strand. A repeat library consisting of de novo repeats by RepeatModeler[102] and repeats in RepBase was constructed. Gene models were predicted by homology-based predictors, FGENESH+[103], FGENESH_EST (similar to FGENESH+, but using expressed sequence tags, ESTs, to compute splice sites and intron input instead of protein/translated open reading frames, ORFs), EXONERATE[104] and PASA assembly ORFs (in-house homology constrained ORF finder). The best-scored predictions for each locus were selected using multiple positive factors including EST and protein support, and one negative factor, overlap with repeats. The selected gene predictions were improved by PASA. Improvement included adding untranslated regions, splicing correction and adding alternative transcripts. PASA-improved gene model proteins were subject to protein similarity analysis to obtain Cscore and protein coverage. Cscore is a protein BLASTP score ratio to mutual best hit BLASTP score, and protein coverage is the highest percentage of protein aligned to the best homologue. PASA-improved transcripts were selected based on Cscore, protein coverage, EST coverage and its coding sequence (CDS) overlapping with repeats. A transcript was selected if its Cscore was ≥0.5 and protein coverage was ≥0.5, or it had EST coverage, but its CDS overlapping with repeats was <20%. For a gene model whose CDS overlaps with repeats for more than 20%, its Cscore must be at least 0.9 and homology coverage at least 70% to be selected. The selected gene models were subject to Pfam analysis, and gene models whose proteins were >30% in Pfam transposable element domains were removed. Incomplete gene models, gene models with low homology support without full transcriptome support and gene models based on short single exons (<300 bp CDS) without protein domain or good expression were manually filtered out.

Completeness of the euchromatic portion of the assembly was assessed using 39,078 annotated genes from the v.1.0 release of *Ceratopteris*. The aim of this analysis is to obtain a measure of completeness of the assembly, rather than a

comprehensive examination of gene space. The transcripts were aligned to the assembly using BLAT[105], and alignments with ≥95% base pair identity and ≥95% coverage were retained. The screened alignments indicate that 38,458 (98.40%) of the previously annotated *Ceratopteris* genes aligned to the v.2.0 release. Of the remaining annotated genes, 126 (0.32%) aligned at <50% coverage and 68 (0.19%) were not found in the v.2.0 release. The predicted proteome showed that there are 37,263 putative proteins and 3,841 predicted signal peptides in *Ceratopteris* (Supplementary Table 7).

Transcriptome analysis. Clean reads of different transcriptome tissues were mapped to the genome reference by HISAT2 v.2.2.1 (ref. [106]), then sorted by samtools (v.1.11)[107]. Stringtie v.2.1.4 (ref. [108]) and the R package ballgown v.2.20.0 (ref. [109]) were used in quantification of each tissue based on the bam file generated by HISAT2. Tissue/developmental stage-specific genes were those that were only expressed in a single tissue/developmental stage. For example, the 273 gametophyte-specific genes identified in Fig. 4a were expressed in either the mature or immature gametophyte RNA-seq data, but not at all expressed in any of the sporophyte RNA-seq samples. The clusterProfiler package v.3.16.1 (ref. [110]) in R was used to enrich cluster analysis and the Vennerable v.3.0 package (https://R-Forge.R-project.org/projects/vennerable/) was used in Venn diagram analysis. The heatmaps of transcripts per million in different tissues were generated by the R package pheatmap v.1.0.12 (https://CRAN.R-project.org/package=pheatmap).

Comparative genomics. Syntenic orthologous and homeologous regions were defined with the GENESPACE pipeline (https://code.jgi.doe.gov/plant/genespace-r)[111]. In short, GENESPACE accomplishes synteny-constrained orthology inference across multiples species, permitting variable ploidy by parsing protein similarity scores into syntenic blocks with MCScanX[112] and runs Orthofinder[113] on synteny-constrained BLAST results. The resulting block coordinates and syntenic orthogroups give high-confidence anchors for evolutionary inference. We used default parameters for intergenomic comparisons and very relaxed thresholds for within-*Ceratopteris* WGD tests (no constraint to orthogroups, minimum block size, 5; maximum gaps, 25).

K_s distributions. K_s distribution analysis was implemented using the wgd (https://github.com/arzwa/wgd) pipeline[114]. Briefly, all-versus-all BLASTp[115] was used for similarity searches, and the results were then clustered in paralogous families by MCL[116]. Estimation of K_s for all pairs of paralogous genes and tandemly duplicated paralogous genes was performed using the codeml program in the PAML[117] package with the F3X4 model. In further analyses, we used only gene pairs with K_s values in the range of 0.05–3.0. Histograms of K_s distributions were generated by the ggplot2 package v.3.3.3 (ref. [118]) in R.

MAPS and NOTUNG analyses. To determine the phylogenetic placement of ancient WGD events detected in the *Ceratopteris* genome, we used the *Ceratopteris richardii* genome and other fern transcriptomes for MAPS and NOTUNG analyses. For MAPS, orthologous groups for the selected species were obtained from Orthofinder[113] with the default parameters and only retained gene families that contained at least one gene copy from each taxon. The phylogenetic trees for gene families constructed by PASTA[119] were analysed by MAPS. Both null and positive simulations of the background gene birth and death rates were performed to compare with the observed number of duplications at each node. For null simulations, we estimated the gene birth rate (λ) and death rate (μ) for the selected species with WGDgc[120]. Gene count data of each gene family for all species were obtained from Orthofinder[113]. The species tree for each MAPS analysis was obtained based on previous phylogenetic analyses[31]. The estimated parameters λ and μ were configured in MAPS, and the gene trees were then simulated within the species tree using the 'GuestTreeGen' program from GenPhyloData[121]. For each species tree, we simulated 3,000 gene trees with at least one tip per species: 1,000 gene trees at the estimated λ and μ, 1,000 gene trees at half of the estimated λ and μ, and 1,000 trees at three times λ and μ[31,122]. We then randomly resampled 1,000 trees without replacement from the total pool of gene trees 100 times to provide a measure of uncertainty on the percentage of subtrees at each node. A Fisher's exact test was used to identify locations with significant increases in gene duplication compared with a null simulation. For positive simulations, we simulated gene trees using the same methods described above. However, we incorporated WGDs at the location in the MAPS phylogeny with significantly larger numbers of gene duplications compared with the null simulation. We allowed at least 20% of the genes to be retained following the simulated WGD[31,122].

We also performed gene tree reconciliation using NOTUNG v.2.9.1.5 (ref. [123]) with a model of gene duplication and loss without HGTs. We used protein alignments from PASTA[119] as described above. RAxML v.8.2.11 (ref. [124]) was used to generate gene family phylogenies with 100 bootstraps. The species trees were rerooted with an outgroup under the '-reroot' function in NOTUNG. The 80% bootstrap value was used as a threshold to rearrange branches with low support on gene trees based on the species tree topology under the '-rearrange' function. The gene tree reconciliation was performed using all gene family phylogenies. The cost of loss was set to 0.1 to account for missing data of transcriptomes[125].

Paleologue identification using Frackify. We used Frackify to identify paleologues in the *Ceratopteris* genome originating from the WGD peak at K_s 1.3 (ref. [53]). Frackify is a machine learning approach that uses multiple features from gene age distributions and synteny to identify paleologues in genomes[53]. Syntenic blocks were identified using MCScanX[112] set to a minimum match size (-s) of three based on Zhoa and Shranz[126]. This analysis recovered 193 syntenic blocks representing 923 collinear genes. Given that *Ceratopteris* does not have a closely related outgroup with a high-quality reference genome available, we used a version of Frackify trained without an outgroup (https://gitlab.com/barker-lab/frackify). For this analysis, Frackify identified a total of 395 paleologues within the syntenic blocks identified by MCScanX.

To assess the relative rate of fractionation in the *Ceratopteris* genome compared with other plant genomes, we compared the distribution of syntenic block sizes and WGD age for *Ceratopteris, Azolla filiculoides, Salvinia cucullata, Ginkgo biloba* and 27 flowering plant genomes (Supplementary Table 8)[53]. We used MCScanX set to minimum match size (-s) of three to identify synteny blocks in the 31 plant species[126]. We then compared the distribution of syntenic block lengths in *Ceratopteris* against the entire dataset (Fig. 2c). Finally, we used the fitdistr() function from the MASS R library to fit a log normal distribution to the distribution of syntenic block lengths in each species[127]. We compared the standard deviation of the log normal distribution in each species against the mean K_s of the focal WGD (Fig. 2c).

