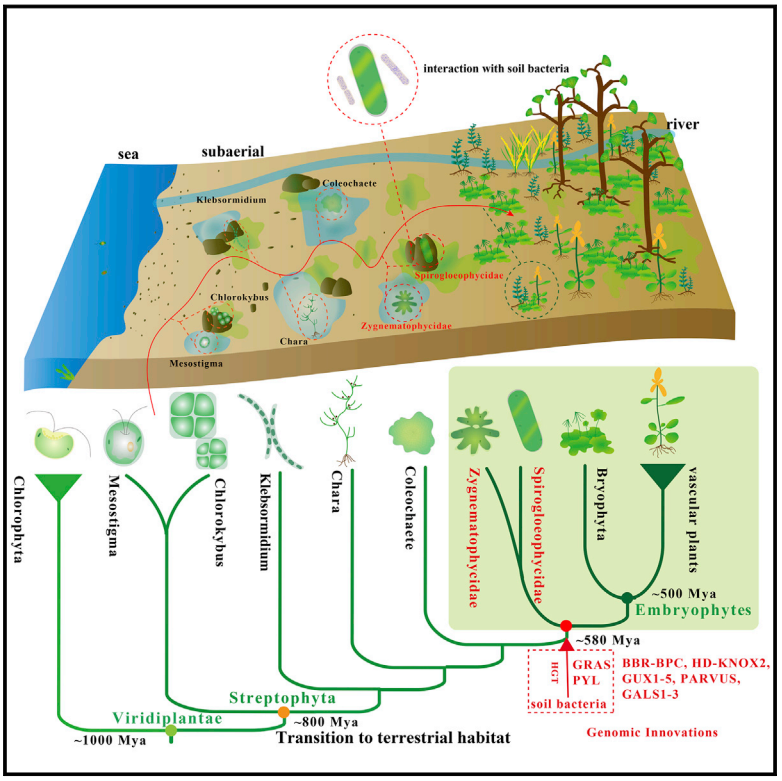
Article

Cell

Genomes of Subaerial Zygnematophyceae Provide Insights into Land Plant Evolution

Graphical Abstract



Highlights

- Genomes of two subaerial Zygnematophyceae highlight terrestrial adaptation
- A novel lineage is described that is closest to the origin of embryophytes
- Genes acquired by HGT from soil bacteria regulate plant development and stress
- A recent whole genome triplication is reported for *Spirogloea muscicola*

Authors

Shifeng Cheng, Wenfei Xian, Yuan Fu, ..., Barbara Melkonian, Gane Ka-Shu Wong, Michael Melkonian

Correspondence

gane@ualberta.ca (G.K.-S.W.), michael.melkonian@uni-koeln.de (M.M.)

In Brief

The genomes of two streptophyte algal species, including a newly identified lineage placed closest to the branch point separating green algae and land plants, are reported and provide evidence that genes thought to be important for resistance to desiccation were gained by horizontal gene transfer from soil bacteria approximately 580 million years ago.

Cheng et al., 2019, Cell *179*, 1057–1067
November 14, 2019
https://doi.org/10.1016/j.cell.2019.10.019

CellPress

Article

Cell

Genomes of Subaerial Zygnematophyceae Provide Insights into Land Plant Evolution

Shifeng Cheng,[1,11] Wenfei Xian,[1,11] Yuan Fu,[1] Birger Marin,[2] Jean Keller,[3] Tian Wu,[4,5] Wenjing Sun,[4,5] Xiuli Li,[1] Yan Xu,[4,5] Yu Zhang,[1] Sebastian Wittek,[2] Tanja Reder,[2] Gerd Günther,[6] Andrey Gontcharov,[7] Sibo Wang,[4,5] Linzhou Li,[4,5] Xin Liu,[4,5] Jian Wang,[4,5] Huanming Yang,[4,5] Xun Xu,[4,5] Pierre-Marc Delaux,[3] Barbara Melkonian,[2,10] Gane Ka-Shu Wong,[4,8,9,*] and Michael Melkonian[2,10,12,*]

[1]Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Area, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, 518120, China

[2]Botanical Institute, Cologne Biocenter, University of Cologne, 50674 Cologne, Germany

[3]Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, CNRS, UPS, Castanet Tolosan, France

[4]BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China

[5]China National GeneBank, BGI-Shenzhen, Jinsha Road, Shenzhen 518120, China

[6]Knittkuhler Str. 61, 40629 Düsseldorf, Germany

[7]Federal Scientific Center of the East Asia Terrestrial Biodiversity, Far Eastern Branch, Russian Academy of Sciences, Vladivostok RUS-690022, Russia

[8]Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada

[9]Department of Medicine, University of Alberta, Edmonton, AB T6G 2E1, Canada

[10]Present address: University of Duisburg-Essen, Campus Essen, Faculty of Biology, Universitätsstr. 5, 45141 Essen, Germany

[11]These authors contributed equally

[12]Lead Contact

*Correspondence: gane@ualberta.ca (G.K.-S.W.), michael.melkonian@uni-koeln.de (M.M.)

https://doi.org/10.1016/j.cell.2019.10.019

SUMMARY

The transition to a terrestrial environment, termed terrestrialization, is generally regarded as a pivotal event in the evolution and diversification of the land plant flora that changed the surface of our planet. Through phylogenomic studies, a group of streptophyte algae, the Zygnematophyceae, have recently been recognized as the likely sister group to land plants (embryophytes). Here, we report genome sequences and analyses of two early diverging Zygnematophyceae (*Spirogloea muscicola* gen. nov. and *Mesotaenium endlicherianum*) that share the same subaerial/terrestrial habitat with the earliest-diverging embryophytes, the bryophytes. We provide evidence that genes (i.e., GRAS and PYR/PYL/RCAR) that increase resistance to biotic and abiotic stresses in land plants, in particular desiccation, originated or expanded in the common ancestor of Zygnematophyceae and embryophytes, and were gained by horizontal gene transfer (HGT) from soil bacteria. These two Zygnematophyceae genomes represent a cornerstone for future studies to understand the underlying molecular mechanism and process of plant terrestrialization.

