Fern genomes elucidate land plant evolution and cyanobacterial symbioses


Ferns are the closest sister group to all seed plants, yet little is known about their genomes other than that they are generally colossal. Here, we report on the genomes of Azolla filiculoides and Salvinia cucullata (Salviniales) and present evidence for episodic whole-genome duplication in ferns—one at the base of ‘core leptosporangiates’ and one specific to Azolla. One fern-specific gene that we identified, recently shown to confer high insect resistance, seems to have been derived from bacteria through horizontal gene transfer. Azolla coexists in a unique symbiosis with N₂-fixing cyanobacteria, and we demonstrate a clear pattern of cospeciation between the two partners. Furthermore, the Azolla genome lacks genes that are common to arbuscular mycorrhizal and root nodule symbioses, and we identify several putative transporter genes specific to Azolla-cyanobacterial symbioses. These genomic resources will help in exploring the biotechnological potential of Azolla and address fundamental questions in the evolution of land plants.

The relatively small genome (0.75 Gb) of Azolla is exceptional among ferns, a group that is notorious for genomes as large as 148 Gb and averaging 12 Gb. Azolla is one of the fastest-growing plants on the planet, with demonstrated potential to be a significant carbon sink. Data from the Arctic Ocean show that, ~50 Myr ago, in middle-Eocene sediments, an abundance of fossilized Azolla characterizes an ~800,000-year interval known as the ‘Azolla event’. This period coincides with the shift from the early Eocene greenhouse world towards our present icehouse climate, suggesting that Azolla had a role in abrupt global cooling by sequestering atmospheric carbon dioxide. Azolla is also remarkable in harbouring an obligate, N₂-fixing cyanobacterium, Nostoc azollae, within specialized leaf cavities. Because of this capability, Azolla has been used as ‘green manure’ for over 1,000 years to bolster rice productivity in Southeast Asia. 

The advent of land plants ~474–515 Myr ago led to complex vegetational innovations that shaped emerging terrestrial and freshwater ecosystems. Bryophytes, lycophytes, ferns and gymnosperms dominated the global landscape before the ecological radiation of flowering plants 90 Myr ago. The first complete plant genome sequence (Arabidopsis thaliana) was published in 2000, followed by reference genomes for all other major lineages of green plants, except ferns. A dearth of genomic information from this entire lineage has limited not only our knowledge of fern biology but also the processes that govern the evolution of land plants.

The relatively small genome (0.75 Gb) of Azolla is exceptional among ferns, a group that is notorious for genomes as large as 148 Gb and averaging 12 Gb. Azolla is one of the fastest-growing plants on the planet, with demonstrated potential to be a significant carbon sink. Data from the Arctic Ocean show that, ~50 Myr ago, in middle-Eocene sediments, an abundance of fossilized Azolla characterizes an ~800,000-year interval known as the ‘Azolla event’. This period coincides with the shift from the early Eocene greenhouse world towards our present icehouse climate, suggesting that Azolla had a role in abrupt global cooling by sequestering atmospheric carbon dioxide. Azolla is also remarkable in harbouring an obligate, N₂-fixing cyanobacterium, Nostoc azollae, within specialized leaf cavities. Because of this capability, Azolla has been used as ‘green manure’ for over 1,000 years to bolster rice productivity in Southeast Asia. 

The relatively small genome (0.75 Gb) of Azolla is exceptional among ferns, a group that is notorious for genomes as large as 148 Gb and averaging 12 Gb. Azolla is one of the fastest-growing plants on the planet, with demonstrated potential to be a significant carbon sink. Data from the Arctic Ocean show that, ~50 Myr ago, in middle-Eocene sediments, an abundance of fossilized Azolla characterizes an ~800,000-year interval known as the ‘Azolla event’. This period coincides with the shift from the early Eocene greenhouse world towards our present icehouse climate, suggesting that Azolla had a role in abrupt global cooling by sequestering atmospheric carbon dioxide. Azolla is also remarkable in harbouring an obligate, N₂-fixing cyanobacterium, Nostoc azollae, within specialized leaf cavities. Because of this capability, Azolla has been used as ‘green manure’ for over 1,000 years to bolster rice productivity in Southeast Asia.
The Azolla symbiosis is unique among plant–bacterial endosymbiotic associations because the cyanobiont is associated with the fern throughout its life cycle, being vertically transmitted during sexual reproduction to subsequent generations\(^{10}\). In all other land plant symbiotic associations, the relationship must be renewed each generation. The *Nostoc* symbiont is not capable of autonomous growth and its genome shows clear signs of reduction, with several housekeeping genes lost or pseudogenized\(^{11}\). With a fossil record that extends back to the mid-Cretaceous period, *Azolla* probably shares a ~100-Myr-old co-evolutionary relationship with *Nostoc*\(^{12}\).

To better understand genome size evolution in *Azolla* and its closely related lineages, we obtained genome size estimates for all five genera of Salviniales (Supplementary Table 1). We found them to be at least an order of magnitude smaller than any other fern species (Fig. 1a), and, most notably, the genome of *Salvinia cucullata*, which belongs to the sister genus to *Azolla*, is only 0.26 Gb, the smallest genome size ever reported in ferns. This unanticipated discovery afforded us the opportunity to include a second fern genome for comparison.

### Results

#### Genome assembly and annotation.

To gain insight into fern genome evolution, as well as plant–cyanobacterial symbioses, we sequenced the genomes of *A. filiculoides* (Fig. 1b) and *S. cucullata* (Fig. 1c) using Illumina and PacBio technologies. The assembled *Azolla* and *Salvinia* genomes have N50 contig size of 964.7 Kb and 719.8 Kb, respectively. The BUSCO (Benchmarking Universal Single-Copy Orthologs) assessment and Illumina read-mapping results indicate high completeness for both assemblies (Supplementary Fig. 1 and Supplementary Table 2). We identified 20,201 and 19,914 high-confidence gene models in *Azolla* and *Salvinia*, respectively, that are supported by transcript evidence or had significant similarity to other known plant proteins (Supplementary Figs. 1–3, Supplementary Table 3 and Supplementary Discussion). *Salvinia* genes are much more compact, with a mean intron length half of that in *Azolla* (Supplementary Fig. 1). In addition to introns, differences in the repetitive content explain some of the nearly threefold difference in genome size. *Azolla* has more of every major category of repeat, but 191 Mb of the 233-Mb difference in the total...
Insights into gene family evolution in land plants. The genomes of Azolla and Salvinia offer a new opportunity to examine the evolution of plant genes and gene families across all Viridiplantae (land plants plus green algae). We classified genes into orthogroups from 23 genomes (12 angiosperms, 2 gymnosperms, 2 ferns, 1 lycophyte, 2 mosses, 2 liverworts, 1 charophyte and 1 chlorophyte; Supplementary Table 5) and reconstructed the gene family evolution—gain, loss, expansion and contraction—across the green tree of life (Supplementary Fig. 5 and Supplementary Table 5). To investigate the origin of genes linked to seed development, we examined orthogroups containing 48 transcription factors that express exclusively in Arabidopsis seeds[12]. Homologues of 39 of them were detected in ferns or other seed-free plants, indicating that many seed transcription factors were present before the origin of seeds (Supplementary Table 6). Similarly, only a handful of transcription factor families arose along the branch that led to seed plants (Supplementary Table 7); rather than relying on entirely novel transcription factors, it seems instead that an expansion of pre-existing transcription factor families had a greater role in seed plant evolution[13]. Indeed, ancestral gene number reconstructions of MADS-box transcription factors had a greater role in seed plant evolution than does Salvinia plus seed plants; Supplementary Table 5).

