Fern Genomes Elucidate Land Plant Evolution and Cyanobacterial Symbioses

Fay-Wei Li  
Cornell University

Paul Brouwer  
Utrecht University

Lorenzo Carretero-Paulet  
Ghent University

Shifeng Cheng  
BGI-Shenzhen

Jan de Vries  
Dalhousie University

Pierre-Marc Delaux  
Université de Toulouse

See next page for additional authors

Follow this and additional works at: https://digitalcommons.usu.edu/biology_facpub

Part of the Biology Commons, and the Ecology and Evolutionary Biology Commons

Recommended Citation

Li, Fay-Wei; Brouwer, Paul; Carretero-Paulet, Lorenzo; Cheng, Shifeng; de Vries, Jan; Delaux, Pierre-Marc; Eily, Ariana; Koppers, Nils; Kuo, Li-Yaung; Li, Zheng; Simenc, Mathew; Small, Ian; Wafuta, Eric; Angarita, Stephany; Barker, Michael S.; Bräutigam, Andrea; dePamphilis, Claude; Gould, Sven; Hosmani, Prashant S.; Huang, Yao-Moan; Huettel, Bruno; Kato, Yoichiro; Liu, Xin; Maere, Steven; McDowell, Rose; Mueller, Lukas A.; Nierop, Klaas G. J.; Rensing, Stefan A.; Robison, Tanner; Rothfels, Carl J.; Song, Erin M.; Song, Yue; Timilsena, Prakash R.; Van de Peer, Yves; Wang, Hongli; Wilhelmsson, Per K. I.; Wolf, Paul G.; Xu, Xun; Der, Joshua P.; Schlüpmann, Henriette; Wong, Gane K.-S.; and Pryer, Kathleen M., "Fern Genomes Elucidate Land Plant Evolution and Cyanobacterial Symbioses" (2018). Biology Faculty Publications. Paper 1596.  
https://digitalcommons.usu.edu/biology_facpub/1596

This Article is brought to you for free and open access by the Biology at DigitalCommons@USU. It has been accepted for inclusion in Biology Faculty Publications by an authorized administrator of DigitalCommons@USU. For more information, please contact dylan.burns@usu.edu.
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 symbiosis. These genomic resources will help in exploring the biotechnological potential of Azolla and address fundamental questions in the evolution of plant life.

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.

https://doi.org/10.1038/s41477-018-0188-8
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. 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. With a fossil record that extends back to the mid-Cretaceous period, Azolla probably shares a ~100-Myr-old co-evolutionary relationship with Nostoc.

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

---

Fig. 1 | Genome size evolution in Salviniales. a, Members of Salviniales have smaller genome sizes than other ferns (averaging 1C = 12 Gb). Two whole-genome duplication (WGD) events identified in this study were mapped onto the phylogeny, with divergence time estimates obtained from Testo and Sundue. b, c, Whole genomes were assembled from A. filiculoides (b) and S. cucullata (c). d, e, The genome of S. cucullata has substantially reduced levels of RNA (d) and DNA (e) transposons compared to A. filiculoides. Image in panel c courtesy of P.-F. Lu.
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 seeds15. 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 evolution14. Indeed, ancestral gene number reconstructions of MADS-interacting keratin-like and C-terminal (MIKC)-type MADS box genes (orthogroup 23) show that these important developmental regulators more than doubled in number from 15 in the ancestral vascular plant to 31 in the ancestral eukalyphytale (here, Salviniales plus seed plants; Supplementary Table 5).

In a recent study on the evolution of plant transcription-associated proteins, which include transcription factors and transcriptional regulators24, ferns were exclusively represented by the Pteridium aquilinum transcriptome. The finding that the transcriptional regulator Polycomb group EZ (PcG_EZ) was lost in ferns is corroborated here by our whole-genome data (Supplementary Table 8). Conversely, the transcription factor ULTRAPETALA, which originated at the base of eukalyphytales and is present in P. aquilinum, was apparently secondarily lost in Salviniales (Supplementary Table 8). YABBY, an important transcription factor that patterns leaf polarity in flowering plants, is absent in our fern genomes and in the genome of the lycophyte Selaginella moellendorfii15 (Supplementary Table 8). Interestingly, a YABBY homologue was recently identified in a separate lycophyte species—Huperzia selago16—suggesting that YABBY has been lost at least twice in land plant evolution (in Selaginella and in ferns). How the differential retention of YABBY shaped the evolution of the vascular plant body plan requires further studies.