INTRODUCTION

Recent phylogenomic analyses concluded that a species-rich lineage of streptophyte algae, the Zygnematophyceae, represents the most likely sister group of embryophyte land plants (Gitzendanner et al., 2018; Timme et al., 2012; Wickett et al., 2014; Wodniok et al., 2011). This was unexpected, because Zygnematophyceae are structurally simple, consisting of unicells or simple filaments, whereas the structurally more complex Charophyceae and Coleochaetophyceae, which had previously been favored as closest relatives of embryophytes (Graham, 1993; Karol et al., 2001), were more distantly related to them. Another prevailing notion, namely that transition from water to land (terrestrialization) occurred in the common ancestor of embryophytes (Doyle, 2013), has also recently been challenged (Harholt et al., 2016). Several studies based on transcriptome assemblies documented that the molecular tool kit for life in a terrestrial environment evolved in streptophyte algae before the origin of embryophytes (de Vries et al., 2018; Delaux et al., 2015; Hori et al., 2014; Mikkelsen et al., 2014). These findings are apparently in conflict with the aquatic lifestyle of Charophyceae, Coleochaetophyceae, and most Zygnematophyceae. In Zygnematophyceae, however, early diverging taxa are subaerial/terrestrial, suggesting that the ancestor of this class might have originated on land. Here, we present genome sequences from Zygnematophyceae, choosing two subaerial species (*Spirogloea muscicola* gen. nov. and *Mesotaenium endlicherianum*) (Figure 1A) that represent early divergences in the class.

RESULTS

Genomes of Two Early Diverging Zygnematophyceae: Phylogeny and Taxonomic Implications

Phylogenomic analyses confirmed that Zygnematophyceae are sister to embryophytes (Figure 1B), and a phylogeny based on an extended taxon sampling with a restricted dataset revealed the exceptional position of *S. muscicola* well separated from all

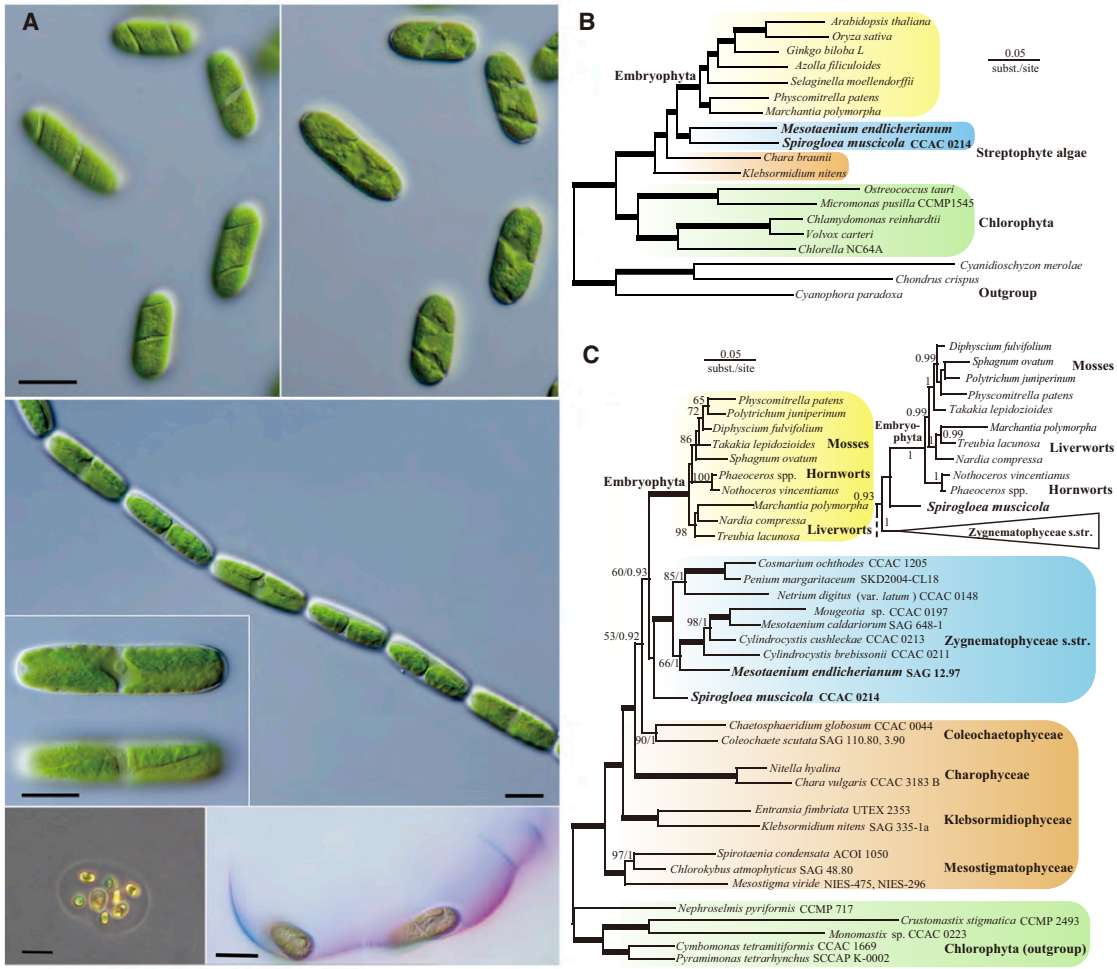


Figure 1. Light Micrographs and Phylogeny Reconstruction of Two Species of Zygnematophyceae

(A) Nomarski differential interference contrast images of *Spirogloea muscicola* (two focal planes: top left, cell surface with spiral chloroplast and top right, cell center) and *Mesotaenium endlicherianum* (middle). Cells of *M. endlicherianum* arranged in loose chains (inset, two focal planes of a single cell in central view [top] and surface view [bottom]), note that there are two chloroplasts per cell. Magnification bars, 10 μm. Phase contrast image (bottom left) and bright field image stained with crystal violet (bottom right) to show the confluent mucilage sheath enclosing groups of cells of *S. muscicola*. Magnification bars, 20 μm.