In a recent study on the evolution of plant transcription-associated proteins, which include transcription factors and transcriptional regulators[2], ferns were exclusively represented by the Pteridium aquilinum transcriptome. The finding that the transcriptional regulator Polycomb group EZ (PcG_EZ) was apparently secondarily lost in Salviniales (Supplementary Fig. 4). DNA transposon profiles are similar for the two ferns except that Azolla has substantially more SOLA elements than does Salvinia (Fig. 1e).

The history of whole-genome duplication in ferns. Our MultitAXon Paleopolyploidy Search (MAPS) phylogenomic analyses of the Azolla and Salvinia genomes (Fig. 3a), together with all available transcriptome data from other ferns, support two whole-genome duplication (WGD) events: a recent WGD event occurring in Azolla following its divergence from Salvinia and an earlier WGD predating the origin of ‘core leptosporangiates’ (sensu Pryer et al.[21]), a large clade comprising the heterosporous, tree and polypod ferns. The observed peaks of duplication associated with the inferred WGDs exceeded the 95% confidence intervals of our birth and death simulations for gene family evolution in the absence of WGDs. This high number of shared gene duplications is readily explained by a significant episodic birth event, such as a WGD. The discovery that Azolla experienced a genome duplication independent of other heterosporous ferns is not entirely surprising because Azolla has nearly twice the number of chromosomes of other heterosporous ferns, including Salvinia and Pilularia[30,31] (Fig. 1a).

To further substantiate the two WGD events identified by MAPS, we examined the distribution of synonymous distances ($K_s$) between syntenic paralogues within each of the genomes, as well as syntenic orthologues between Azolla and Salvinia. In the Azolla genome, we detected 242 syntenic blocks comprising 988 syntelog pairs. By contrast, only 83 syntenic blocks with 254 syntelog pairs could be found in Salvinia. Between Azolla and Salvinia, 3,587 pairs of syntenic orthologues were detected, clustering into 356 syntenic genomic blocks. We fit Gaussian mixture models to identify peaks in the distribution of synonymous distances ($K_s$) (Fig. 3b and Supplementary Fig. 7). The main peak for Azolla–Salvinia orthologous pairs centres at $~1.0$, which marks the species divergence between the two genera. To the left of this peak is the major Azolla intragenomic peak ($~0.8$), whose position confirms the Azolla-specific WGD event (Fig. 3b). To the right of the Azolla–Salvinia divergence peak is the Salvinia intragenomic $K_s$ peak ($~1.2$–1.3), which matches a minor Azolla intragenomic peak, consistent with the proposed pre-core leptosporangiates WGD (Fig. 3b).

Moreover, despite the antiquity of the WGDs and species divergence (Fig. 1a), we were still able to detect Azolla–Salvinia syntenic regions in a 2:1 or 2:2 syntenic relationship (Fig. 3c), respectively, corroborating the Azolla-specific and the older WGD events. The confirmation of these two WGDs in ferns further allows us to characterize patterns of gene retention following WGD. We found that Azolla syntenic paralogues are enriched for transcription-related genes (Supplementary Table 9), similar to what was observed in Arabidopsis and other angiosperms[32]. Likewise, protein kinases, another functional category commonly retained after WGD in seed plants, are significantly enriched in Salvinia syntenic paralogues (Supplementary Table 9). Additional genomic data are needed to better characterize the distribution of WGD events across the fern tree of life and to compare patterns of post-WGD gene fractionation with those documented in seed plants.

The pentatricopeptide repeat family and RNA editing. The pentatricopeptide repeat (PPR) family is the largest gene family found in the Azolla and Salvinia genomes, with the Azolla genome encoding over 2,000 PPR proteins and the Salvinia genome over 1,700 PPR proteins. PPRs are implicated in organellar RNA processing[33, and
the large repertoire of PPRs correlates well with the extensive RNA editing we observed in the organellar genomes of Salviniales: 1,710 sites in Azolla organelles and 1,221 sites in Salvinia (Supplementary Table 10). These editing events include both C-to-U conversions (~70%) and U-to-C conversions (~30%). The number of PPR genes and the degree of RNA editing greatly exceed that found in seed plants and most bryophytes26. Of the sequenced plant genomes, only that of S. moellendorffii15 has more PPR genes7, correlating with the hyperediting seen in lycophytes8. However, there are no U-to-C editing events in Selaginella, making the Azolla and Salvinia genome sequences a novel and valuable resource for identifying the unknown factors catalysing these events.

More than half of the plastid transcripts and two-thirds of the mitochondrial transcripts in Azolla and Salvinia require start codon creation by C-to-U editing or stop codon removal by U-to-C editing before translation is possible. Most stop codon edits (76%) and start codon edits (62%) are shared between Azolla and Salvinia plastomes (as opposed to only 19% in internal ACG codons; Supplementary Fig. 10). This persistence of start and stop codon edits suggests that their loss is selected against, that is, creating the translatable sequence by RNA editing has an advantage over having it encoded by the genome. This argues that these particular RNA-editing events are not selectively neutral29 and supports editing as a control mechanism for gene expression in fern organelles.

Only ~55–60% of PPR proteins (1,220 in Azolla and 930 in Salvinia) contain domains associated with RNA editing in other plants. Although sufficient to account for the number of editing events observed (assuming each protein can specify one or a few sites as in other plants), this leaves a very large number of PPR proteins (~700 in Azolla and ~600 in Salvinia) with unknown functions. By comparison, flowering plants contain only 200–250 PPR proteins that lack editing domains.