Among the orthogroups specific to seed plants, 1-aminocyclopropane-1-carboxylic acid (ACC) oxidase is of special interest because it converts ACC to ethylene—the last step in the ethylene biosynthetic pathway (Fig. 2). Ethylene is a critical plant hormone that controls various important physiological responses (for example, fruit ripening, flowering time, seed germination and internode elongation). Because ethylene responses are known in bryophytes, lycophytes and ferns17, it is intriguing to find that ACC oxidase only evolved within seed plants, a result confirming that seed-free plants must possess an alternative ethylene-forming mechanism18. Two other mechanisms, found in bacteria and fungi, result in ethylene formation: one via the 2-oxoglutarate-dependent ethylene-forming enzyme and the other through the non-enzymatic conversion of 2-keto-4-methylthiobutyric acid (KMBA) into ethylene17. We did not identify ethylene-forming enzyme in seed-free plant genomes, suggesting the absence of the ethylene-forming enzyme-based biosynthetic pathway. Seed-free plants may possibly synthesize ethylene non-enzymatically via KMBA; however, further biochemical studies are needed to test this hypothesis. Interestingly, ACC synthase (upstream of ACC oxidase) is present in seed-free plants, albeit in a lower copy number (<3) compared to seed plants, which average 9.3 copies (Fig. 2 and Supplementary Fig. 6). Paralogues of ACC synthase in seed plants are differentially regulated in response to varying developmental or environmental stimuli19. Thus, it is plausible that the expansion of the ACC synthase family was coupled with the origin of ACC oxidase in seed plants to create a regulated ethylene biosynthetic pathway.

The history of whole-genome duplication in ferns. Our Multi-tAxon Paleopolyploidy Search (MAPS)20 phylogenetic 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 leptosporangiates21,22, 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 Pitularia23,24 (Fig. 1a).

To further substantiate the two WGD events identified by MAPS, we examined the distribution of synonymous distances (Ks) 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 Ks distributions (Fig. 3b and Supplementary Fig. 7). The main peak for Azolla–Salvinia orthologue 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 Ks 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 2:1 or 2:2 syntenic relationships (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 angiosperms25. 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 processing26, 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. moellendorfii27 has more PPR genes28, correlating with the hyperediting seen in lycophytes29. 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.

Fig. 2 | Evolution of ethylene biosynthesis. The ethylene-forming pathway involves the Yang cycle, where ACC is synthesized from S-adenosyl-methionine (SAM; also known as AdoMet) by ACC synthase. ACC oxidase then catalyses the conversion of ACC to ethylene. We found that ACC oxidase is unique to seed plants (green) and its origin probably drove the expansion of the ACC synthase gene family (orange; Supplementary Fig. 6) to create a regulated ethylene biosynthetic mechanism.

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) fern transcriptomes, but not in Azolla or any other publicly available plant genomes. Phylogenetic analyses position the fern Tma12 sequences together with bacterial sequences, and are most closely related to the chin-in-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 rare32. 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 occur33,34 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 before35,36, we provide unequivocal
Salvinia and Azolla were found in other ferns, the VAPYRIN (SYMRK), except for those that have lost the AM symbiosis40,41, such as Glomeromycota38, 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 plants95, except for those that have lost the AM symbiosis95[41], such as A. thaliana and three aquatic angiosperms96,97.

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 set92 (Supplementary Table 11). Although DMI2 (also known as SYMRK), DMI3 (also known as CCaMK), 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 families95 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 paralogs 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 N₂ fixation in AzCy+ (Supplementary Fig. 10), 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)38, 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 plants95, except for those that have lost the AM symbiosis95[41], such as A. thaliana and three aquatic angiosperms96,97.