Whole-genome bisulfite sequencing and DNA methylation patterns on genes and repeats. The *Ceratopteris* bisulfite sequencing data used in this study were processed by Methylpy v.1.4.2 as described in Schultz et al.[128]. Quality filtering and adaptor trimming were performed using cutadapt v.1.18 (ref. [129]). Qualified reads were aligned to the *Ceratopteris* v.2.1 reference genome using bowtie v.2.4.1 (ref. [130]). Only uniquely aligned and non-clonal reads were retained (Supplementary Table 9). Lambda genomic DNA was used as a control to calculate the sodium bisulfite reaction conversion rate of unmodified cytosines, which was >99.9% for this sample. A binomial test was used to determine the methylation status of cytosines with a minimum coverage of three reads.

The gene or repeat body was divided into 20 windows. Additionally, regions 1,000 bp upstream and downstream were each divided into twenty 50 bp windows. Methylation levels were calculated for each window according to previous recommendations[128]. The mean methylation level for each window was then calculated for all genes and all repeats, respectively. Locations of genes and repeats were obtained from the annotated gff files of the *Ceratopteris* v.2.0 reference.

The two *CMT* genes (*Ceric.34G031800*, *Ceric.12G007000*) were identified from our annotations (PFAM, PANTHER, KEGG) and possess the BAH, CHROMO and C-5 cytosine-specific DNA methylase domains.

Evolutionary analysis of gene families. Unless noted otherwise, comparative genetic similarity analysis of gene families across the major green plant lineages and algae was described in Chen et al.[131]. Briefly, the candidate protein sequence was searched in the *Arabidopsis* protein sequences database with the criteria of E-value $<10^{-5}$ and paired with the *Arabidopsis* protein in the first hit (with highest BLAST score). The number and protein similarity of gene families in each plant species were calculated using the *Arabidopsis* gene family as a reference. Genome sequence data of all species were obtained from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) and Ensembl Plants (http://plants.ensembl.org/index.html). Transcriptomic sequence data of plants were obtained from the One Thousand Plant Transcriptomes (1KP) database[31] and a recent fern transcriptome analysis[83]. The sequences were aligned with MAFFT[132], the best model was estimated with IQ-Tree, and the phylogeny was constructed with 1,000 fast-bootstrap replicates[133].

MADS-box identification and phylogeny. To identify MADS-box genes in *Ceratopteris*, we first searched for MADS domains in the predicted proteins using HMMER 3.1b2 with a customized hidden Markov model for the MADS domain[134]. To also detect MADS-box genes that may have escaped automatic annotation, the *Ceratopteris* genome was translated in all six possible reading frames into amino acid sequences, and HMMER 3.1b2 searches were repeated on the translated genome. Genomic regions for which new MADS domains were recognized were extracted, including 100,000 nucleotides up- and downstream of the sequence potentially encoding a MADS domain. Genes were predicted on these genomic regions using AUGUSTUS 3.3.2 (ref. [135]) with the parameters '--strand=forward--codingseq=on --species=arabidopsis'.

BLAST searches[136] of the identified MADS-domain proteins were conducted on NCBI using 'non-redundant' protein sequences as the database. In cases where the BLAST results indicated suboptimal gene models, an improved similarity-based gene prediction was attempted using FGENESH+[137] with the most similar protein as a guide. In some cases, the similarity-based gene prediction also failed, presumably due to the presence of large introns. In these cases, we manually annotated the MADS-box genes. To do so, we compared the genomic regions with the most similar MADS-domain protein using BLAST with the option 'Align two or more sequences' and then annotated exons complying with conventional splice sites and keeping an open reading frame. The plausibility of the

manually annotated MADS-box genes was checked by searching these sequences in short-read archive data[138] for *Ceratopteris* transcriptomes on NCBI. We initially identified 39 MADS-box genes in *Ceratopteris* (Supplementary Table 10). Four of these were classified as potential pseudogenes, whereas 16 of the 35 potential genuine MADS-box genes correspond to previously identified MADS-box genes in *Ceratopteris* (Supplementary Table 10)[68–70,139].

Protein sequences of MADS-domain proteins were aligned using Probalign[140] on the CIPRES Science Gateway v.3.3 (ref. [141]). Alignment of the complete set of MADS-domain proteins was trimmed using trimAl 1.2rev59[142] with the parameters '-gt .9 -st .00001' to remove positions with low conservation. For the alignment of the Type II MADS-domain proteins, we used the parameters '-gt .8 -st .0001' for the first trimming. Both trimmed alignments were then trimmed with the parameters '-seqoverlap 75 -resoverlap 0.7' to remove sequences with low conservation. Phylogenies were reconstructed using RAxML v.8.2.12 (ref. [124]) on the CIPRES Science Gateway[141]. Based on the phylogeny of the complete set of MADS-domain proteins, Type I and Type II MADS-domain proteins were separated, and a phylogeny of Type II MADS-domain proteins was reconstructed.

Identification of *Tma12* in ferns. We used BLASTp to search for *Tma12* (JQ438776 in GenBank) homologues in 1KP transcriptomes[31] and 69 fern transcriptomes[83], as well as the published water fern genomes of *Azolla filiculoides* and *Salvinia cucullata*[20], using an E-value $<10^{-10}$. HMMER 3.1b2 was also used in conserved domain building based on the fungal *Tma12* orthologue sequences and predicted protein searching. After removing the alternative isoforms and filtering the sequences in which the length is shorter than one-third of the reference sequence length, we found six *Tma12* orthologous genes in the fern transcriptomes database, and two in *Ceratopteris* (Ceric.24G052800, Ceric.1Z115600) that have the same amino acid sequences but no hits were found in the 1KP database. In *Salvinia cucullata*, the only hit was found in the genome data, but not in annotated protein sequences. To explore the phylogeny of *Tma12* in ferns, we selected three fungal *Tma12* sequences (*Streptosporangium subroseum*, *Thermopolyspora flexuosa*, *Actinomadura echinospora*) as the outgroup. The protein sequences were aligned with MAFFT[132], the best model was estimated with IQ-Tree, and the phylogeny was constructed with 1,000 fast-bootstrap replicates[133].

Identification of aerolysin-like genes and evolution. We initially identified the aerolysin-like genes on chromosome 9 while investigating blocks of tandemly duplicated genes based on our Panther annotations (PTHR34007). We used BLASTp to search for aerolysin-like homologues for other fern species in 69 fern transcriptomes[83], for algae in 1KP transcriptomes[31] and for bacteria in NCBI using an E-value $<10^{-5}$. We then restricted our search to the *Ceratopteris* genome and published genomes as described in the gene family analysis section. The sequences were aligned with MAFFT[132], the best model was estimated with IQ-Tree and the phylogeny was constructed with 1,000 fast-bootstrap replicates[133].

Identification of PADs genes and evolution. We used BLASTp to search for PAD-like homologues in 69 fern transcriptomes[83] and published genomes (species used in the gene family analysis section) using an E-value $<10^{-5}$. The sequences were aligned with MAFFT[132], the best model was estimated with IQ-Tree, and the phylogeny was constructed with 1,000 fast-bootstrap replicates[133].

Analysis of metabolites. Young sporophytes of *Ceratopteris* were collected with six replicates, and metabolites were extracted with 1:1 methanol:water buffer. The samples were stored at −80°C before liquid chromatography–mass spectrometry analysis. Pooled quality control (QC) samples were also prepared by combining 10 µl of each extract. All samples were analysed using a TripleTOF 5600 Plus high-resolution tandem mass spectrometer (SCIEX) with both positive and negative ion modes. Chromatographic separation was performed using an ultra-performance liquid chromatography system (SCIEX). An ACQUITY UPLC T3 column (100 mm × 2.1 mm, 1.8 µm) was used for the reversed-phase separation.

The TripleTOF 5600 Plus system was used to detect metabolites eluted from the column. The ion spray floating voltages were set at 5 and −4.5 kV for the positive ion mode and negative ion mode, respectively. The mass spectrometry data were acquired in the IDA mode. The TOF mass range was 60–1,200 Da. A QC sample was analysed every ten samples to evaluate the stability of the liquid chromatography–mass spectrometry.

Raw data files were converted into mzXML format and then processed using the XCMS, CAMERA and metaX toolbox in R. Each ion was identified by the comprehensive information of retention time and *m*/*z*. The open access databases, KEGG[143] and HMDB[144], were used to annotate the metabolites by matching the exact molecular mass data (*m*/*z*) to those from the database within a threshold of 10 ppm. The peak intensity data were further preprocessed using metaX. Features that were detected in <50% of QC samples or 80% of test samples were removed, and values for missing peaks were extrapolated with the *k*-nearest neighbour algorithm to improve the data quality further. Data normalization was performed on all samples using the probabilistic quotient normalization algorithm. Then, QC-robust spline batch correction was performed using the QC samples. *P* values from Student's *t*-tests were adjusted for multiple tests using an FDR correction (Benjamini–Hochberg) for the metabolite selection. The VIP cut-off value of 1.0 was set to select important features (Supplementary Table 11).