**Origin and evolution of a fern insecticidal protein.** Ferns are remarkable for their high levels of insect resistance compared to flowering plants30. Recently, Shukla et al.31 isolated a novel insecticidal protein, Tma12, from the fern Tectaria macrodonta. Transgenic cottons carrying Tma12 exhibit outstanding resistance to whitefly, yet show no decrease in yields, demonstrating tremendous agricultural potential. Tma12 has a high similarity to chitin-binding proteins (Pfam PF03067), but its evolutionary origin is unknown. Here, we found a Tma12 homologue to be present in the Salvinia genome (henceforth ScTma12), as well as in a few 1,000 Plants (1KP)32 fern transcriptomes, but not in Azolla or any other publicly available plant genomes. Phylogenetic analyses position the fern ScTma12 sequences together with bacterial sequences, and are most closely related to the chitin-binding proteins from Chloroflexi (Fig. 4). We investigated whether this insecticidal protein was more likely a result of horizontal gene transfer (HGT) from bacteria to ferns or produced by fern-associated microorganisms. ScTma12 is in a 646,687-bp scaffold (Sacu_v1.1_s0099) and has a 247-bp intron. The genes upstream and downstream of ScTma12 are all clearly plant genes, and we found no abnormality in read-mapping quality, nor an abrupt change in read coverage (Supplementary Fig. 9), which together speak against the sequence being a contamination from a bacterial source. It has been argued that differential loss of genes in eukaryotes is the rule and gene acquisition by HGT rather rare34. The concerted loss of Tma12 in each of the other Viridiplantae lineages is unlikely but cannot entirely be ruled out. However, functional HGT into eukaryotes does occur34,35 and ScTma12 might represent such a case that contributed to the well-documented resistance of ferns against phytophagous insects.

**Azolla–cyanobacterial symbiosis.** To explore the co-evolutionary history of the Azolla–Nostoc symbiosis, we resequenced five other Azolla species and assembled each of their cyanobiont genomes. We then compared the cyanobiont phylogeny to the host species phylogeny and found a clear cospeciation pattern, with just one exception (the placement of Azolla caroliniana; Fig. 5a). Although such a pattern has been hinted at before36,37, we provide unequivocal
Fig. 3 | The history of WGD in Azolla and Salvinia. **a,** MAPS analysis identified two WGD events: one specific to Azolla (orange circle) and one predating the core leptosporangiates (green circle). The blue line illustrates the percentage of subtrees indicative of a gene duplication shared by the descendants at each node. The grey lines display the gene birth–death simulation results without WGD. The species divergence dates are from Testo and Sundue129. **b,** Density plots from fitting Gaussian mixture models to 

\( K \) distributions estimated from pairs of syntenic paralogues within the Azolla and Salvinia genomes, as well as of syntenic orthologues between Azolla and Salvinia. **c,** Examples of synteny between Azolla and Salvinia genomic regions. The left and right panels display a 2:1 and 2:2 syntenic relationship between Azolla and Salvinia regions, respectively. Each subpanel represents a genomic region in Azolla or Salvinia, with gene models on both strands shown above and below the dashed line. High-scoring sequence pairs (HSPs) in protein-coding sequences are marked by short vertical bars above the gene models. Selected HSP links between genomic regions are depicted as coloured lines crossing the subpanels, whereas others (for example, the HSP links between the two Azolla genomic regions in the left panel) are left out for clarity. Collinear series of HSPs across genomic regions indicates a syntenic relationship between the regions concerned. Genomic regions conserved in duplicate after the WGD that occurred prior to the divergence between Azolla and Salvinia should show a 2:2 syntenic relationship, whereas regions conserved in duplicate after the Azolla-specific WGD should show a 2:1 syntenic relationship with Salvinia regions. The left and right panels can be regenerated at https://genomevolution.org/r/ujll and https://genomevolution.org/r/ukys, respectively.

evidence from whole-genome data. The genetic basis for this persistent symbiosis is undetermined. In plants, two other mutualistic associations—the arbuscular mycorrhizal (AM) and the nitrogen-fixing root nodule (RN) symbioses—have been well characterized. Whereas the AM symbiosis is formed between almost all land plants and a single fungal clade (Glomeromycota), the RN symbiosis is restricted to a few angiosperm lineages (mostly legumes) that associate with various nitrogen-fixing bacterial symbionts (for example, *Rhizobium* and *Frankia*). Despite these distinct differences, both symbioses require that a common symbiosis pathway (CSP) be established. This pathway is highly conserved in all land plants, except for those that have lost the AM symbiosis, such as *A. thaliana* and three aquatic angiosperms.

We investigated whether the CSP might have been co-opted during the evolution of the Azolla–Nostoc symbiosis by searching for six essential CSP genes in the Azolla and Salvinia genomes, as well as in transcriptomic data from other ferns in the 1KP data set (Supplementary Table 11). Although *DMI2* (also known as *SYM1*), *DMI3* (also known as *CgAMK*), *IPD3* (also known as *CYCLOPS*) and *VAPYRIN* were found in other ferns, the Azolla and Salvinia genomes completely lacked orthologues (Fig. 5b). *IPD3* and *VAPYRIN* do not belong to multigene families and homologues were not detected. Although homologues of *DMI2* and *DMI3* were identified, phylogenetic analyses confirmed that they are not orthologous to the symbiotic genes (Supplementary Data). In addition, for *DMI3*, we searched the Azolla and Salvinia homologues for two motifs (threonine 271 and the calmodulin-binding domain) that are critical for symbiosis. Both motifs are missing from these sequences, confirming the absence of *DMI3*. *CASTOR* and *POLLUX* are paralogues resulting from a gene duplication event in the ancestor of seed plants, and although pre-duplicated homologues are present in Salvinia and other seed-free plants, they are absent in Azolla (Fig. 5b). The co-elimination of the CSP genes suggests the lack of AM symbiosis in Azolla and Salvinia and that the nitrogen-fixing Azolla–Nostoc symbiosis does not rely on this pathway.

To identify genes important for the Azolla–Nostoc symbiosis, we treated *A. filiculoides* with erythromycin to remove the cyanobiont (*AzCy*) and compared its gene expression patterns with the wild type (*AzCy*). Experiments were carried out in conditions where the nitrogen nutrient (ammonium nitrate) was either supplied (N+) or withheld (N−) from the growth media. Results from *nifH* real-time PCR confirmed the complete absence of cyanobacteria in *AzCy*− and showed that the addition of the nitrogen nutrient suppresses symbiotic N\(_2\) fixation in *AzCy*+ (Supplementary Fig. 10),
consistent with an earlier study48. A large portion of the transcriptome is affected by the presence or absence of cyanobionts, with 6,210 and 2,125 genes being differentially transcribed under N− and N+ conditions, respectively (Fig. 5c and Supplementary Discussion). Of these, over 33% have at least a twofold expression difference. In response to nitrogen starvation, the Azolla transcriptomes remained moderately stable when the cyanobiont was present, but shifted drastically once it was absent (Fig. 5d). This finding suggests that the presence of the cyanobiont buffers the transcriptomic profile of Azolla from fluctuations in environmental nitrogen availability.