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 set92 (Supplementary Table 11). Although DMI2 (also known as SYMRK), DMI3 (also known as CCaMK), 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 families95 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 paralogs 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 N₂ fixation in AzCy+ (Supplementary Fig. 10),
consistent with an earlier study. 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 two-fold 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 N fixation. 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 symbioses. There is also a parologue of the molybdate transporter gene family (AfMOT1; Azfi_s0167.g054529) that is most likely specialized for supplying molybdenum, a required co-factor for nitrogenase, to the cyanobiont. One of the legume MOT1 genes was recently found to facilitate nitrogenase activity in RN symbiosis. In addition to these two transporters, we identified a chalcone synthase paralogue among cyanobacteria sequences (Supplementary Fig. 12). Although the ancient and intimate nature of the Azolla–Nostoc relationship suggests that gene transfer from Nostoc to the Azolla nuclear genome may have occurred over time, a thorough homology search found no evidence of Nostoc-to-Azolla HGT. However, we did discover a cyanobacteria-derived gene in the Azolla genome, but one that is shared with other ferns. This gene encodes a squalene–hopene cyclase (SHC), which mediates the cyclization of squalene into hopene, and is thought to be the evolutionary progenitor of many classes of eukaryotic and prokaryotic sterol cyclases. We found SHC homologues in both the Azolla and Salvinia genomes, as well as in 40 fern 1KP transcriptomes. Our reconstructed gene phylogeny clearly shows that the fern SHCs are nested among cyanobacteria sequences (Supplementary Fig. 12). Although
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 test48, 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 host species phylogeny; furthermore, the SHC genes were not assembled into stray scaffolds in the genomes of Arabidopsis and Physcomitrella. The SHC phylogenies are also consistent with the SHC genes in other organisms, such as ferns but are embedded in other bacterial SHC lineages. This suggests that SHC genes have been transferred from different prokaryotic lineages to land plants, and that the SHC genes in land plants are not derived from a common ancestor. However, further studies are needed to confirm this hypothesis.

We anticipate that the availability of the first genomic data from ferns will continue to lead to vital insights into the processes that govern the evolution of plant genes and gene families. The implementation of fern data into the existing comparative genomic framework will enhance our understanding of the plant tree of life.

Methods

Flow cytometry and genome size estimation. To estimate the genome sizes of S. cucullata, P. americana, Regnellidium diphyllum and Marsilea minuta (Supplementary Table 1), we used the Beckman chopping buffer to extract nuclei from fresh leaves, following the protocol of Kuo and Huang56. 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. cucullata, we used A. thaliana (0.165 pg per C) as the standard, and for all other samples, we used Zea 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 and used to calculate the 2C-value of S. cucullata, we used a formula of: $2C = \frac{\text{peak area of sample}}{\text{peak area of standard}} \times 2C_{\text{standard}}$.
Articles