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

Data availability

Raw genomic sequences and assemblies have been deposited in the NCBI SRA under BioProject PRJNA729743. Genome and transcriptome assemblies and annotations can be found in Phytozome (https://phytozome-next.jgi.doe.gov/info/Crichardii_v2_1). Publicly available data were collected from Ensembl Plants (plants.ensembl.org), NCBI (ncbi.nlm.nih.gov), Swiss-Prot (uniprot.org), RepBase (girinst.org/repbase), One Thousand Plant Transcriptomes (1KP) database (OTPT Initiative, 2019), fern transcriptome database[83], water fern genomes (fernbase.org), spruce genome (congenie.org), TAIR (arabidopsis.org). Source data are provided with this paper.

Code availability

All custom codes are available for research purposes from the corresponding authors upon request.

Received: 12 January 2022; Accepted: 15 July 2022;
Published online: 1 September 2022

References

1. Kenrick, P. & Crane, P. R. The origin and early evolution of plants on land. *Nature* **389**, 33–39 (1997).
2. Lloyd, R. M. Mating systems and genetic load in pioneer and non-pioneer Hawaiian Pteridophyta. *Bot. J. Linn. Soc.* **69**, 23–35 (1974).
3. Ellwood, M. D. F. & Foster, W. A. Doubling the estimate of invertebrate biomass in a rainforest canopy. *Nature* **429**, 549–551 (2004).
4. de León, S. G., Briones, O., Aguirre, A., Mehltreter, K. & Pérez-García, B. Germination of an invasive fern responds better than native ferns to water and light stress in a Mexican cloud forest. *Biol. Invas.* **20**, 3187–3199 (2021).
5. Raja, W., Rathaur, P., John, S. A. & Ramteke, P. W. *Azolla*: an aquatic pteridophyte with great potential. *Int. J. Res. Biol. Sci.* **2**, 68–72 (2012).
6. PPG I. A community-derived classification for extant lycophytes and ferns. *J. Syst. Evol.* **54**, 563–603 (2016).
7. Mehltreter, K., Walker, L. R. & Sharpe, J. M. (eds) *Fern Ecology* (Cambridge Univ. Press, 2010).
8. Cao, H. et al. Phytochemicals from fern species: potential for medicine applications. *Phytochem. Rev.* **16**, 379–440 (2017).
9. Goswami, H. K., Sen, K. & Mukhopadhyay, R. Pteridophytes: evolutionary boon as medicinal plants. *Plant Genet. Resour.* **14**, 328–355 (2016).
10. Shukla, A. K. et al. Expression of an insecticidal fern protein in cotton protects against whitefly. *Nat. Biotechnol.* **34**, 1046–1051 (2016).
11. Sessa, E. B. & Der, J. P. in *Advances in Botanical Research* Vol. 78, (ed. Rensing, S. A.) 215–254 (Elsevier, 2016).
12. Klekowski, E. & Baker, H. Evolutionary significance of polyploidy in the Pteridophyta. *Science* **153**, 305–307 (1966).
13. Haufler, C. H. Ever since Klekowski: testing a set of radical hypotheses revives the genetics of ferns and lycophytes. *Am. J. Bot.* **101**, 2036–2042 (2014).
14. Soltis, D. E. & Soltis, P. S. Polyploidy and breeding systems in homosporous Pteridophyta: a reevaluation. *Am. Nat.* **130**, 219–232 (1987).
15. Nakazato, T., Jung, M.-K., Housworth, E. A., Rieseberg, L. H. & Gastony, G. J. Genetic map-based analysis of genome structure in the homosporous fern *Ceratopteris richardii*. *Genetics* **173**, 1585–1597 (2006).
16. Marchant, D. B. et al. The C-Fern (*Ceratopteris richardii*) genome: insights into plant genome evolution with the first partial homosporous fern genome assembly. *Sci. Rep.* **9**, 18181 (2019).
17. Hickok, L. G., Warne, T. R., Baxter, S. L. & Melear, C. T. Education: sex and the C-Fern: not just another life cycle. *Bioscience* **48**, 1031–1037 (1998).
18. Sessa, E. B. et al. Between two fern genomes. *Gigascience* **3**, 15 (2014).
19. Marchant, D. B. Ferns with benefits: incorporating *Ceratopteris* into the genomics era. *Am. Fern J.* **109**, 183–191 (2019).
20. Li, F.-W. et al. Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat. Plants* **4**, 460–472 (2018).
21. Xiong, X. et al. The *Taxus* genome provides insights into paclitaxel biosynthesis. *Nat. Plants* **7**, 1026–1036 (2021).
22. Zonneveld, B. J. M. Conifer genome sizes of 172 species, covering 64 of 67 genera, range from 8 to 72 picogram. *Nord. J. Bot.* **30**, 490–502 (2012).
23. Scott, A. D. et al. A reference genome sequence for giant sequoia. *G3* **10**, 3907–3919 (2020).
24. Lisch, D. How important are transposons for plant evolution? *Nat. Rev. Genet.* **14**, 49–61 (2013).
25. Hawkins, J. S., Proulx, S. R., Rapp, R. A. & Wendel, J. F. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proc. Natl Acad. Sci. USA* **106**, 17811–17816 (2009).
26. Galindo-González, L., Mhiri, C., Deyholos, M. K. & Grandbastien, M.-A. LTR-retrotransposons in plants: engines of evolution. *Gene* **626**, 14–25 (2017).