We focused primarily on those genes that are differentially expressed between the nitrogen treatments when the cyanobiont is present, and to a lesser extent on when the cyanobiont is absent (Fig. 5e and Supplementary Discussion). Because the cyanobacterial N2-fixation rate is strongly induced in the N− condition, we expect these genes to be candidates involved in nutrient exchange or in communication with the cyanobiont to promote Nfixation. A total of 88 upregulated and 72 downregulated genes were identified (Fig. 5e). Among the upregulated genes is a parologue of the ammonium transporter 2 subfamily (AfAMT2-4; Azfi_s0034. g025227; Fig. 5e and Supplementary Fig. 11) that is probably dedicated to ammonium uptake from the Azolla leaf cavity where the cyanobiont resides; homologous ammonium transporters have been implicated to participate in the AM and RN symbioses49,50. There is also a parologue of the molybdate transporter gene family (AfMOTI; Azfi_s0167.g054529) that is most likely specialized for supplying molybденum, a required co-factor for nitrogenase, to the cyanobiont. One of the legume MOTI genes was recently found to facilitate nitrogenase activity in RN symbiosis51. In addition to these two transporters, we identified a chalcone synthase paralogue among cyanobacteria sequences (Supplementary Fig. 12). Although these genes were recently found to facilitate nitrogenase activity in RN symbiosis51, and the thickened branches indicate B > 70. The tree is rooted based on the result from a broader phylogenetic analysis of PF03067 and PF08329 (Supplementary Data). The pink star denotes the sequence from the S. cucullata genome.

![Fig. 4 | Origin of a fern insecticidal protein. Phylogenetic analysis of the chin-binding domain Pfam PF03067 shows that the fern Tma12 insecticidal protein was probably derived from bacteria through an ancient HGT event. The numbers above the branches are bootstrap values (B5 = 100) support values (B5 = 100) and the blackened branches indicate B > 70. The pink star denotes the sequence from the S. cucullata genome.](image-url)

### Notes
- **Fig. 4 | Origin of a fern insecticidal protein.** Phylogenetic analysis of the chin-binding domain Pfam PF03067 shows that the fern Tma12 insecticidal protein was probably derived from bacteria through an ancient HGT event. The numbers above the branches are bootstrap values (B5 = 100) support values (B5 = 100) and the blackened branches indicate B > 70. The pink star denotes the sequence from the S. cucullata genome.
no homologue can be found in seed plants or in green algae, the SHC is also present in bryophyte (moss and liverwort) genomes and transcriptomes. Interestingly, these bryophyte SHCs are not related to those of ferns but are embedded in other bacterial SHC lineages (the monophyly of land plant SHCs is rejected by the Swofford–Olsen–Waddell–Hillis test \( P < 0.005 \)). This finding implies a complex evolutionary history for SHCs in land plants, possibly featuring independent transfers of SHC from different prokaryotic lineages to mosses, liverworts and ferns. We are confident that these SHC genes are not from contaminants because the gene phylogeny largely mirrors the species phylogeny; furthermore, the SHC genes were not assembled into stray scaffolds in the genomes of \( A. \) \textit{mexicana} and \( A. \) \textit{nilotica} are absent from the plastome tree. Scale bars represent 0.01 substitutions per site.

Methods
Flow cytometry and genome size estimation. To estimate the genome sizes of \( S. \) \textit{cucullata}, \( P. \) \textit{americana}, \( R. \) \textit{diphylleum} and \( M. \) \textit{minuta} (Supplementary Table 1), we used the Beckman chopping buffer to extract nuclei from fresh leaves, following the protocol of Kuo and Huang. The nuclei extractions were mixed with those from standards, stained with 1/50 volume of propidium iodide solution (2.04 mg ml\(^{-1}\)) and incubated at 4°C in darkness for 1 h. For each species, three replicates were performed on the BD FACScan system. For \( S. \) \textit{cucullata}, we used \( A. \) \textit{thaliana} (0.165 pg per C) as the standard, and for all other samples, we used \( Z. \) \textit{mays} CE-777 (2.785 pg per C). For each peak (in both standard and sample), over 1,000 nuclei were collected with cross-validation. We used a formula of:
(0.66 × (F − S2n) + 0.33 pg × (S4n − F)) / (S4n − S2n). For all other samples, we used: 5.57 pg × F/S2n, where 0.66, 0.33 and 5.57 pg are the 4C-value of A. thaliana, the 2C-value of A. thaliana and the 2C-value of Z. mays CE-777, respectively. S2n, S4n and F are the relative fluorescence amount (that is, the peak mean value) of the standard 2n nuclei, standard 4n nuclei and the sample 2n nuclei, respectively.

**Genome and transcriptome sequencing.** _A. filiculoides_ was collected from the Galgenwaard ditch in Utrecht, the Netherlands, and propagated directly or sterilized as described in Dijkhuizen et al. _A. filiculoides_ (sterilized without cyanoantioxidant) DNA was extracted, then sequenced on PacBio RSII at 51x coverage and Illumina HiSeq2000 (100 bp paired-end; ~86x coverage; Supplementary Table 12) with library insert sizes of 175 bp and 340 bp. RNA sequencing (RNA-seq) data from _A. filiculoides_ of the Galgenwaard ditch used for annotation included the following RNA profiles: (1) at four time points during the die cycle of fern spores; (2) at multiple stages of development with or without 2 mM ammonium nitrate for 1 week; (2) of different reproductive stages comparing fern sporophytes, megasporocarps and megaspores collected at noon; (3) of roots treated with cytokinin, indole-3-acetic acid (IAA) or none; and (4) of sporophytes with or without cyanobacterial symbionts grown with or without ammonium nitrate for 2 weeks then collected at noon. Plant materials of _A. caroliniana_, _Azolla mexicana_, _Azolla microphylla_, _Azolla nilotica_ and _Azolla rubra_ were obtained from the International Rice Research Institute (Supplementary Table 1) and DNA was extracted by a modified cetrltrimethylammonium bromide (CTAB) procedure. Illumina libraries with a 500-bp insert size were prepared and sequenced on Illumina HiSeq2000 (100 bp paired-end; ~75x coverage; Supplementary Table 1). _S. cucullata_ was originally collected from Bangladesh and subsequently cultured at Taiwan Forestry Research Institute, Dr. Cecilia Koo Botanic Conservation Center and Duke University (Supplementary Table 1). Genomic DNA was purified using a modified CTAB procedure and sequenced on both PacBio RSII (10 SMRT cells; 46x coverage) and Illumina HiSeq2000 platforms (1 lane of 125 bp paired-end; ~20x coverage). Genomic DNA was used to construct plastid DNA libraries from the floating and submerged leaves was separately extracted using the Sigma Spectrum Plant Total RNA kit, each with three biological replicates. To examine patterns of RNA editing, one library per leaf type was prepared by the Illumina Ribozero Plant kit (that is, not poly-A enriched), whereas the other two were done by the Kapa Stranded mRNA-seq kit. These six RNA libraries were pooled and sequenced in one lane of Illumina HiSeq2000 (125 bp paired-end).