(B) Species phylogeny reconstruction for 19 species of Plantae by genome-wide concatenated orthologous genes. In total, 85 low-copy orthologous gene families (1–2 copies for each genome) were selected, and multiple sequence alignment was performed by MAFFT. The tree was built by RAxML8.2.4 (bold branches received maximal bootstrap support, 1,000 replicates) with LG4X amino acid substitution model. The two Zygnematophyceae were recovered as sisters to embryophytes.

(C) The phylogenetic position of *S. muscicola* and *M. endlicherianum* within the radiation of streptophyte algae and embryophytes (represented by Bryophyta) based upon complete nuclear and plastid-encoded rRNA operon sequences with an extended taxon sampling. Shown is the RAxML phylogeny (large tree); numbers at branches are RAxML bootstrap percentages/Bayesian posterior probabilities. Bold branches received maximal support (100/1). The substitution models used were GTRCAT (maximum likelihood) and GTR+I+G (Bayesian). The analysis recovered the paraphyletic divergence of five clades of streptophyte algae (Mesostigmatophyceae, Klebsormidiophyceae, Charophyceae, Coleochaetophyceae, and Zygnematophyceae). *S. muscicola* was well separated from all other Zygnematophyceae (Zygnematophyceae s. str.) in an unresolved position close to the branch point between Zygnematophyceae and embryophytes. The Bayesian tree (small tree) even positioned S. *muscicola* on the embryophyte branch, albeit without support.

See also Video S1, a video of live *S. muscicola* cells showing rotation of the spiral chloroplast (6× time lapse).

other Zygnematophyceae studied and closest to the branch point separating Zygnematophyceae from embryophytes (Figure 1C). In consequence, *Spirogloea gen. nov.* is placed in a new subclass, Spirogloeophycidae, which is also supported by plastid genome features including a canonical inverted repeat with operon structure of the ribosomal RNA genes (Data S1A–S1F).

Taxonomic Acts and Revisions

Class Zygnematophyceae Round ex Guiry emend. Melkonian, Gontcharov, and Marin.

Emended Diagnosis: Coccoid or filamentous streptophyte green algae; flagellate stages and centrioles entirely absent; sexual reproduction by conjugation, meiosis upon germination of zygote; freshwater or subaerial; loss of protein synthesis

elongation factor Tu gene (tufA) from chloroplast genome; group II introns in trnI (GAU) and trnV (UAC) genes lost from chloroplast genome.

Type Genus: Zygnema C Agardh (1817, Synopsis Algarum Scandinaviae, pp. XXXII, 98).

Comments: We prefer the use of Zygnematophyceae over Conjugatophyceae (recommended by Guiry [2013]) for three reasons. First, the "Conjugatae" were first established formally for conjugating green algae (both unicellular [Desmidieae] and filamentous [Zygnemeae and Mesocarpeae]) by de Bary (1858) as a descriptive name in accordance with Article 16.1(b) (International Code of Nomenclature for algae, fungi, and plants [Shenzhen Code]) (Turland et al., 2018). In his diagnosis, de Bary wrote "Fructification: Durch Copulation entsteht eine von ihren Mutterzellen verschieden gebaute Zygospore" (translated: fructification: a zygospore is formed by copulation differing in structure from its parental cells). The name "Conjugatae" de Bary thus derives from the Latin "conjugatae," meaning "those that are connected by a yoke." Because the taxon Conjugatae when erected by de Bary was descriptive and not typified (i.e., not derived from the genus *Conjugata* Vaucher 1803, a rejected name in favor of *Spirogyra* Link in Nees [1820], *nom. cons.*), recommendation 16A of the Code does not apply ("In choosing among typified names for a taxon above the rank of family, authors should generally follow the principle of priority."). Second, the term "conjugation" is ambiguous and also applies to organisms that are not Conjugatophyceae, e.g., bacteria, ciliates, diatoms, and most recently, the streptophyte algal genus *Spirotaenia*, which has been shown to be a member of the Mesostigmatophyceae (Gontcharov and Melkonian 2004; Wickett et al., 2014) (Figure 1C). Conjugation in streptophyte green algae is thus homoplasious. Third, the name Zygnematophyceae is much more widely used in the literature. A quick search in Clarivate's Web of Science (https://webofknowledge.com; retrieved October 11, 2019) returned 201 articles including the name Zygnematophyceae compared to only 64 articles for Conjugatophyceae (some articles used both names) since 1991 (the first time that Zygnematophyceae was used). Furthermore, articles with Zygnematophyceae are cited more than four times as often than those with Conjugatophyceae, reflecting the increasing use of the name Zygnematophyceae in experimental and molecular work.

Subclass Zygnematophycidae Melkonian, Gontcharov, and Marin subclass. nov.

Diagnosis: With the characteristics of the class Zygnematophyceae; loss of operon structure in rRNA genes on chloroplast genome; loss of linkage between trnI (GAU) and trnA (UGC) genes on chloroplast genome; loss of ribosomal protein gene rpl32 from the chloroplast genome.

Subclass Spirogloeophycidae Melkonian, Gontcharov, and Marin subclass. nov.

Diagnosis: With the characteristics of the class Zygnematophyceae; canonical operon structure of rRNA genes in inverted repeats present on chloroplast genome; trnI (GAU) and trnA (UGC) located between rrL and rrS genes on chloroplast genome; ribosomal protein gene rpl32 located on chloroplast genome.

Comments: The separation of class Zygnematophyceae into two subclasses Zygnematophycidae and Spirogloeophycidae is warranted, because (1) of the sister group relationship between *Spirogloea muscicola* and all other Zygnematophyceae; (2) the large phylogenetic distance between *S. muscicola* and the next-diverging taxon in the Zygnematophyceae (*M. endlicherianum*); and (3) the unique characteristics of the chloroplast genome of *S. muscicola* among Zygnematophyceae.

Order Spirogloeales Melkonian, Gontcharov, and Marin ord. nov.