27. 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).
28. Wegrzyn, J. L. et al. Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics* **196**, 891–909 (2014).
29. Ren, X.-Y., Vorst, O., Fiers, M. W. E. J., Stiekema, W. J. & Nap, J.-P. In plants, highly expressed genes are the least compact. *Trends Genet.* **22**, 528–532 (2006).
30. Callis, J., Fromm, M. & Walbot, V. Introns increase gene expression in cultured maize cells. *Genes Dev.* **1**, 1183–1200 (1987).
31. One Thousand Plant Transcriptomes InitiativeOne thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
32. Soltis, P. S., Marchant, D. B., Van de Peer, Y. & Soltis, D. E. Polyploidy and genome evolution in plants. *Curr. Opin. Genet. Dev.* **35**, 119–125 (2015).
33. Haufler, C. H. & Soltis, D. E. Genetic evidence suggests that homosporous ferns with high chromosome numbers are diploid. *Proc. Natl Acad. Sci. USA* **83**, 4389–4393 (1986).
34. Haufler, C. H. Electrophoresis is modifying our concepts of evolution in homosporous pteridophytes. *Am. J. Bot.* **74**, 953–966 (1987).
35. Wagner, W. H. Reticulate evolution in the Appalachian Aspleniums. *Evolution* **8**, 103–118 (1954).
36. Wagner, W. H. in *Distributional History of the Biota of the Southern Appalachians* (eds Holt, P. C. & Paterson R. A.) 147–192 (Virginia Polytechnic Institute, 1971).
37. Klekowski, E. Genetical features of ferns as contrasted with seed plants. *Ann. Mo. Bot. Gard.* **59**, 138–151 (1972).
38. Barker, M. S., Vogel, H. & Schranz, M. E. Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biol. Evol.* **1**, 391–399 (2009).
39. Clark, J. et al. Genome evolution of ferns: evidence for relative stasis of genome size across the fern phylogeny. *New Phytol.* **210**, 1072–1082 (2016).
40. Barker, M. S. & Wolf, P. G. Unfurling fern biology in the genomics age. *Bioscience* **60**, 177–185 (2010).
41. Li, Z. et al. Early genome duplications in conifers and other seed plants. *Sci. Adv.* **1**, e1501084 (2015).
42. Chen, K., Durand, D. & Farach-Colton, M. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.* **7**, 429–447 (2000).
43. Lai, H., Stolzer, M. & Durand, D. Fast heuristics for resolving weakly supported branches using duplication, transfers, and losses. In *RECOMB International Workshop on Comparative Genomics* (eds Meidanis, J. & Nakhleh, L.) 298–320 (Springer, 2017).
44. Schuettpelz, E. & Pryer, K. M. Evidence for a Cenozoic radiation of ferns in an angiosperm-dominated canopy. *Proc. Natl Acad. Sci. USA* **106**, 11200–11205 (2009).
45. Rothfels, C. J. et al. The evolutionary history of ferns inferred from 25 low-copy nuclear genes. *Am. J. Bot.* **102**, 1089–1107 (2015).
46. Huang, X. et al. The flying spider-monkey tree fern genome provides insights into fern evolution and arborescence. *Nat. Plants* **8**, 500–512 (2022).
47. Chen, H. et al. Revisiting ancient polyploidy in leptosporangiate ferns. Preprint at *bioRxiv* https://doi.org/10.1101/2022.03.12.484015 (2022).
48. Vanneste, K., Baele, G., Maere, S. & Van de Peer, Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Res.* **24**, 1334–1347 (2014).
49. Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433–438 (2003).
50. Dodsworth, S., Chase, M. W. & Leitch, A. R. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Bot. J. Linn. Soc.* **180**, 1–5 (2016).
51. Mandáková, T. & Lysak, M. A. Post-polyploid diploidization and diversification through dysploid changes. *Curr. Opin. Plant Biol.* **42**, 55–65 (2018).
52. Li, Z. et al. Patterns and processes of diploidization in land plants. *Annu. Rev. Plant Biol.* **72**, 387–410 (2021).
53. McKibben, M. T. W. & Barker, M. S. Applying machine learning to classify the origins of gene duplications. Preprint at *bioRxiv* https://doi.org/10.1101/2021.08.12.456144 (2021).
54. Cokus, S. J. et al. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
55. Lister, R. et al. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
56. Gouil, Q. & Baulcombe, D. C. DNA methylation signatures of the plant chromomethyltransferases. *PLoS Genet.* **12**, e1006526 (2016).
57. Bewick, A. J. et al. The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants. *Genome Biol.* **18**, 65 (2017).
58. You, C. et al. Conservation and divergence of small RNA pathways and microRNAs in land plants. *Genome Biol.* **18**, 158 (2017).
59. Niederhuth, C. E. et al. Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* **17**, 194 (2016).
60. Takuno, S., Ran, J.-H. & Gaut, B. S. Evolutionary patterns of genic DNA methylation vary across land plants. *Nat. Plants* **2**, 15222 (2016).
61. Vasco, A. et al. Challenging the paradigms of leaf evolution: class III HD-Zips in ferns and lycophytes. *New Phytol.* **212**, 745–758 (2016).
62. Zumajo-Cardona, C., Vasco, A. & Ambrose, B. A. The evolution of the *KANADI* gene family and leaf development in lycophytes and ferns. *Plants* **8**, 313 (2019).
63. Zumajo-Cardona, C., Pabón-Mora, N. & Ambrose, B. A. The evolution of *euAPETALA2* genes in vascular plants: from plesiomorphic roles in sporangia to acquired functions in ovules and fruits. *Mol. Biol. Evol.* **38**, 2319–2336 (2021).
64. Zumajo-Cardona, C., Little, D. P., Stevenson, D. & Ambrose, B. A. Expression analyses in *Ginkgo biloba* provide new insights into the evolution and development of the seed. *Sci. Rep.* **11**, 21995 (2021).
65. Rajkumar, K. et al. Understanding perspectives of signalling mechanisms regulating PEBP1 function. *Cell Biochem. Funct.* **34**, 394–403 (2016).
66. Jin, S., Nasim, Z., Susila, H. & Ahn, J. H. Evolution and functional diversification of FLOWERING LOCUS T/TERMINAL FLOWER 1 family genes in plants. *Semin. Cell Dev. Biol.* **109**, 20–30 (2021).
67. Smaczniak, C., Immink, R. G. H., Angenent, G. C. & Kaufmann, K. Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. *Development* **139**, 3081–3098 (2012).
68. Münster, T. et al. Floral homeotic genes were recruited from homologous MADS-box genes preexisting in the common ancestor of ferns and seed plants. *Proc. Natl Acad. Sci. USA* **94**, 2415–2420 (1997).
69. Hasebe, M., Wen, C.-K., Kato, M. & Banks, J. A. Characterization of MADS homeotic genes in the fern *Ceratopteris richardii*. *Proc. Natl Acad. Sci. USA* **95**, 6222–6227 (1998).
70. Kofuji, R. & Yamaguchi, K. Isolation and phylogenetic analysis of MADS genes from the fern *Ceratopteris richardii*. *J. Phytogeogr. Taxon.* **45**, 83–91 (1997).
71. Gramzow, L. & Theissen, G. A hitchhiker's guide to the MADS world of plants. *Genome Biol.* **11**, 214 (2010).
72. Theissen, G. et al. A short history of MADS-box genes in plants. *Plant Mol. Biol.* **42**, 115–149 (2000).
73. Sheldon, C. C., Conn, A. B., Dennis, E. S. & Peacock, W. J. Different regulatory regions are required for the vernalization-induced repression of FLOWERING LOCUS C and for the epigenetic maintenance of repression. *Plant Cell* **14**, 2527–2537 (2002).
74. Sieburth, L. E. & Meyerowitz, E. M. Molecular dissection of the AGAMOUS control region shows that cis elements for spatial regulation are located intragenically. *Plant Cell* **9**, 355–365 (1997).
75. Kooiker, M. et al. BASIC PENTACYSTEINE1, a GA binding protein that induces conformational changes in the regulatory region of the homeotic *Arabidopsis* gene SEEDSTICK. *Plant Cell* **17**, 722–729 (2005).
76. Schauer, S. E. et al. Intronic regulatory elements determine the divergent expression patterns of AGAMOUS-LIKE6 subfamily members in *Arabidopsis*. *Plant J.* **59**, 987–1000 (2009).
77. Akhter, S. et al. Integrative analysis of three RNA sequencing methods identifies mutually exclusive exons of MADS-box isoforms during early bud development in *Picea abies*. *Front. Plant Sci.* **9**, 1625 (2018).
78. Markham, K., Chalk, T. & Stewart, C. N. Jr Evaluation of fern and moss protein-based defenses against phytophagous insects. *Int. J. Plant Sci.* **167**, 111–117 (2006).
79. Zaynab, M. et al. Role of secondary metabolites in plant defense against pathogens. *Microb. Pathog.* **124**, 198–202 (2018).
80. Cao, C. et al. Single-molecule sensing of peptides and nucleic acids by engineered aerolysin nanopores. *Nat. Commun.* **10**, 4918 (2019).
81. Cao, C. et al. Mapping the sensing spots of aerolysin for single oligonucleotides analysis. *Nat. Commun.* **9**, 2823 (2018).
82. Moran, Y., Fredman, D., Szczesny, P., Grynberg, M. & Technau, U. Recurrent horizontal transfer of bacterial toxin genes to eukaryotes. *Mol. Biol. Evol.* **29**, 2223–2230 (2012).
83. Shen, H. et al. Large-scale phylogenomic analysis resolves a backbone phylogeny in ferns. *Gigascience* **7**, gix116 (2018).
84. Sheng, X., Lind, M. E. S. & Himo, F. Theoretical study of the reaction mechanism of phenolic acid decarboxylase. *FEBS J.* **282**, 4703–4713 (2015).
85. Yahara, Y. et al. Pterosin B prevents chondrocyte hypertrophy and osteoarthritis in mice by inhibiting Sik3. *Nat. Commun.* **7**, 10959 (2016).
86. Jannat, S. et al. Inhibition of β-site amyloid precursor protein cleaving enzyme 1 and cholinesterases by pterosins via a specific structure–activity relationship with a strong BBB permeability. *Exp. Mol. Med.* **51**, 1–18 (2019).