**Genome assembly.** We assembled the PacBio reads from _A. filiculoides_ and _S. cucullata_ genomes using PBcR, and the resulting drafts were then polished by Quiver (v2.2.31) blastn or blastp, and then refined using the functions est2genome and protein2genome from the splice-site aware alignment program Exonerate (v2.2.0). We included the previously described _A. 21_ and _S. cucullata_ RNA seq data from the floating and submerged leaves that were assembled de novo by Trinity (v2.0.6) and in a reference-guided manner using HISAT2 (v2.0.4) and StringTie (v1.2.2), except for nine libraries published in de Vries et al. for which only a reference-guided approach was used. All programs used default parameters, and Trinity was run with the additional ‘--optimصلا_1’ option. StringTie results were merged using StringTie --merge, combined with the Trinity output, and were purged of redundant sequences using the GenomeTools sequnq utility.

Putative centromere sequences were first identified by searching the genome assemblies with Tandem Repeat Finder to identify very high copy (>100 repeats) tandem repeats with a motif length in the range of 185–195 bp. Motif sequences were extracted from the Tandem Repeat Finder output and clustered using USEARCH. A single major cluster was identified for each species and the sequences were separately aligned using MAFFT. Multiple sequence alignments for each species were used to generate a profile HMM representing the putative centromere sequences. We next used hmmsearch to search the genome assemblies with regards to identity alignments to identify other similar high copy repeat profile HMMs. Genomic regions with significant HMM matches were identified and these regions were annotated in a GFF3 format.

**Gene prediction.** Protein-coding genes were predicted using MAKER-P (v2.3.8), and three MAKER-P iterations were performed (1) by masking and creation of initial gene models from transcript and homologous protein evidence; (2) refinement of initial models with SNAP (ab initio gene predictor trained on initial models); and (3) final models generated using SNAP and the ab initio gene predictor AUGUSTUS trained on gene models from the second iteration. Masking was performed by RepeatMasker (v4.0.5) using the previously described species-specific repeat libraries and the full RepBase v22.0 database. After masking, gene models were inferred from transcripts and homologous protein sequences by first aligning to the genomes using BLAST+ (v2.2.31) blastn or blastp, and then refined using the functions est2genome and protein2genome from the splice-site aware alignment program Exonerate (v2.2.0). We included the previously described _A. filiculoides_ and _S. cucullata_ transcriptome set of protein sequences consisting of the full Swiss-prot database (downloaded 18 June 2016), Amborella trichopoda v1.0 proteins, _A. thaliana_ TAIR10 proteins, _Chlamydomonas reinhardtii_ v5.5 proteins, _Oryza sativa_ v7.0 proteins and _Physcomitrella patens_ v3.3 proteins (from Phytozome). Gene models with an annotation edit distance (AED) score of <0.2 were used to train SNAP, which was used during the second iteration of MAKER-P SNAP was retrained for the final iteration using gene models from the second iteration with an AED score of <0.2 and a translated protein length of >200 amino acids. Prior to training AUGUSTUS, redundant sequences, defined as those sharing ≥70% sequence similarity in significant (E-value < 1 × 10−10) HSPs from an all-by-all blast alignment, were removed from the training set. Final non-redundant sets of 5,013 (_A. filiculoides_) or 6,475 (_S. cucullata_) gene models were used to train AUGUSTUS.

**Phylogenomic inference and placement of WGDs from nuclear gene trees.** To infer ancient WGDs, we used a gene-tree sorting and counting algorithm, implemented in the MAPS tool. We selected four species of heterosporous ferns (two _Azolla_, _O. salvinia_ and _O. philaradia_ and representatives of three other leptosporangiate lineages (_Blechnum_, _Lygodium_ and _Dieperia_). The MAPS algorithm uses a given species tree to filter collections of nuclear gene trees for subtrees consistent with relationships at each node in the species tree. Using this filtered set of subtrees, MAPS was run with additional parameters for the identification of gene duplications shared by descendant taxa. To infer and locate a potential WGD, we plotted the percentage of gene duplications shared by descendant taxa by node: a WGD will produce a large burst of shared duplications, appearing as an increase in the percentage of shared gene duplications.

We circumscribed and constructed nuclear gene family phylogenies from multiple species for each _MAPS_ analysis. We translated each transcriptome into amino acid sequences using the TransPipe pipeline. Using these translations, we performed reciprocal protein BLAST (blastp) searches among data sets for each _MAPS_ analysis using an E-value cut-off of 10−10. We clustered gene families from the _BLAST_ results with the default parameters and kept only gene families that contained at least one gene copy from each taxon in a given _MAPS_ analysis. We discarded the remaining OrthoFinder clusters. We used PASTA for automatic alignment and phylogenetic reconstitution of gene families, employing MAFFT for constructing alignments, MUSCLE for merging alignments and RAXML for tree estimation. The parameters for each software were the default options for PASTA. For each gene family phylogeny, we ran PASTA until we reached three iterations without an improvement in the likelihood score using a centroid breaking strategy. We used the best-scoring _PASTA_ tree for each multi-species nuclear gene family to infer and locate WGDs using _MAPS_.

For the null simulations, we first estimated the mean background gene duplication rate (\( \lambda \)) and the gene loss rate (\( \mu \)) with WGDge. Gene count data were obtained from OrthoFinder clusters associated with each species tree. \( \lambda = 0.0031 \) and \( \mu = 0.0039 \) were estimated using only gene clusters that spanned the root of their respective species trees, which has been shown to reduce biases in the maximum likelihood estimates of \( \lambda \) and \( \mu \). We chose a maximum gene family
size of 100 for parameter estimation, which was necessary to provide an upper bound for numerical integration of node states. We provided a prior probability distribution of 1.3 on the number of genes at the root of each species tree, such that ancestral gene family sizes followed a shifted geometric distribution with a mean equal to the average number of genes per gene family across species.

Gene trees were then simulated within each MAPS species tree using the GuestTreeGen program from GenPhyloData. We developed ultrametric species trees from the topological relationships inferred by the 1KP Consortium analyses and median branch lengths from TimeTree. For each species tree, we simulated 4,000 gene trees with at least one tip per species: 1,000 gene trees at the λ and μ maximum likelihood estimates, 1,000 gene trees at half the estimated λ and μ, 1,000 trees at three times λ and μ, and 1,000 trees at five times λ and μ.