Diagnosis: With the characteristics of the subclass Spirogloeophycidae; unicellular with a single, spiral chloroplast per cell.

Family Spirogloeaceae Melkonian, Gontcharov, and Marin fam. nov.

Diagnosis: With the characteristics of the order Spirogloeales; chloroplast/protoplast rotates spontaneously with oscillations.

Genus *Spirogloea* Melkonian gen. nov.

Diagnosis: With the characteristics of the family Spirogloeaceae; several to many cells embedded in a common, confluent mucilaginous envelope.

Type Species: *Spirogloea muscicola* (de Bary, 1858) Melkonian nov. comb.

Etymology: Spiro- (from Greek "speira" [Σπείρα]), a coil (referring to the spiral shape of the chloroplast); -gloea from Greek "gloia" [γλοια], mucus (referring to the confluent mucilaginous envelope surrounding groups of cells).

Basionym: *Spirotaenia muscicola* (de Bary, 1858; Untersuchungen über die Familie der Conjugaten, p. 75)

Lectotype (hic designatus): de Bary 1858; Plate VIIF, Figure 1

Representative Strain: CCAC 0214 deposited at the Central Collection of Algal Cultures (CCAC) (http://www.ccac.uni-koeln.de/).

Representative DNA Sequence: nuclear-encoded small subunit (SSU) rDNA (accession MN585752).

Comments: A new combination for *Spirotaenia muscicola* de Bary, 1858 became necessary because genus *Spirotaenia* is polyphyletic, the species type (*S. condensata* Brébisson ex Ralfs 1848, The British Desmidieae, p. 179; lectotypified by Silva [1952], p. 252) is sister to genus *Chlorokybus* and thus a member of the Mesostigmatophyceae (Figure 1C). Several other *Spirotaenia* spp. (including *S. minuta* Thuret in Brébisson 1856) also belong to Mesostigmatophyceae (Gontcharov and Melkonian 2004; Wickett et al., 2014). In keeping with Article 13.1.(e) of the Code (the starting date for valid publications of names for organisms in Zygnematophyceae is January 01, 1848 [Ralfs, British Desmidieae]), a new genus name for *S. muscicola* de Bary was searched. In principle, the name *Endospira* Brébisson in Desmazières Crypt. de France is available (Silva, 1952, p. 252). de Brébisson deposited two exsiccates named *Endospira bryophila* (from the 1st edition, fascicle XL, exsiccate no. 1954 and from the 2nd edition, fascicle XXXIV, exsiccate no. 1654). According to Stafleu et al. (A Selective Guide to Botanical Publications and Collections with Dates, Commentaries and Types; https://www.sil.si.edu/DigitalCollections/tl-2/browse.cfm?vol=1#page/679), fascicle XXXIV was published in 1848 and fascicle XL in 1850. Therefore, *Endospira bryophila* de Brébisson, in Desmazières Crypt. de France 2nd edition, exsiccate no. 1654 (1848) should be the type species of genus *Endospira*. There is some confusion in the literature (summarized by Lütkemüller [1903]) about the taxonomic

identity of the two exsiccates of de Brébisson. According to the testimony of Rabenhorst (cited in Lütkemüller, 1903), exsiccate no. 1654 of de Brébisson corresponds to *S. muscicola* de Bary, whereas exsiccate no. 1954 investigated by Lütkemüller (1903) is a different species, *E. bryophila* Brébisson. One of the authors of this contribution (M.M.) visited the Muséum national d'Histoire naturelle (MNHN), 57 rue Cuvier, Cedex 05 Paris, France, in October 2017 to investigate both exsiccates. They are both labeled "ENDOSPIRA BRYOPHILA, De Bréb, in herb." and have identical text descriptions and overall appearance. It is concluded that they were prepared from the same natural material at the same time. M.M. also had the chance to see two original exquisite watercolor drawings of de Brébisson in the Muséum national d'Histoire naturelle (MNHN), labeled *E. bryophila* and *E. truncorum* (the latter identical to *S. muscicola* De Bary) that clearly represent two different species. de Brébisson apparently knew both species well, and he would not have labeled both exsiccates "E. bryophila," if they represented two different species. We therefore initially considered the organism, described here as *Spirogloea muscicola* (de Bary, 1858) Melkonian comb. nov., to be a second species of *Endospira*, *E. muscicola*. *E. bryophila* was recently isolated from a natural sample near the original sampling site of *S. muscicola* CCAC 0214 by one of us (M.M.). It conforms in morphology with the description of de Brébisson in Desmazières and de Brébisson's watercolor drawing as well as Lütkemüller's (1903) diagnosis of *Spirotaenia bryophila*. When its nuclear-encoded SSU rDNA sequence was determined, the organism, to our surprise, turned out to be a Trebouxiophyceae (Chlorophyta), related to genera *Koliella*/*Raphidonema* (unpublished data), and so belongs to yet another class of green algae (in the Chlorophyta). This made it impossible to use the generic name *Endospira* for strain CCAC 0214 and necessitated the erection of a new genus for *Spirotaenia muscicola* de Bary, namely *Spirogloea muscicola* nov. comb.

Sequencing and Assembly of the Genomes

Traditional hierarchical shotgun whole-genome deep sequencing was performed on the Illumina platform, resulting in 354X and 412X genome coverage for S. *muscicola* and *M. endlicherianum*, for which haploid nuclear genome sizes of 174 Mb and 163 Mb were estimated, respectively (Tables S1A and S1B; Data S1G). *S. muscicola* genome sequences were assembled into 19,678 scaffolds, covering 98.3% of the predicted genome size with a scaffold N50 reaching 566 kb (Table S1C), the contiguity accuracy was also validated by paired-end and mate-paired libraries from the randomly selected scaffolds (Data S2A). Transcriptome sequencing reads, assembled transcripts, and the eukaryotic BUSCO dataset were aligned against the genome assembly, with 95%, 94%, and 87.1% of these datasets successfully mapped, respectively, indicating that an appropriate genome completeness for further analyses was captured (Table S1D; Data S1H and S1J). A comparable quality of genome assembly was obtained for *M. endlicherianum* (Table S1D; Data S1I, S1J, and S2B).