1. **Genome and transcriptome sequencing.** A *filioloides* was collected from the Galgenwaard ditch in Utrecht, the Netherlands, and propagated directly or sterilized as described in Dykhuizen et al. A *filioloides* (sterilized without cyanobactin) 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. DNA 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 diel cycle of fern spore germination; (2) of plastome regions with or without 2 mM ammonium nitrate for 1 week; (2) of different reproductive stages comparing fern sporophytes, microsporocarps and megasporocarps 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. azollae* (Azolla *sp.*), *A. nitolica* and *A. rubra* were obtained from the International Rice Research Institute (Supplementary Table 1) and DNA was extracted by a modified cetyltrimethylammonium bromide (CTAB) procedure, Illumina libraries with a 500-bp insert size were prepared and sequenced on Illumina HiSeq2000 (100 bp paired-end; ~50x coverage; Supplementary Table 2). *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; 2x Illumina sequencing Table) at Taiwan Forestry Research Institute, Dr. Cecilia Koo Botanic Conservation Center and Duke University (Supplementary Table 1). Genome assembly. We assembled the PacBio reads from *A. filiculoides* and *S. cucullata* genomes using PBcR (v.3.3), and the resulting drafts were then polished by Quiver (v.2.1.6) and BlobTools (v.2016.15) in combination with SILVA and UniProt databases to infer the taxonomy for each scaffold. We removed all scaffolds that were classified as sterilized as described in Dijkhuizen et al. Five million pairwise BLASTX (v.2.3.11) blastn or blastp, and then refined using the functions est2genome and protein2genome from the splice-site aware alignment program Exonerate (v.2.2.0). We included the previously described *A. filiculoides* RNA 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). Genomes with an annotation edit distance (AED) score of < 0.2 were used to train SNAP, which was used during 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 x 10^-15) HSPs from an all-by-all blast alignment, were removed from the training set. 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 Mapsr tool. We selected four species of heterosporous ferns (two *Azolla*, one *Salvinia* and one *Philodila*) and representatives of three other leptosporangiate lineages (Blechnum, Lygodium and Diplopterus). The Mapsr 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 filtering, we cut and reconstructed the entire dataset with a translated protein sequence consisting of the full Swiss-prot database, 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 Mapsr 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 Mapsr analysis using an E-value cut-off of 10^-10. We clustered gene families from the BLAST results with the default parameters and only kept gene families that contained at least one gene copy from each taxon in a given Mapsr analysis. We discarded the remaining OrthoFinder clusters. We used PASTA for automatic alignment and phylogenetic reconstruction of gene families, employing MAFFT for constructing alignments, MUSCLE for merging alignments and RAXML for tree estimation. The parameters for each software package 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 Mapsr.
size of 100 for parameter estimation, which was necessary to provide an upper bound for numerical integration of node states\(^4\). 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\(^5\). We developed ultrametric species trees from the topological relationships inferred by the 1KP Consortium analyses and median branch lengths from TimeTree\(^6\). 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 syntetic duplicates and microsynten analysis. To distinguish gene duplicates as syntenic or tandem, we used the SynMap\(^7\) tool from the CoGe\(^8\) 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 genomic region were identified as tandem duplicates (Supplementary Discussion). Results for within Azolla and Salvinia genome comparisons, as well as between Azolla and Salvinia, can be regenerated using the links https://genomevolution.org/r/tozk, https://genomevolution.org/r/toz7 and https://genomevolution.org/r/toy, respectively. Microsynten analysis were performed using the GEno tool from CoGe\(^8\). We used the default setting to define a minimum number of collinear 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://genomevolution.org/r/ujll and https://genomevolution.org/r/ukys.

Gaussian mixture model analysis of Kd distributions. Estimates of Kd were obtained for all pairs of syntenic paralogous and orthologous genes using the CODEML program\(^9\) in the PAML package (v4.8)\(^10\) 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\(^11\) (setting -codon) to align coding DNA, always skipping insertions (F). Only gene pairs with Kd values in the range of 0.05–5 were considered for further analyses. Gaussian mixture models were fitted to the resulting frequency distributions of Kd values by means of the densityMclust function in the R mclust version 5.3 package\(^12\).

The Bayesian information criterion was used to determine the best-fitting model for the data, including the optimal number of Gaussian components to a maximum of nine. For each component, several parameters were computed including the mean and the variance, as well as the density mixing probabilities and the total number of gene pairs.

Gene family classification and ancestral reconstruction. The OrthoFinder\(^13\) clustering method was used to classify complete proteomes of 23 sequenced green plant genomes, including A. filiculoides and S. cucullata (Supplementary Table 5), into orthologous gene lineages (that is, orthogroups). We selected taxa that represented all of the major land plant and green algal lineages, including six core eudicot species: A. thaliana, (Montbretia), Populus trichocarpa, Erythranthe guttata, and Vitis vinifera, four monocots (O. sativa, Sorghum bicolor, Musa acuminata, Zostera marina and Spirodella polycirrhiza), one basal angiosperm (A. trichopoda), two gymnosperms (Pinus taeda and Picea abies), two ferns (A. filiculoides and S. cucullata), one lycophyte (S. moellendorfii), four bryophytes (Pseudogynium fallax, P. patens, Marchantia polymorpha and Jungermannia fusca) and two green algae (Klebsormidium flaccidum and C. reinhardtii). In total, 16,817 orthogroups containing at least two genes were circumscribed, 8,680 of which contain at least one gene from either A. filiculoides or S. cucullata. Of the 20,203 annotated A. filiculoides genes and the 19,780 annotated S. cucullata genes, 17,941 (89%) and 16,807 (84%) were classified into orthogroups, respectively. The details for each orthogroup, including gene counts, secondary clustering of the orthogroups (that is, super-orthogroups)\(^14\) and functional annotations, are reported in Supplementary Table 5.