87. Hsu, F.-L. et al. Antidiabetic effects of pterosin A, a small-molecular-weight natural product, on diabetic mouse models. *Diabetes* **62**, 628–638 (2013).
88. Zhang, X. et al. Metabolite profiling for model cultivars of wheat and rice under ozone pollution. *Environ. Exp. Bot.* **179**, 104214 (2020).
89. Mondal, S. & Rahaman, S. T. Flavonoids: a vital resource in healthcare and medicine. *Pharm. Pharmacol. Int. J.* **8**, 91–104 (2020).
90. Rubinstein, C. V., Gerrienne, P., de la Puente, G. S., Astini, R. A. & Steemans, P. Early Middle Ordovician evidence for land plants in Argentina (eastern Gondwana). *New Phytol.* **188**, 365–369 (2010).
91. Davies, K. M. et al. The evolution of flavonoid biosynthesis: a bryophyte perspective. *Front. Plant Sci.* **11**, 7 (2020).
92. Spiro, M. D. & Knisely, K. I. Alternation of generations and experimental design: a guided-inquiry lab exploring the nature of the her1 developmental mutant of *Ceratopteris richardii* (C-Fern). *CBE Life Sci. Educ.* **7**, 82–88 (2008).
93. Xiao, C.-L. et al. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat. Methods* **14**, 1072–1074 (2017).
94. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
95. Chin, C.-S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
96. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
97. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3397 (2013).
98. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
99. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
100. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
101. Smit, A. F. A., Hubley, R. & Green, P. 2013–2015. RepeatMasker Open-4.0 (2013); http://www.repeatmasker.org
102. Smit, A. F. A. & Hubley, R. RepeatModeler Open-1.0 (2008); https://www.repeatmasker.org
103. Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
104. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
105. Kent, W. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
106. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
107. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
108. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
109. Frazee, A. C. et al. Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat. Biotechnol.* **33**, 243–246 (2015).
110. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
111. Lovell, J. T. et al. The genomic landscape of molecular responses to natural drought stress in *Panicum hallii*. *Nat. Commun.* **9**, 5213 (2018).
112. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
113. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
114. Zwaenepoel, A. & Van de Peer, Y. wgd—simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* **35**, 2153–2155 (2019).
115. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
116. Altschul, S. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
117. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
118. Wickham, H. *ggplot2* 189–201 (Springer, 2016).
119. Mirarab, S. et al. PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *J. Comput. Biol.* **22**, 377–386 (2015).
120. Rabier, C.-E., Ta, T. & Ané, C. Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Mol. Biol. Evol.* **31**, 750–762 (2014).
121. Sjöstrand, J., Arvestad, L., Lagergren, J. & Sennblad, B. GenPhyloData: realistic simulation of gene family evolution. *BMC Bioinformatics* **14**, 209 (2013).
122. Li, Z. & Barker, M. S. Inferring putative ancient whole-genome duplications in the 1000 Plants (1KP) initiative: access to gene family phylogenies and age distributions. *Gigascience* **9**, giaa004 (2020).
123. Stolzer, M. et al. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* **28**, i409–i415 (2012).
124. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
125. Koenen, E. J. M. et al. The origin of the legumes is a complex paleopolyploid phylogenomic tangle closely associated with the Cretaceous–Paleogene (K–Pg) mass extinction event. *Syst. Biol.* **70**, 508–526 (2021).
126. Zhao, T. & Schranz, M. E. Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proc. Natl Acad. Sci. USA* **116**, 2165–2174 (2019).
127. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* 271–300 (Springer, 2002).
128. Schultz, M. D. et al. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**, 212–216 (2015).
129. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
130. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
131. Chen, Z.-H. et al. Molecular evolution of grass stomata. *Trends Plant Sci.* **22**, 124–139 (2017).
132. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
133. 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).
134. Gramzow, L. & Theißen, G. Phylogenomics of MADS-box genes in plants—two opposing life styles in one gene family. *Biology* **2**, 1150–1164 (2013).
135. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
136. Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
137. Solovyev, V., Kosarev, P., Seledsov, I. & Vorobyev, D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **7**, S10 (2006).
138. Sayers, E. W. et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **39**, D38–D51 (2010).
139. Kwantes, M., Liebsch, D. & Verelst, W. How MIKC* MADS-box genes originated and evidence for their conserved function throughout the evolution of vascular plant gametophytes. *Mol. Biol. Evol.* **29**, 293–302 (2011).
140. Roshan, U. & Livesay, D. R. Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics* **22**, 2715–2721 (2006).
141. Miller, M. A., Pfeiffer, W. & Schwartz, T. The CIPRES science gateway: a community resource for phylogenetic analyses. In *Proc. 2011 TeraGrid Conference: Extreme Digital Discovery* 1–8 (Association for Computing Machinery, 2011).
142. 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).
143. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
144. Wishart, D. S. et al. HMDB: the human metabolome database. *Nucleic Acids Res.* **35**, D521–D526 (2007).

Acknowledgements

We thank L. Hickok for spores of *Ceratopteris richardii*. We thank G. Jin (Guhe Technology, China) and Central Laboratories of Zhejiang Academy of Agricultural Sciences for sequencing and bioinformatics service. We also thank D. Randall, S. Shan, C. Zhao, C. Solis and M. Yong for plant maintenance and photography. The work (proposal: 10.46936/10.25585/60001405) conducted by the US Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the US Department of Energy operated under Contract No. DE-AC02-05CH11231. This work was supported by US National Science Foundation (NSF; grant PRFB IOS-1907220 to D.B.M.; grant MCB-1856143 to R.J.S.; grant DEB-1911459 to P.G.W.; grant MRI-1828479 to B.A.A.); the Natural Science Foundation of China (NSFC; grant 32001456 to G.C.; grant 31771687 to S.C.; grant 41571049 to Y. Wu); China Agricultural Research Systems (grant CARS-05 to S.C.); the Key R&D Program of Zhejiang Province (grant 2021C02009 to S.X.) and Zhejiang Provincial Natural Science Foundation of China (grant LY21C130007 to D.X.).the Australian Research Council (grants FT210100366, DE1401011143 to Z.-H.C.); the Horticulture Innovation Australia (grants LP18000, VG17003 to Z.-H.C.). B.A.A. and D.W.S. acknowledge funding for this project from the Ambrose Monell Foundation.

Author contributions

D.B.M., P.S.S., D.E.S., J.H.L.-M. and Z.-H.C. conceived and supervised the project. D.B.M., G.C., S.C., J.G., S.S., D.X., F.C., T.T. and F.X. conducted experimental work. J.S. and J.H.L.-M. led the genome sequencing work. J.G. and S.S. undertook the annotation of the genome assembly. D.B.M., G.C., S.C., J.G., S.S., E.S., R.R.D., A. Harkess, P.G.W., A. Healey, F.W.L., P.S., A.B., L.G., F.C., T.T., Y. Wang, G.T., F.X., J.H.L.-M., C.Z.-C., B.A.A., J.S., P.S.S., D.E.S., Z.L., J.T.L., M.S.B. and Z.-H.C. conducted data analysis. Z.-H.C., J.H.L.-M., D.W.S., B.A.A., J.S., P.S.S., D.E.S., D.X., Y. Wu, F.D., S.H., H.W., S.X. and H.D. sourced the funding to support the project. Z.-H.C., D.B.M., J.H.L.-M., J.S., P.S.S. and D.E.S. wrote the article with contributions of all authors.

Competing interests

The authors declare no competing interests.

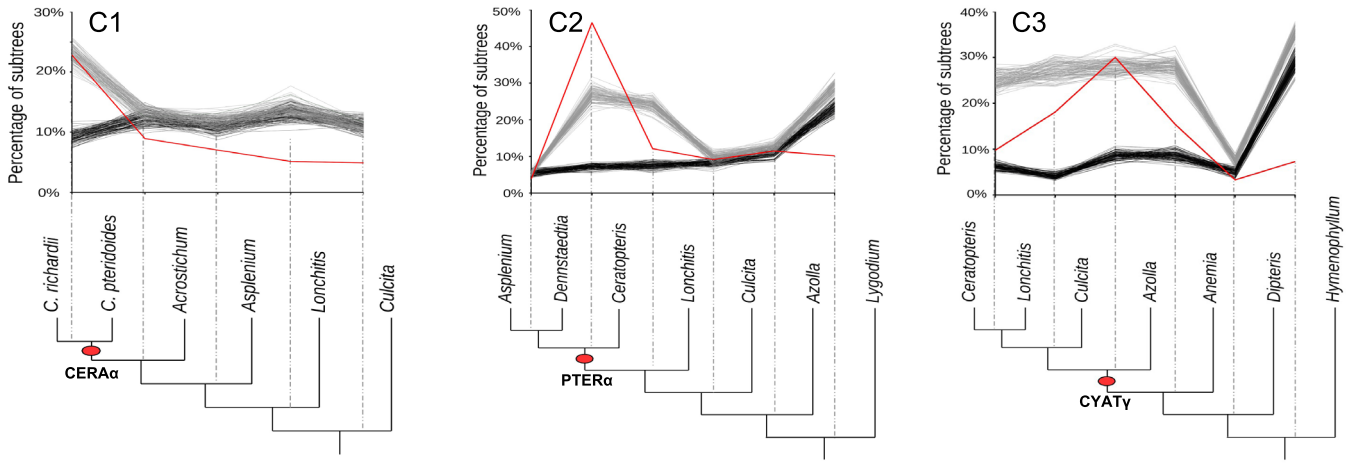
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41477-022-01226-7.