Structural and Comparative Genomics

Genome annotation and evaluation (Tables S1E–S1G; Data S1K) revealed a high-confidence gene set, presenting 27,137 (*S. muscicola*) and 11,080 (*M. endlicherianum*) genes for downstream analyses (Tables S1H–S1J). The much larger gene set in *S. muscicola* is attributed to a recent whole genome triplication event, which is supported by a large-scale burst of triplicated segments identified (Figures 2 and S1; Data S1L and S1M), as well as tripled orthologs and collinear blocks in *S. muscicola* corresponding to only one in *M. endlicherianum* (Data S1N–S1Q). Comparative phylogenomic analyses were performed among 16 representative genomes of Chlorophyta, Klebsormidiophyceae, Charophyceae, Zygnematophyceae, bryophytes, and vascular plants (Table S1K). We built a core Viridiplantae gene set through genome-wide ortholog gene clustering and recovered 5,076 orthogroups shared by all Viridiplantae and 376 orthogroups shared between Zygnematophyceae and embryophytes (Figure 3A).

Gene Family Innovations and Dynamics

The homolog matrix of orthogroups (Table S1L) was further analyzed to infer the ancestral and lineage-specific gene content along the phylogenetic tree, resulting in 373 gained orthogroups (Table S1M) and 232 expanded orthogroups (Table S1N) in the common ancestor of Zygnematophyceae and embryophytes (Figure 3B). To bioinformatically filter out false positives and retrieve false negatives, further analyses were done using the Viridiplantae dataset by re-Blast, HMMER search, as well as phylogenetic analysis (see STAR Methods for details) for each gene member of the 373 gained orthogroups, resulting in a total of 902 genes in 22 orthogroups. These genes are innovations that likely evolved in the common ancestor of Zygnematophyceae and embryophytes (Figure 3C, subset; Table S1M). Many of these genes have been studied by forward and reverse genetics in flowering plants and are known to play roles in response to biotic and abiotic stresses in embryophytes. They include transcription factor (TF) genes (Figure 4A; Table S1O), as well as genes involved in phytohormone signaling (Figure 4B; Table S1P; Data S1R–S1AE) and biosynthesis of cell wall constituents and their remodeling (Table S1Q; Data S1AF–S1AM). Three TF families (GRAS, HD-KNOX2, and BBR/BPC) (Figure 4C; Data S1AN and S1AO), a homolog of the PYR/PYL/RCAR-like abscisic acid (ABA) receptor (Figures 4B and S2), and genes involved in 1,4 β-xylan formation (GUX1-5, PARVUS) and galactan/RG I pectin synthesis (GALS1-3), were likely gained in the common ancestor of Zygnematophyceae and embryophytes (Data S1AF, S1AG, and S1AL). Expanded orthogroups in the common ancestor of Zygnematophyceae and embryophytes (see STAR Methods), were also recorded; most refer to transcription factors, regulators of gene expression, receptors, and signaling components involved in abiotic and biotic stress responses (Figure 3D, subset; Table S1N). Some of these gains and expansions of genes had already been deduced from transcriptomic analyses (Bowman et al., 2017; de Vries et al., 2018; Jensen et al., 2018; Wilhelmsson et al., 2017) and are confirmed here genomically or are corroborated by previous biochemical data (pectins) (Domozych et al., 2014). The origin of the GRAS gene family in streptophytes, however, remains currently unresolved: although the presence of a GRAS homolog in transcriptomes of *Chaetosphaeridium globosum* (Cooper

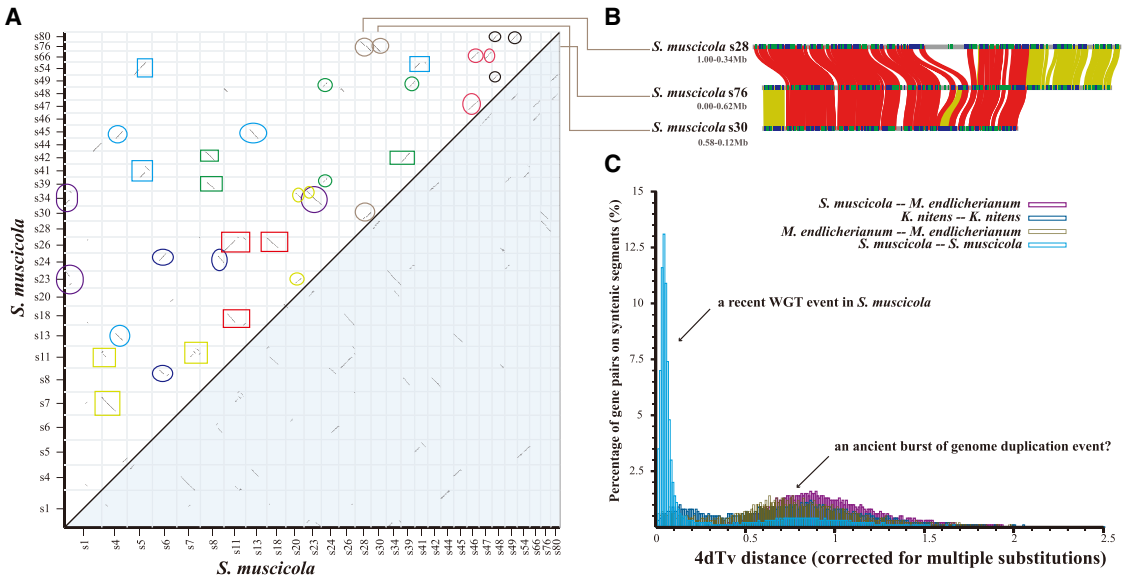


Figure 2. A Recent Whole Genome Triplication Event Observed in *S. muscicola* through Comparative Genome Analysis

(A) Whole-genome self-alignment and syntenic blocks within *S. muscicola* were present, strings of paralogous genes that correspond to triplicated regions are highlighted using the same circles or boxes in the same color.