We used Wagner parsimony implemented in the program Count\(^15\) with a weighted gene gain penalty of 1.2 to reconstruct the ancestral gene content at key nodes in the phylogeny of the 23 land plants and green algae species (Supplementary Table 5). The ancestral gene content dynamics—gains, losses, expansions and contractions—are depicted in Supplementary Fig. 5. Complete details of orthogroup dynamics for the key ancestral nodes that include seed plants, such as Selasniaceae, euflyophytes and vascular plants, are reported in Supplementary Table 5.

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\(^16,17\). 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 (v.6.5.7) package\(^18\) with a minimum size restriction of 300 nucleotides. These open reading frames were searched for PPR motifs using the hmmsearch tool from the HHMER package\(^19\). The PPR motif models and parameters used follow those of Cheng et al.\(^20\). Motifs were assembled into full PPR tracts and the best model for each PPR was determined\(^21\).

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 ≥5%Ns. Clean reads were aligned against the assembled plastid and mitochondrial genome assemblies using TopHat 2.0 (ref.\(^22\)). One of the inverted repeat regions in the plastid genomes was removed before mapping. Only uniquely mapped reads were retained as input for SAMtools\(^23\) to call mismatches between the reads and the reference. 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\(^24\) to search for Tma12 (Genbank accession: 3Q438776) homologues in Phytomyzoz\(^25\), 1KP transcriptomes\(^26\) 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 cleaved leader domain that begins to the PF03067 Pfam family, but this family 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\(^27\). We then used this preliminary phylogeny (Supplementary Data 3) to construct a more focused data set to narrow down the phylogenetic affinity of Tma12. PartitionFinder\(^28\) was used to infer the optimal codon partition scheme and substitution models, and RAxML\(^29\) 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 polyphylgenomic 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\(^28\) to identify the optimal data partition scheme and the associated nucleotide substitution models. RAxML\(^29\) 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\(^30\) to extract the exon sequences from each of the resequenced species. The T. mauritianum instead of T. trichopoda was used as the outgroup. 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\(^31\), with the default setting, except that RAxML\(^29\) was used to estimate the best tree on the final alignment. To infer the species tree from these gene trees, we used the multi-species coalescent method implemented in ASTRAL-III (v5.6.1)\(^32\). 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 DNA, 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\(^33\) (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\(^34\); Genbank accession: NC_014248) using BLASTN (alignment criteria: E-value ≤ 1 × 10\(^{-15}\), sequenced reads and aligned coverage of ≥90%). Only the aligned reads were assembled by Mitobim\(^35\) (iterations = 5) using N. azollae 0708 (ref.\(^36\)) as a reference. Gene prediction for each cyanobiont assembly was performed by the Prodigal program\(^37\). Transfer RNAs were predicted by tRNAscan-SE\(^38\) using a bacterial tRNA gene structure model. The presence of tRNA 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 mugy10 to generate the whole-genome alignment, which resulted in a nucleotide matrix of 5,354,640 characters. IQ-TREE11 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 STRIDE12 implemented in OrthoFinder13 to locate the root by reconciling gene trees. STRIDE was run with the default setting, except that MAFFT14 was used for alignment and RAxML15 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/POLLYX and VAPYRIN sequences were used as queries, as in a previous study14, to search against the genomes and transcriptomes from species listed in Supplementary Table 11 using BLASTN15. 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. flaciloides and S. cucullata, both the annotated gene models and the unannotated genome contigs 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 used. All hits that matched already annotated gene models were discarded prior to subsequent analyses. No homologues were identified in the two fern genomes used.