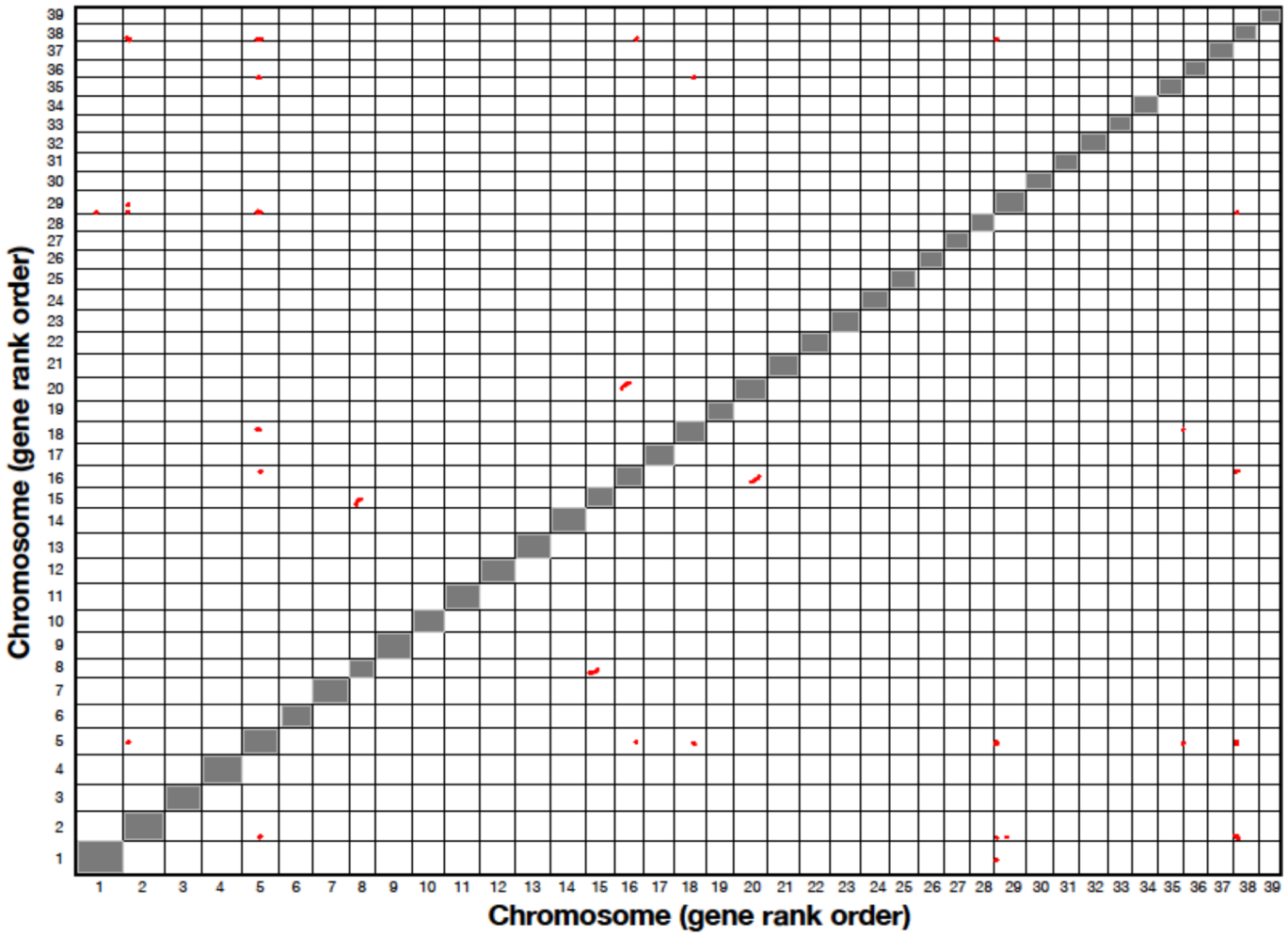
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-022-01226-7.

Correspondence and requests for materials should be addressed to D. Blaine Marchant, James H. Leebens-Mack, Pamela S. Soltis, Douglas E. Soltis or Zhong-Hua Chen.

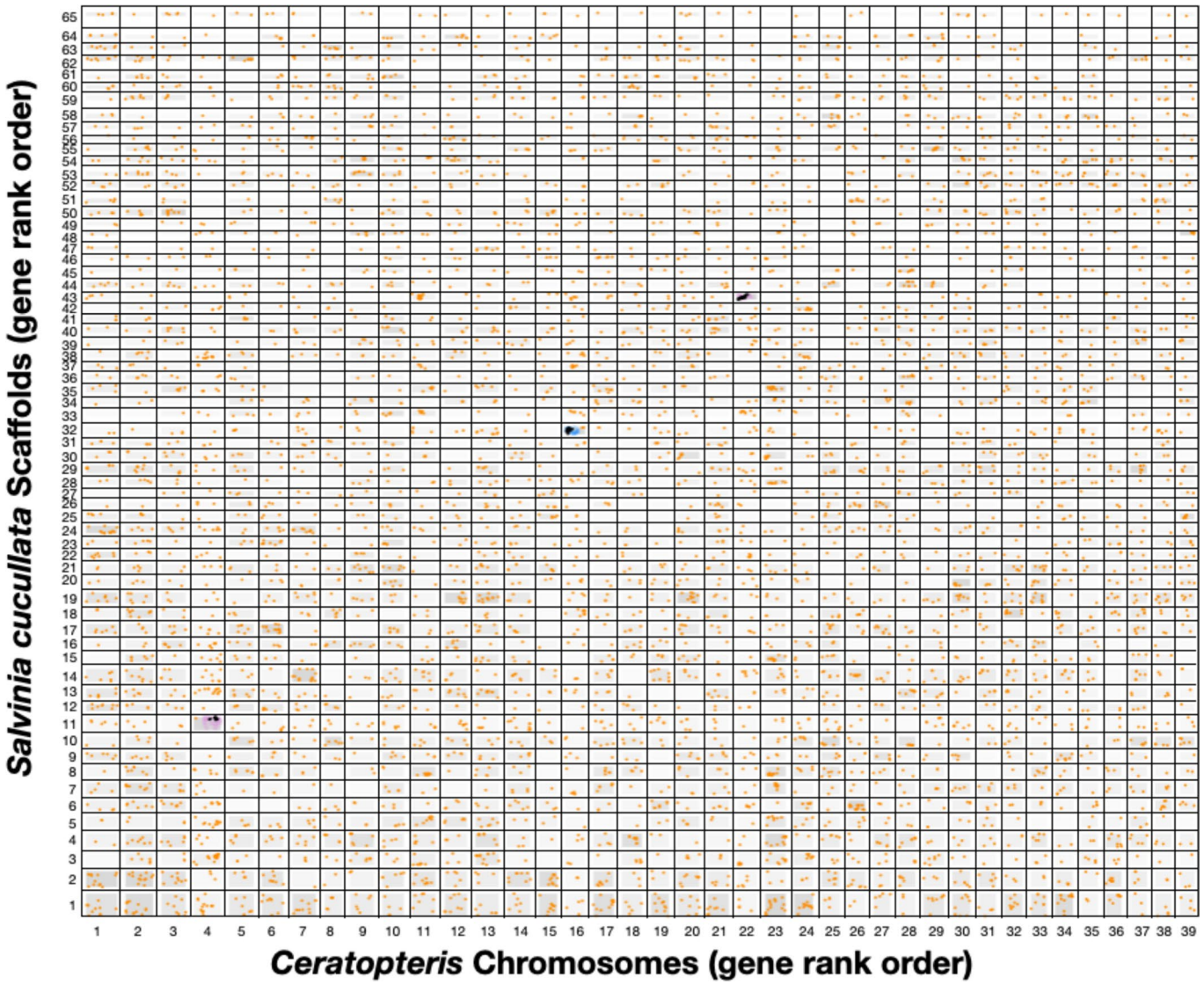
Peer review information *Nature Plants* thanks Jo Ann Banks and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