(B) An exemplary set of triplicated blocks derived from scaffold28, scaffold76, and scaffold30. Red, microsynteny shared by all three scaffolds. Yellow, syntenic blocks and paralogs were only retained between two scaffolds.

(C) Age distribution of all duplicated paralogous genes in *S. muscicola* (cyan) and *M. endlicherianum* (brown), *Klebsormidium* (blue), as well as orthologous gene pairs between *S. muscicola* and *Mesotaenium* (purple), respectively.

See also Figure S1.

and Delwiche, 2016) was confirmed through phylogenetic analyses, suggesting that GRAS homologs may have originated earlier than suggested here, the position of the *C. globosum* sequence in the tree is compatible with either a genuine Chaetosphaeridium GRAS gene or a bacterial gene (i.e., a contamination) (Data S1AY–S1BB). Furthermore, motif 1, which is present in all GRAS domains of embryophytes, Zygnematophyceae, and bacteria GRAS group 1, is missing from this sequence (as in bacteria GRAS group 2) (Figure S3). We tried unsuccessfully to amplify the putative GRAS sequence from genomic DNA of axenic *C. globosum* by PCR using a variety of primer combinations derived from the accession that successfully amplified GRAS genes from *Spirogloea muscicola* with the same parameters (Data S1BD). GRAS genes appear to be absent from *Coleochaete* spp., both at the transcriptome and genome levels (Table S1O). A final conclusion must be deferred until draft genomes of *C. globosum* and other Coleochaetophyceae become available. In any case, a significant expansion of GRAS genes apparently occurred in the common ancestor of Zygnematophyceae and embryophytes (Figure 4C).

However, there were also losses of genes in the ancestor of the Zygnematophyceae including all genes involved in structure and function of flagella/basal bodies, in accordance with their loss in this class (Data S1AP). Although sexual reproduction has not been reported for the two species, we found all core meiosis-specific genes (10 and 11, respectively) in their genomes (Table S1R).

Symbiosis-Related Genes

Surprisingly, Zygnematophyceae lack some genes involved in innate immunity such as LysM receptor-like kinases (LysM-RLKs), which function in pattern-triggered immunity and symbioses. The disease resistance (R) protein NB-ARC, that responds to effectors secreted by pathogens to help establish successful infections, is also absent in Zygnematophyceae (Tables S1S and S1T). Because both genes are present in other streptophyte algae (Han, 2019; Nishiyama et al., 2018), we can only speculate that evolution of the mucilaginous coat that surrounds most Zygnematophyceae, and which may increase desiccation tolerance, released the selective pressure to retain these genes. Whereas genes involved in early steps of AM-symbiosis previously detected in transcriptomes of Zygnematophyceae such as CCaMK and CYCLOPS (Delaux et al., 2015) were found in the genomes, no orthologs of later steps in the AM-symbiosis such as VAPYRIN were recovered confirming their evolution in embryophytes (Table S1U; Data S1AQ–S1AX).

GRAS Genes and HGT

Genes of the GRAS gene family are essential regulators of plant growth and development but also have functions in response to biotic and abiotic stresses and in symbiosis (Gonzalez, 2015). Until recently, the GRAS gene family was thought to be restricted to embryophytes, but transcriptomic evidence for its presence in Zygnematophyceae has been presented (Cooper and Delwiche, 2016; Delaux et al., 2015; Wilhelmsson et al., 2017). Evolutionary analyses of the GRAS gene family in angiosperms identified