Classification of syntenic duplicates and microsynteny analysis. To distinguish gene duplicates as syntenic or tandem, we used the SynMap tool from the CoGe platform, with default parameters and the Quota Align algorithm to merge syntenic blocks. Sets of syntenic paralogues or orthologues (defined by a collinear series of putative homologous genes) were extracted using the DAGChainer algorithm, whereas duplicates within ten genes apart in the same genome were defined as synteny regions (Supplementary Discussion). Results for within Azolla and Salvinia genome comparisons, as well as between Azolla and Salvinia, can be regenerated using the links https://genomeworkflow.org/rt/tozk, https://genomeworkflow.org/toz7 and https://genomeworkflow.org/e/toy, respectively. Microsynteny analyses were performed using the GEvo tool from CoGe, which was used in setting the default minimum number of conserved genes for two regions to be called syntenic. Non-coding regions were masked in the two genomes to include only the protein-coding sequences. The two example microsynteny shown in Fig. 5c can be regenerated at https://genomeworkflow.org/ri/afl and https://genomeworkflow.org/ri/uky.

Gaussian mixture model analysis of K estimates. Distances of K were obtained for all pairs of syntenic paralogous and orthologous genes using the CODEML program in the PAML package (version 4.8) on the basis of codon sequence alignments. We used the GY model with stationary codon frequencies empirically estimated by the F3 x 4 model. Codon sequences were aligned with PRANK (version 100701) using the empirical codon model (setting -codon) to align coding DNA, always skipping insertions (-F). Only gene pairs with K values in the range of 0.05–5 were considered for further analyses. Gaussian mixture models were fitted to the resulting frequency distributions of K values by means of the densityMclust function in the R mclust version 5.3 package.

The Bayesian information criterion was used to determine the best-fitting model for the data, including the optimal number of Gaussian components to classify the complete proteomes of 23 sequenced green plant genomes, such as Salviniaceae, euphyllophytes and vascular plants, are reported in Supplementary Table 5. For each species tree, we simulated 4,000 gene trees with at least one tip per species: 1,000 gene trees at the λ and μ maximum likelihood estimates, 1,000 gene trees at half the estimated λ and μ, 1,000 trees at three times λ and μ, and 1,000 trees at five times λ and μ.

Transcription-associated protein characterization. Transcription-associated proteins comprise transcription factors that bind in a sequence-specific manner to cis-regulatory DNA elements and transcriptional regulators that act via protein–protein interaction or chromatin modification. We conducted genome-wide, domain-based annotation of transcription-associated proteins according to previous studies. A total of 1,206 (6%, Azolla) and 983 (7%, Salvinia) proteins were sorted into families; this amount is similar to Selaginella but less than in gymnosperms or angiosperms (Supplementary Table 8).

PPR annotation. We conducted a targeted annotation for PPR genes because they are generally only weakly expressed and thus often lack transcriptome support. Open reading frames from the nuclear genome assemblies were translated into amino acid sequences using the “getorf” tool from the EMBOSS package with a minimum size restriction of 300 nucleotides. These open reading frames were searched for PPR motifs using the hmmsearch tool from the HMMER3 package. The PPR motif models and parameters used follow those of Cheng et al.

Motifs were assembled into full PPR tracts and the best model for each PPR was determined.

To study the prevalence and location of RNA editing, non-poly(A)-enriched RNA data were filtered to remove adapters, low-quality reads and reads with ≤25% Ns. Clean reads were aligned against the assembled plastid and mitochondrial genome assemblies using TopHat 2.0 (ref. 111). One of the inverted repeat regions in the plastid genomes was removed before mapping. Only uniquely mapped reads were retained as input for SAMtools to call mismatches between RNA and the corresponding DNA. Differences between corresponding RNA and DNA sequences were identified as the putative RNA-editing sites. The RNA-editing level was defined as the number of altered reads divided by the total mapped reads for each site.

Phylogeny of the insecticidal protein Tma12. We used BLASTP to search for Tma12 (Genbank accession: 3Q438776) homologues in Phytomyzoz, 1KP transcriptomes and the NCBI Genbank non-redundant protein database. Although Tma12 homologues are present in fern transcriptomes and in the S. cucullata genome, no significant hit was found in any other plant genomes or transcriptomes. In addition, the majority of the Tma12 protein is composed of a characteristic domain that begins at the PF03067 Pfam family domain signature and does not contain any plant genes but is predominantly represented in the genomes of Actinobacteria, insects and fungi. To trace the origin of fern Tma12 genes, we downloaded representative sequences containing PF03067 and PF08329 (as the outgroup) from UniProt and Genbank and reconstructed the phylogeny using IQ-TREE. We then used this preliminary phylogeny (Supplementary Data) to construct a more focused data set to narrow down the phylogenetic affinity of Tma12. PartitionFinder was used to infer the optimal codon partition scheme and substitution models, and RAxML was used for maximum likelihood phylogeny inference and to calculate bootstrap branch support.

Azolla phylogeny. From the resequencing data (Supplementary Table 12), we compiled both plastome and nuclear phylogenomic data sets to infer the Azolla species phylogeny. S. cucullata was used as the outgroup. For the plastome phylogeny, we concatenated nucleotide alignments from 83 protein-coding genes and used PartitionFinder to identify the optimal data partition scheme and the associated nucleotide substitution models. RAxML was used for maximum likelihood phylogeny inference and to calculate bootstrap branch support. For the nuclear data set, we focused on genes that, based on the gene family classification results, are single copy in both A. filiculoides and S. cucullata genomes. We used HybPiper to extract the exon sequences from each of the resequenced species. To minimize the number of orthologs instead of orthogroups, we then filtered out genes with more than two species missing or having an average sequence length shorter than 75% of the one in A. filiculoides. This resulted in a final data set of 2,108 genes. Sequence alignments and gene tree inferences were done in PASTA, with the default setting, except that RAxML was used to estimate the best tree on the final alignment. To infer the species tree from these gene trees, we used the multispecies coalescent method implemented in ASTRAL-III (v5.6.1)117. The tree topology from the plastome and nuclear data sets were identical, and all nodes received bootstrap support of 100 and a local posterior probability of 1.0.