Quantitative real-time PCR of nifH. Quantitative real-time PCR for the N. azollae nifH gene was conducted using total RNA extracted from A. flaciloides. Primers were designed from Brunner et al.16. Thermocycler Superscript II17 was used to generate complementary DNA from the RNA. The cDNA was then used for quantitative PCR with the Roche SYBR Green Master Mix on a Cfx96 real-time PCR machine with the Opticon platform. The relative gene expression was calculated using the 2ΔΔCT method, with the cyanobacteria present/nitrogen absent condition as the reference.

Azolla symbiosis transcriptome analysis. We used RNA-seq to compare gene expression patterns of AzCy+ and AzCy− individuals grown with or without ammonium nitrate. Each condition and treatment combination has three biological replicates. RNA-seq reads were mapped to the A. flaciloides genome by HISAT218 and read counts for each gene were calculated using the HTSeq software package19. We used the rlog function in the DESeq2 package20 for data normalization 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. flaciloides genome, we first investigated a potential orthologous relationship between any Azolla genes and cyanobacteria. For this, we used the Azolla genome assemblies as a query for Blastn against the cyanobacterial genome 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 BLASTn21 against the Azolla genome; 340 Azolla contigs had reciprocal hits. To investigate whether these represent possible Noatoc-to-Azolla transfers or just examples of plastid-to-nucleus transfers, we used the 340 Azolla contigs against another BLASTn 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 database 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 Phytozome95, 1KP transcriptomes32 and the NCBI Genbank database. Homologues of SHC and oxidosqualene cyclase were obtained by searching against Phytozome95, 1KP transcriptomes32 and the NCBI Genbank database. Using a Bayesian approach, we analyzed the SHC-nonsynthesized triterpenes. Freeze-dried S. cucullata biomass was Soshlet 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)trifluoroacetamide (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 Thermo Trace GC Ultra Trace DSQ gas chromatography mass spectrometry (GC–MS) onto a CP-sil 5CB-fused silica column (30 m × 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 PRJNA405527 and PRJNA405459. The sequence alignments and tree files can be found in the Supplementary Data.

References


Smit, A. & Hubley, R. RepeatModeler Open 1.0 (Institute for Systems Biology, 2015); http://www.repeatmasker.org


Andrews, S. FastQC: a quality control tool for high throughput sequence data v0.11.7 (Ibabrahim Institute, 2018); http://www.bioinformatics. babraham.ac.uk/projects/fastqc
91. Smit, A. F., Hubley, R. & Green, P. RepeatMasker Open 4.0 (Institute for Genomic Research, 2015).}

**Articles**


**Acknowledgements**

We are grateful to 123 backers from Experiment.com who crowdfunded the initial work. We thank Z. Fei for providing comments and suggestions, M. Harrison for discussion on Azolla transporter genes, J. Shaw and D. Weston for providing access to the *Sphagnum* genome and T. Nishiyama for the *Jungermannia* genome, and P.-F. Lu for providing the image for Fig. 1c. This project was partly supported by the Shenzhen Municipal Government and the Institute of China (no. JCYJ2017030109405946), the National Science Foundation Doctoral Dissertation Improvement Grant DEB-1407158 (to K.M.P. and F.-W.L.) and the German Research Foundation Research Fellowship VR132/1-1 (to J.d.V.). Computational support was provided by the Duke Compute Cluster and the Centre for Information and Media Technology at University of Düsseldorf.

**Author contributions**


**Competing interests**

The authors declare no competing interests.

**Additional information**

Supplementary information is available for this paper at https://doi.org/10.1038/s41477-018-0184-8.

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

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.
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.

Sample size
For phylogenetic analyses, bootstrapping datasets were sampled between 100 to 1000 times, which is the field standard.

Data exclusions
Gene models without transcript or homology supports were excluded (see the supplementary discussion).

Replication
The RNA-seq experiments were done with three biological replicates per treatment.

Randomization
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.

Blinding
Blinding is not applicable in this study.

Reporting for specific materials, systems and methods

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

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

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 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.