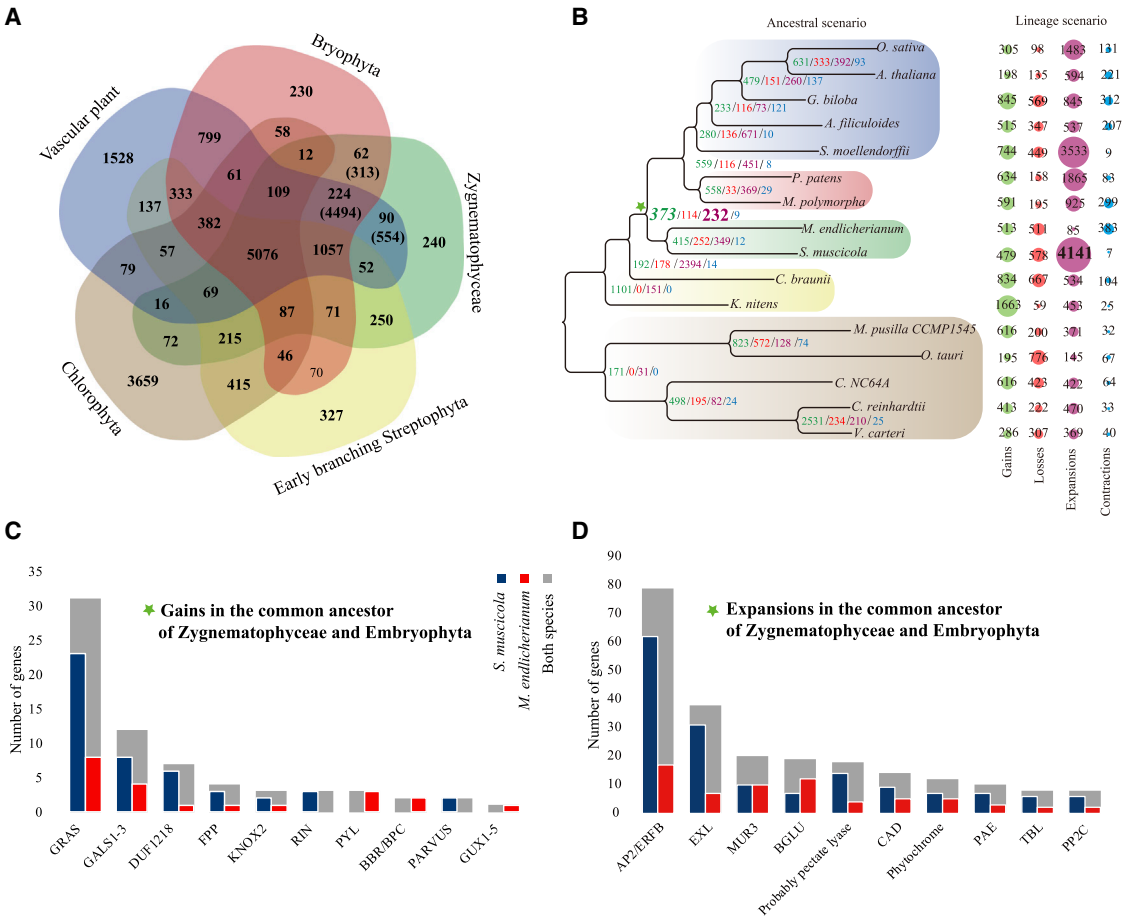


Figure 3. Gene Family Evolution

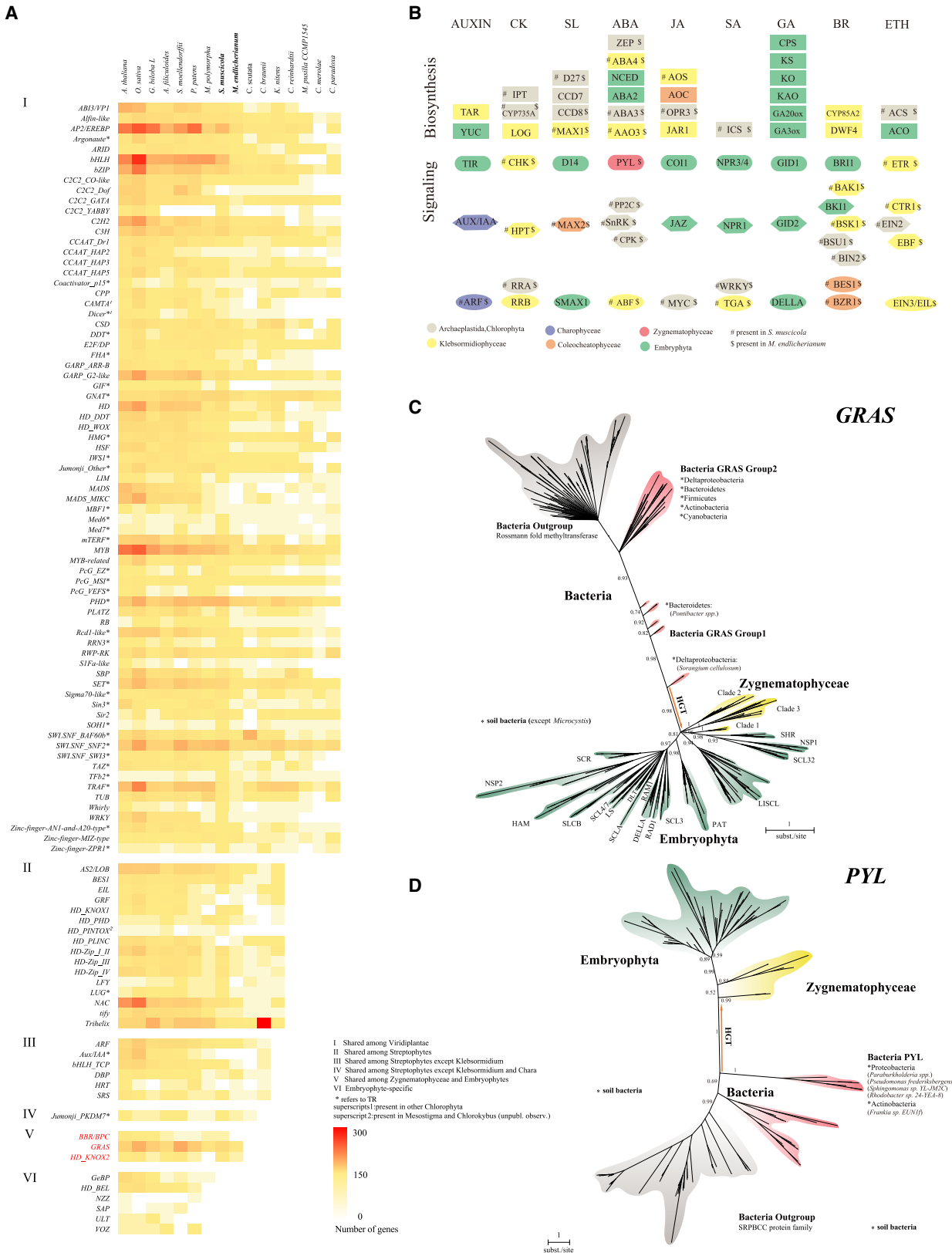
(A) Venn diagram to show shared and unique orthogroups between five groups of Viridiplantae. Gene lists of orthogroups clustered by Orthofinder 2.2.6 (default parameters) from 16 genomes used in this study are summarized in Tables S1K and S1L. Numbers of orthogroups and genes (the latter in parentheses), that are exclusively shared between Bryophyta and Zygnematophyceae, between vascular plants and Zygnematophyceae, and between Bryophyta, vascular plants, and Zygnematophyceae are highlighted in bold.

(B) Evolutionary analyses of gains (green), losses (red), expansions (purple), and contractions (blue) of orthogroups, in the context of phylogenetic profiles, both for the reconstruction of the ancestral gene content in key nodes and the dynamic changes of the lineage-specific gene characteristics. The Count software (Method Details) was implemented to define and calculate gains, losses, expansions, and contractions of orthogroups for each branch. The ancestral state was reconstructed and compared with the node of the closest outgroup using a phylogenetic birth-and-death statistic model. The size of the circles is proportional to the number of orthogroups (the large number of orthogroup expansions in *S. muscicola* refers to its triploid genome).

(C and D) A representative list of genes in orthogroups that were gained (C) or expanded (D) in the common ancestor of Zygnematophyceae and embryophytes. Both (C) and (D) were based on the results derived from Figure 3B. The 373 orthogroups gained and 232 orthogroups expanded as defined in (B) were further analyzed by re-blast, HMMER searching, and phylogenetic analyses (Tables S1M and S1N); a subset of these genes is shown in (C) and (D) (for details see Method Details). See also Figure S2.