Cyanobiont phylogeny. To compare the host and symbiont phylogenies, we assembled the cyanobiont genomes from five additional Azolla species (Supplementary Table 12) using the resequencing data generated from total DNAs, including sequences derived from both the host and the cyanobiont. To extract the cyanobiont genomes from each of the Azolla species, we first filtered out chloroplast sequences by using BWA (default parameters) to map the total clean DNA reads against each chloroplast genome reference. In this step, ~3–4% of the reads were filtered out, which is necessary to remove plastid ribosomal RNAs that are highly similar to ones in the cyanobionts. For each of the five Azolla species, we then mapped the filtered reads to the published cyanobiont reference (N. azollae 0708 isolated from A. filiculoides; Genbank accession: NC_014248) using BLAST (alignment criteria: E value ≤1×10−5, sequence identity ≥90% and an aligned coverage of ≥90%). Only the aligned reads were assembled by Mbitom (iterations = 5) using N. azollae 0708 (ref. 111) as a reference. Gene prediction for each assembled cyanobiont was performed by the Prodigal program. Transfer RNAs were predicted by RNAscan-SE using a bacterial tRNA gene structure model. The presence of rRNA sequences (gene number and structure) for each cyanobiont
was confirmed by mapping the rRNAs of *N. azollae* 0708 against each assembled genome cyanobiont sequence using BLAST. We used mugy2 to generate the whole-genome alignment, which resulted in a nucleotide matrix of 5,354,840 characters. IQ-TREE was used for model testing and maximum likelihood tree inference. Because the *N. azollae* genome is reduced in size and is significantly diverged from other cyanobacteria, we could not find an appropriate outgroup to root the cyanobiont tree. To overcome this, we used STRIDE implemented in OrthoFinder to locate the root by reconciling gene trees. STRIDE was run with the default setting, except that MAFFT was used for alignment and RAxML for tree inference. The root was found to be the node placing the *A. nilotica* cyanobiont as sister to a clade comprising all other cyanobionts. The reconciled species tree is identical to the tree reconstructed from the whole-genome alignment.

**Identification of the CSP genes.** The *Medicago truncatula* DM12, DM13, IPD3, CASTER/POLLUX and VAPYRIN sequences were used as queries, as in a previous study10, to search against the genomes and transcriptomes from species listed in SUPplementary Table 11 using tBLASTn. For liverworts and ferns from the 1KP data set, non-annotated transcriptomes were used as targets, with the longest open reading frame of each contig extracted and translated. For *A. filiculoides* and *S. cucullata*, both the annotated gene models and the unassembled scaffolds were used. All hits that matched already annotated gene models were discarded prior to subsequent analyses. No homologues were identified in the two fern genomes for IPD3 and VAPYRIN. Protein sequences for DM12/SYMK, DM13/CaMK and CASTER/POLLUX were aligned using MAFFT11. The best substitution model for each alignment (JTT for all alignments) was determined using MEGA612. Phylogenetic trees were generated using RAxML13 on the CIPRES platform. Phylogenetic trees were generated using RAxML on the CIPRES platform and node support was assessed with 100 rapid bootstrap pseudoreplicates.

**Quantitative real-time PCR of nifH.** Quantitative real-time PCR for the *N. azollae* nifH gene was conducted using total RNA extracted from *A. filiculoides*. Primers were designed from Brown et al.14. Transcript levels for each species were quantified using the *Azolla* genome by HISAT2, and read counts for each gene were calculated using the HTSeq software package15. To identify cyanobiont-derived genes in the *Azolla* genome; 340 genes and cyanobacteria. For this, we used the *Arabidopsis* genome as a query for *A. filiculoides* genome by HISAT2, and read counts for each gene were calculated using the HTSeq software package. We used the rlog function in the DESeq2 package for data normalization and carried out differential expression analysis in DESeq2 to identify upregulated and downregulated genes with an adjusted P value of 0.005. Distance clustering and principal component analysis were used to examine the relatedness of samples and conditions as a quality-control measure.

**Azolla-cyanobacteria HGT.** To identify cyanobiont-derived genes in the *A. filiculoides* genome, we first investigated a potential orthologous relationship between any *Azolla* genes and cyanobacteria. For this, we used the *Azolla* genome assembly as a query for *D. salina* BLASTx against a protein data set of 11 cyanobacterial genomes. This resulted in 30,312 *Azolla* genome contigs hitting 8,779 different cyanobacterial proteins that were used as a query in a BLASTn against the *Azolla* genome; 340 *Azolla* contigs had reciprocal hits. To investigate whether these represent possible *Nostoc*-to-*Azolla* transfers or just examples of plastid-to-nucleus transfers, we used the 340 *Azolla* contigs for another BLASTx against the cyanobacteria and extracted all 51,743 BLASTx-aligned *Azolla* sequences. These highly redundant protein sequences were used for a DIAMOND BLASTp against the non-redundant dataset of NCBI. Almost all of the sequences had streptophyte proteins as the top hit, and when not, phylogenetic analysis clearly placed them within streptophytes.

**Phylogeny of SHC.** Homologues of *SHC* and oxidosqualene cyclase were obtained by searching against Phytozome, IPK transcriptomes and the NCBI Genbank non-redundant protein database. Protein alignment was done in MUSCLE. We used iQ-TREE to find the best-fitting amino acid substitution model and infer the phylogeny using maximum likelihood. Bootstrap support was assessed with 1,000 pseudoreplicates. To test whether the monophyly of fern, lycophyte, moss and liverwort SHC could be rejected, we conducted a Swofford–Olsen–Waddell–Hillis test using SOWHAT. We compared the best maximum likelihood topology against the topology with all land plant SHC constrained to be monophyletic. SOWHAT was run with 1,000 replicates.

**Detection of SHC-synthesized triterpenes.** Freeze-dried *S. cucullata* biomass was Soxhlet extracted in a 9:1 DCM:MeOH mixture for 24 h. The total lipid extracts obtained were dried over Na2SO4 followed by evaporation of the solvent by a gentle stream of N2. Aliquots of the total lipid extracts were methylated with diazomethane to convert the acid groups into corresponding methyl esters, purified over a SiO2 column and silylated using bis(trimethylsilyl)trifluoracetamide (BSTFA) in pyridine at 60°C for 20 min to convert the hydroxy groups into the corresponding trimethylsilyl ethers. The total lipid extracts were on-column injected on a ThermoTrace GC Ultra Trace DSQ gas chromatography mass spectrometry (GC–MS) onto a CP-sil 5CB-fused silica column (30 m x 0.32 mm internal diameter, film thickness: 0.10 μm). The GC–MS was operated at a constant flow of 1.0 ml min–1. The GC oven was programmed starting at 70°C to rise to 130°C at a rate of 20°C per min and then to 320°C at a rate of 4°C per min, followed by an isothermal hold for 20 min.

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

**Data availability.** The genome assemblies and annotations can be found at www.fernbase.org. The raw genomic and transcriptomic reads generated in this study were deposited in the NCBI SRA under the BioProject PRJNA435027 and PRJNA430459. The sequence alignments and tree files can be found in the Supplementary Data.