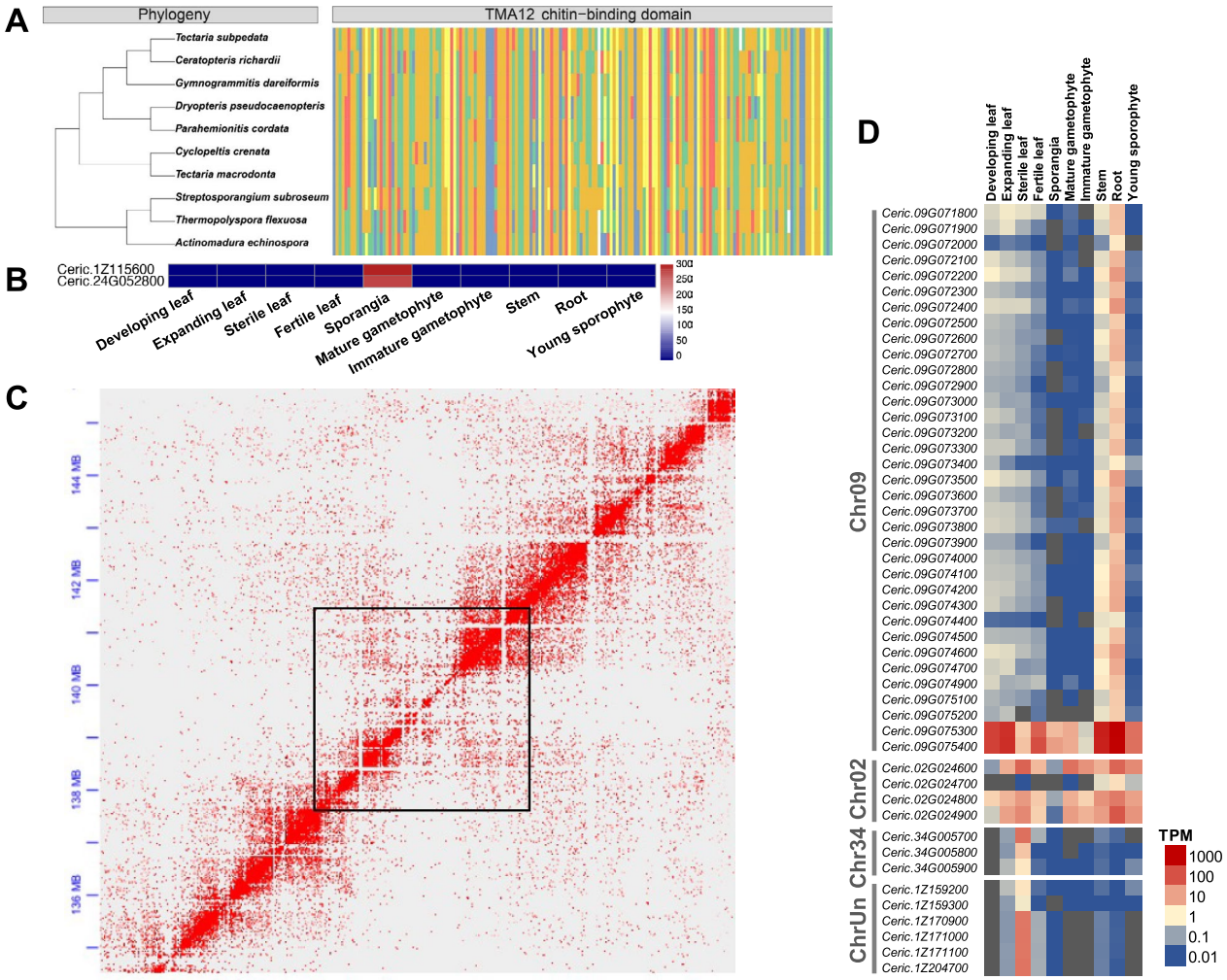
Extended Data Fig. 1 | MAPS; observed results and null and positive simulations. Percentage of subtrees that contain a gene duplication shared by descendant species at each node, results from observed data (red line), 100 resampled sets of null simulations (black lines), and positive simulations (gray lines). The red oval corresponds to the putative ancient WGD.



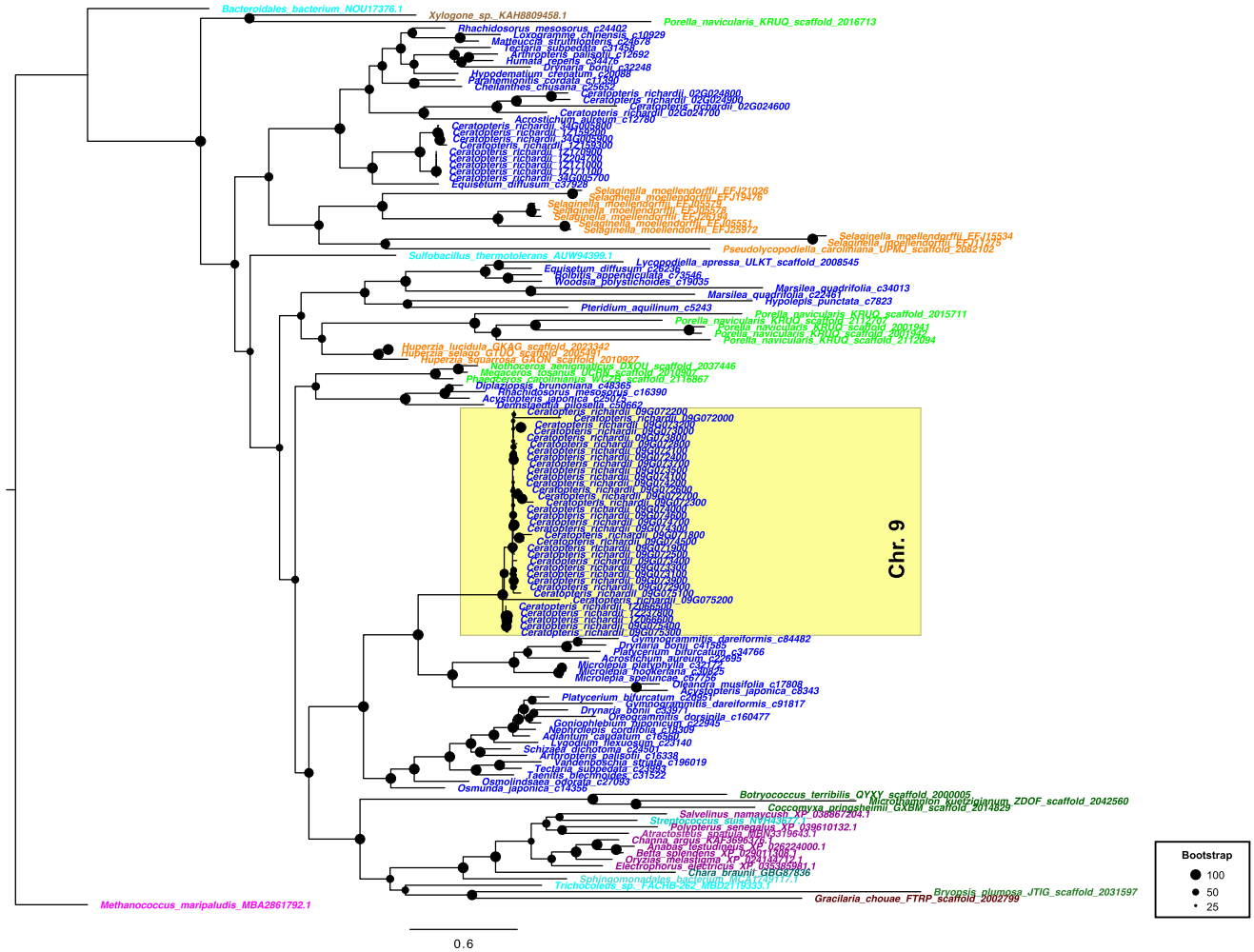
Extended Data Fig. 2 | Self-synteny analysis of *Ceratopteris* by chromosome (gene rank order). Syntenic blocks (≥10 genes) are identified in red.



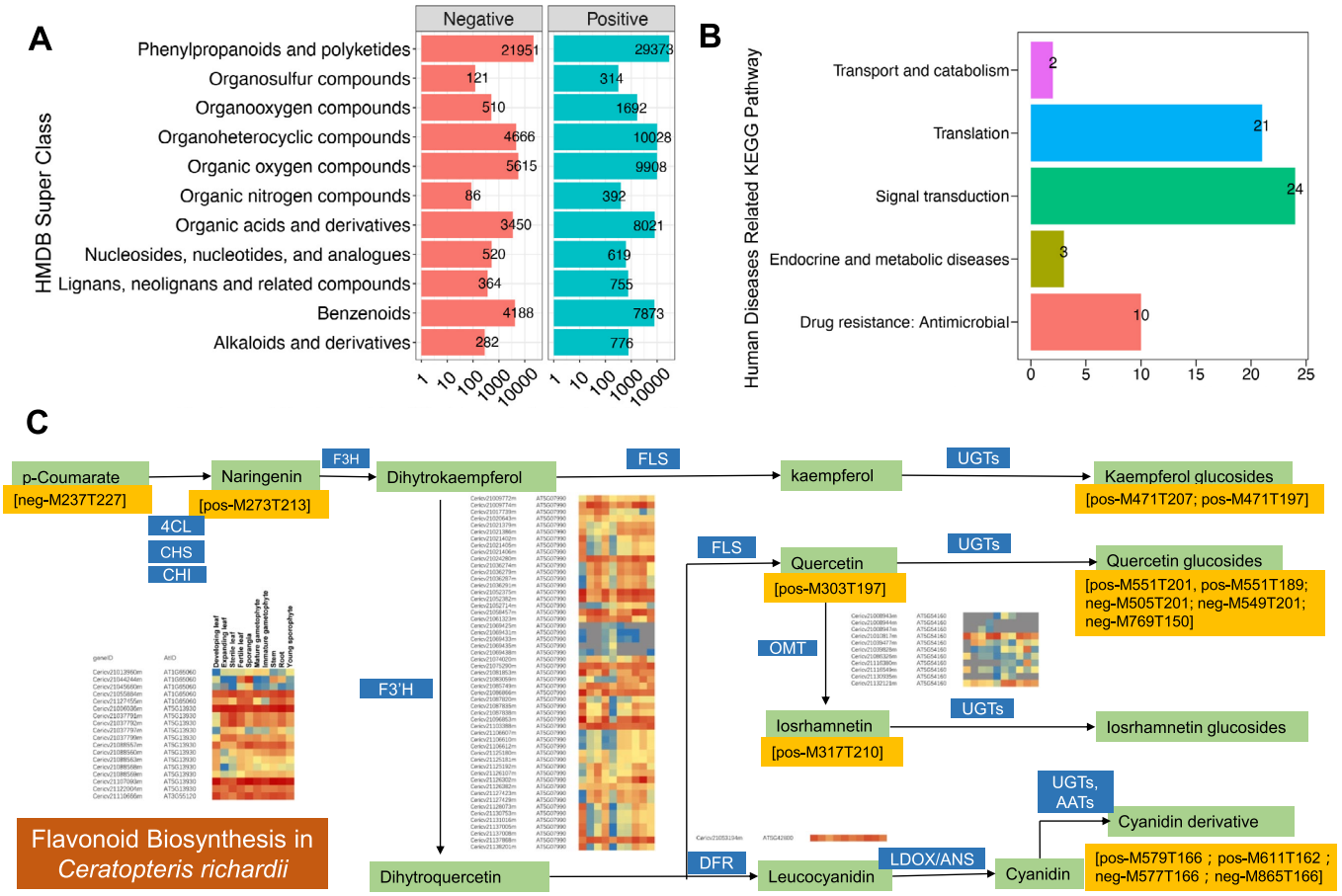
Extended Data Fig. 3 | Synteny analysis of *Ceratopteris* chromosomes and *Salvinia cucullata* scaffolds. Syntenic anchor BLAST hits are black points, nearby non-anchor BLAST hits are blue points, and non-syntenic reciprocal best score BLAST hits are orange points.