Received: 9 February 2018; Accepted: 24 May 2018;
Published online: 02 July 2018

### References

20. Supplementary Data. www.fernbase.org. The raw genomic and transcriptomic reads generated in this study were deposited in the NCBI SRA under the BioProject PRJNA435027 and PRJNA430459. The sequence alignments and tree files can be found in the Supplementary Data.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41477-018-0188-8.
Reprints and permissions information is available at www.nature.com/reprints.
Correspondence and requests for materials should be addressed to F.-W. L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Reporting Summary

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

Statistical parameters

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

<table>
<thead>
<tr>
<th>n/a</th>
<th>Confirmed</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐</td>
<td>☑</td>
</tr>
<tr>
<td>☐</td>
<td>☑</td>
</tr>
<tr>
<td>☐</td>
<td>☑</td>
</tr>
<tr>
<td>☐</td>
<td>☑</td>
</tr>
<tr>
<td>☐</td>
<td>☑</td>
</tr>
<tr>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>☑</td>
<td>☑</td>
</tr>
</tbody>
</table>

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection | Provide a description of all commercial, open source and custom code used to collect the data in this study, specifying the version used OR state that no software was used.

Data analysis | We employed a number of software for data analyses in this study, which was described in detail in the materials and methods, including the parameters used, versions (if applicable), and citations.

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

April 2018
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

The genome assemblies and annotations can be found in www.fernbase.org. The raw genomic and transcriptomic reads generated in this study were deposited in NCBI SRA under the BioProject PRJNA430527 and PRJNA430459. The sequence alignments and tree files can be found in Supplementary Data.

Field-specific reporting

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

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

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

Life sciences study design

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

<table>
<thead>
<tr>
<th>Sample size</th>
<th>For phylogenetic analyses, bootstrapping datasets were sampled between 100 to 1000 times, which is the field standard.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data exclusions</td>
<td>Gene models without transcript or homology supports were excluded (see the supplementary discussion).</td>
</tr>
<tr>
<td>Replication</td>
<td>The RNA-seq experiments were done with three biological replicates per treatment.</td>
</tr>
<tr>
<td>Randomization</td>
<td>The plant cultures for RNA-seq were placed on the same growth chamber shelf, but the positions were randomized in terms of nutrient treatments and symbiont types.</td>
</tr>
<tr>
<td>Blinding</td>
<td>Blinding is not applicable in this study.</td>
</tr>
</tbody>
</table>

Reporting for specific materials, systems and methods

Materials & experimental systems

<table>
<thead>
<tr>
<th>n/a</th>
<th>Involved in the study</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐ Unique biological materials</td>
<td></td>
</tr>
<tr>
<td>☑ Antibodies</td>
<td></td>
</tr>
<tr>
<td>☑ Eukaryotic cell lines</td>
<td></td>
</tr>
<tr>
<td>☐ Palaeontology</td>
<td></td>
</tr>
<tr>
<td>☐ Animals and other organisms</td>
<td></td>
</tr>
<tr>
<td>☐ Human research participants</td>
<td></td>
</tr>
</tbody>
</table>

Methods

<table>
<thead>
<tr>
<th>n/a</th>
<th>Involved in the study</th>
</tr>
</thead>
<tbody>
<tr>
<td>☑ ChIP-seq</td>
<td></td>
</tr>
<tr>
<td>☐ Flow cytometry</td>
<td></td>
</tr>
<tr>
<td>☑ MRI-based neuroimaging</td>
<td></td>
</tr>
</tbody>
</table>

Unique biological materials

Policy information about availability of materials

Obtaining unique materials
The plant materials used in this study are available upon request (to F.-W. Li or H. Schluepmann).
Flow Cytometry

Plots

Confirm that:
☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
☐ All plots are contour plots with outliers or pseudocolor plots.
☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

We used flow cytometry to estimate the genome sizes of Pilularia americana, Regnellidium diphyllum, Marsilea minuta, Salvinia cucullata.

1. Prepare buffer for use.
   a. Allocate appropriate amount of Backmen stock buffer to a 50-ml tube based on an estimation of 1-1.5 ml per sample.
   b. Add 0.04 g PVP-40, 5 μl 2-mercaptoethanol, 1 μl RNase per ml of buffer.

2. Extract sample and standard nuclei by chopping leaf tissue
   a. Add 500 μl of buffer to a glass Petri dish.
   b. Add a (~400 mm²) piece of young leaf to the Petri dish, and chop it with a razor on ice until most tissue slices are less than 1 mm in size.
   c. Filter the chopped sample and standard into a 2.0-ml tube through a 30-μm nylon mesh.
   d. Add additional buffer to the sample, and ensure that the filtered leaf nuclei solution is greater than 500 μl in volume or more depending on need.

3. Staining nuclei solutions
   a. Mix sample nuclei and standard leaf nuclei solutions into a 500-μl volume in 2.0-ml tubes.
   b. Add 10 μl PI solution (2.04 mg/ml ) into each of mixed nuclei solutions.
   c. Incubate in the dark at 4 °C for 1 h for staining.

Recipes

Backmen stock buffer
1.0% Triton X-100
50 mM Na2SO3
50 mM Tris-HCl (pH 7.5)
ddH2O (the solvent)

Note: Store at 4 °C up to 1 year.

Instrument BD FACSCan system (BD Biosciences, Franklin Lake, NJ, USA)

Software BD FACSCan system (BD Biosciences, Franklin Lake, NJ, USA)

Cell population abundance

Pilularia americana:
Replicate 1: sample peak particle number = 1514, standard1 peak particle number = 1154.
Replicate 2: sample peak particle number = 1834, standard1 peak particle number = 1371.
Replicate 3: sample peak particle number = 1450, standard1 peak particle number = 1036.

Regnellidium diphyllum:
Replicate 1: sample peak particle number = 1222, standard1 peak particle number = 1737.
Replicate 2: sample peak particle number = 1180, standard1 peak particle number = 1613.
Replicate 3: sample peak particle number = 1137, standard1 peak particle number = 1759.

Marsilea minuta:
Replicate 1: sample peak particle number = 1892, standard1 peak particle number = 1118.
Replicate 2: sample peak particle number = 1850, standard1 peak particle number = 1209.
Replicate 3: sample peak particle number = 1892, standard1 peak particle number = 1227.

Salvinia cucullata:
Replicate 1: sample peak particle number = 1084, standard1 peak particle number = 1484, standard2 peak particle number = 1170.
Replicate 2: sample peak particle number = 1129, standard1 peak particle number = 1552, standard2 peak particle number = 1253.
Replicate 3: sample peak particle number = 1229, standard1 peak particle number = 1584, standard2 peak particle number = 1500.

Gating strategy

For particle acquisition, we set a threshold of FL2-H = 52 for the samples of Pilularia americana, Regnellidium diphyllum, and Marsilea minuta. For Salvinia cucullata, a threshold of FL2-H = 100 is applied.

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