Extended Data Fig. 4 | Horizontal gene transfers (HGTs) and their evolution among ferns. A, Phylogeny of *Tma12* with protein conservation analysis of the chitin-binding domain. **B**, expression heatmap in tissues of *Ceratopteris*. **C**, HiC anchoring of Chromosome 9 with the region where the aerolysin-like genes are located outlined in black, confirming their presence in the *Ceratopteris* genome assembly. **D**, Expression patterns of the aerolysin-like genes in *Ceratopteris*.



Extended Data Fig. 5 | Transcriptome-based phylogeny of aerolysin-like genes. Two horizontal gene transfers are evident, one fern-specific and one in early land plants. Aerolysin-like genes that were tandemly duplicated on Chromosome 9 are highlighted. Fern genes are blue, lycophyte genes are orange, bryophyte genes are light green, fish genes are purple, bacteria genes are cyan, fungal genes are brown, archaea genes are magenta, streptophytic algae genes are black, chlorophytic algae genes are dark green, and red algae genes are rust.



Extended Data Fig. 6 | Analysis of *Ceratopteris* genes and metabolites for applications in environment and medicine. The metabolites identified in all 6 biological replications were used in this study. **A**, Overview of the metabolomics dataset of *Ceratopteris* based on HMDB (Human Metabolome Database, https://hmdb.ca). All compounds were divided into 11 clades according to the HMDB superclass. The compound numbers in each superclass were separated as positive and negative metabolites. **B**, KEGG enrichment of putative compounds related to human diseases. **C**, The Flavonoid Biosynthesis pathway in *Ceratopteris*. 4CL, 4-coumarate:CoA ligase; CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone 3-hydroxylase; F3'H, flavonoid 3'-hydroxylase; FLS, flavonol synthase; OMT1, O-methyltransferase; UGT, UDPdependent glycosyltransferase; DFR, dihydroflavonol 4-reductase; LDOX/ANS, leucoanthocyanidin dioxygenase/anthocyanidin synthase; AAT, anthocyanin acyltransferase.

nature research

Corresponding author(s): D. Blaine Marchant, James H. Leebens-Mack, Pamela S. Soltis, Douglas E. Soltis, Zhong-Hua Chen

Last updated by author(s): Jul 7, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see 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	Illumina and PacBio reads were sequenced at the Department of Energy (DOE) Joint Genome Institute (JGI) in Walnut Creek, California, USA, and the HudsonAlpha Institute for Biotechnology in Huntsville, Alabama, USA. Illumina reads were sequenced using the Illumina NovaSeq platform, and the PacBio reads were sequenced using the SEQUEL platform.
Data analysis	MECAT assembler v1.1 (Xiao et al., 2017) Canu assembler v1.8 (Koren et al., 2017) ARROW (Chin et al., 2013) JUICER (Durand et al., 2016) bwa mem (Li, 2013) GATK (McKenna et al., 2010) PERTRAN GSNAP (Wu and Nacu, 2010) PASA (Haas et al., 2003) EXONERATE (Slater and Birney, 2005) RepeatMasker (Smit et al., 2013) RepeatModeler (Smit and Hubley, 2008) FGENESH+ (Salamov and Solovyev, 2000) BLAT (Kent, 2002) HISAT2 (Kim et al., 2019) v2.2.1 samtools v1.11 (Li et al., 2009) Stringtie v2.1.4 (Pertea et al., 2015) ballgown v2.20.0 (Frazee et al., 2015) clusterProfiler v3.16.1 (Yu et al., 2012) Vennerable v3.0 (https://R-Forge.R-project.org/projects/vennerable/) pheatmap v1.0.12 (https://CRAN.R-project.org/package=pheatmap)

GENESPACE (https://code.jgi.doe.gov/plant/genespace-r)(Lovell et al., 2018)
MCScanX (Wang et al., 2012)
Orthofinder 2.4.0 (Emms and Kelly, 2015)
Wgd (Zwaenepoel and Van de Peer, 2019)
BLASTp (Camacho et al., 2009)
MCL (Altschul et al., 1997)
PAML (Yang, 2007)
ggplot2 v3.3.3(Wickham, 2016)
PASTA (Mirarab et al., 2015)
WGDgc (Rabier et al., 2014)
GenPhyloData (Sjöstrand et al., 2013)
NOTUNG v. 2.9.1.5 (Stolzer et al., 2012)
MAPS (Li et al., 2015)
RAxML v. 8.2.11 (Stamatakis, 2014)
Frackify (https://gitlab.com/barker-lab/frackify)
MASS R (Venables and Ripley, 2002)
Methylpy v. 1.4.2 (Schultz et al., 2015)
cutadapt v1.18 (Martin, 2011)
bowtie v2.4.1 (Langmead and Salzberg, 2012)
MAFFT (Katoh and Standley, 2013)
IQ-Tree (Minh et al., 2020)
HMMER 3.1b2 (hmmer.org)
AUGUSTUS 3.3.2 (Stanke et al., 2006)
BLAST+ (Altschul et al., 1990)
Probalign (Roshan and Livesay, 2006)
trimAl 1.2rev59t (Capella-Gutiérrez et al., 2009)
jModelTest 2 (Darriba et al., 2012)
Heatmapper (Babicki et al., 2016)
CAFE v5.0 (De Bie et al., 2006)
MUSCLE (Edgar, 2004)
pal2nal (http://www.bork.embl.de/pal2nal/)
Hyphy (http://hyphy.org)

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 Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

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

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

Raw genomic sequences have been deposited in the NCBI SRA under BioProject PRJNA729743. Genome and transcriptome assemblies and annotations can be found in Phytozome (https://phytozome-next.jgi.doe.gov/info/Crichardii_v2_1). Publicly available data was collected from Ensembl Plants (plants.ensembl.org), NCBI (ncbi.nlm.nih.gov), Swiss-Prot (uniprot.org), RepBase (girinst.org/repbase), One Thousand Plant Transcriptomes (1KP) database (OTPT Initiative, 2019), fern transcriptome database (Shen et al., 2018), water fern genomes (fernbase.org), spruce genome (congenie.org), TAIR (arabidopsis.org).

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 phylogenetic analyses, bootstrapping values were set to 100-1000, as per the norm in phylogenetics. For gene feature/genome characteristics (intron length, exon length, methylation levels, etc) analyses all features from the genome were incorporated unless otherwise noted. For MAPS, 3000 gene trees were simulated for each species tree with at least one tip per species: 1000 gene trees at the estimated λ and μ, 1000 gene trees at half of the estimated λ and μ, and 1000 trees at three times λ and μ.
Data exclusions	No data were excluded from the analyses.
Replication	Four biological replicates were used for each tissue sampled for RNA-seq. Six biological replicates were used for the metabolite analyses.
Randomization	The genomic, transcriptomic, methylomic, and metabolomic data were all generated from the Hn-n genotype of Ceratopteris richardii. or

Randomization	gene feature/genome characteristics (intron length, exon length, methylation levels, etc) analyses all features from the genome were incorporated unless otherwise noted.
Blinding	No blinding is necessary for genome assembly, characterization, comparisons, or evolutionary analyses.

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