29 orthogroups reflecting large-scale gene expansion and functional diversification (Cenci and Rouard, 2017). Here, we identified 8 and 23 GRAS genes in *M. endlicherianum* and *S. muscicola*, respectively. Through an extensive search in genome databases of bacteria, fungi, animals, protists, algae, and embryophytes using the conserved GRAS domain with hidden Markov models, we found GRAS-like sequences only in embryophytes, Zygnematophyceae, and bacteria (Figure 4C; Table S1V; Data S1AY–S1BC), suggesting that horizontal gene transfer (HGT) might link these unrelated organisms.

The extent and frequency of HGT from bacteria to eukaryotes is a controversial topic (Husnik and McCutcheon, 2018; Ku and Martin, 2016; Martin, 2017; Soucy et al., 2015). In genome analyses, contamination with bacterial sequences has been a frequent problem (Bemm et al., 2016). We used axenic cultures, monitored axenicity until DNA extraction, and followed recommendations for reducing reagent and laboratory contaminations in genome analyses (Salter et al., 2014). Furthermore, we showed co-assembly of the HGT candidate genes with eukaryotic genes on genomic scaffolds, acquisition of introns, and functionality through gene expression using RNA sequencing (Data S2C). Phylogenetic analyses of the GRAS domain (Figure 4C) suggested a single HGT event from a bacterial donor, because the streptophyte sequences had a single origin, nested



(legend on next page)

within a larger bacterial radiation. This corroborates an earlier report that the GRAS domain belongs to the Rossmann fold methyltransferase superfamily that first emerged in bacteria (Zhang et al., 2012). The search for the bacterial donor of the HGT is compounded by the fact that the GRAS domain apparently underwent multiple HGTs among bacteria. Five phyla of bacteria contain GRAS-like sequences (58 sequences, Table S1W). All of these sequences (except two) derive from soil bacteria. HGT among soil bacteria is rampant involving IncP- and IncPromA-type broad host range plasmids (Klümper et al., 2015). One group of four bacterial sequences (*Sorangium cellulosum*) was monophyletic with the streptophyte sequences (Figure 4C). Analysis of the GRAS domain structure identified all 10 motifs present in streptophyte GRAS domains in this group of bacteria, arranged in the same order (Figure S3). A second, more distantly related group of bacteria (group 2) contained 4 motifs in the same order. The stepwise gain of GRAS domain motifs in bacteria correlated well with the GRAS phylogeny. Duplications (possibly three) of the GRAS gene presumably occurred in the common ancestor of Zygnematophyceae and embryophytes with further duplications, in part by WGD/Ts, in each clade after their split (Figure 4C). Three clades of GRAS genes could be distinguished in the Zygnematophyceae, one (clade 1, Figure 4C) monophyletic (posterior probability 0.96) with the GRAS gene subfamilies NSP1, SCL32, and SHR, the other two (clades 2 and 3) in unresolved positions at the base of the streptophyte GRAS radiation.

PYL Genes and ABA Responses

A genome-wide search for additional HGT candidates among the 902 genes from 22 orthogroups (Table S1M) that are common to Zygnematophyceae and embryophytes yielded only three genes (Table S1V) (in *M. endlicherianum*) that are homologs of the PYR/PYL/RCAR proteins, abscisic acid receptors that play important roles in plant responses to biotic and abiotic stresses (Jahan et al., 2019). The phylogenetic tree showed monophyly of the plant PYR/PYL/RCAR-like genes, nested within a larger radiation of bacteria suggesting again HGT (Figure 4D; Table S1X; Data S1BE–S1BJ). The genes were monophyletic (with moderate support) with a clade of six bacterial genes from two phyla again indicating HGTs among these bacteria. All bacteria with PYR/PYL/RCAR homologs (in two species on plasmids) are soil bacteria.

DISCUSSION

We cannot exclude the possibility that the HGTs described occurred earlier in evolution at the time of endosymbiotic gene transfers upon evolution of plastids/mitochondria from ancestral bacterial pangenomes (Ku et al., 2015). This scenario, however, requires numerous gene losses in both bacteria and eukaryotes and is contradicted by estimations of divergence times of HGT candidates between streptophyte genes and their closest bacterial relatives that are similar to the divergence times estimated for the common ancestor of Zygnematophyceae and embryophytes (Data S1BK and S1BL).

The spatial proximity between soil bacteria and the subaerial/terrestrial common ancestor of Zygnematophyceae and embryophytes should have facilitated HGTs, further supporting our HGT scenario. We propose that HGTs from soil bacteria have crossed the domain boundary into eukaryotes already before the divergence of Zygnematophyceae and embryophytes (~580 mya). The transferred genes underwent diversified selection (Data S1BM) and extensive expansions by gene and genome duplications accompanied by neo-functionalization through domain recruitment, shuffling, and loss. More generally, our results corroborate earlier notions that plant terrestrialization was likely accompanied by a widespread impact of horizontal gene transfer from soil bacteria to early land plants involving among others, YUC-genes, PAL, and microbial terpene-synthase-like (MTPSL) genes (Bowman et al., 2017; Emiliani et al., 2009; Jia et al., 2016; Yue et al., 2012). To what extent the GRAS and PYL/PYR/RCAR genes obtained from soil bacteria have played a crucial role in plant terrestrialization remains an open question and requires functional analysis of the respective genes in Zygnematophyceae, which is currently not possible, because of the lack of a well-characterized, genetically tractable model system in this group of algae. We note, however, that GRAS and other

Figure 4. Evolution of TFs, Phytohormone Metabolism and Signaling, and Phylogenetic Trees of GRAS and PYL Genes Revealing Putative HGT Events

(A) A heatmap diagram to show the gene presence/absence variation (PAV) and gene copy number variation (CNV) of major transcription factors across different genomes identified by reciprocal best BLAST and further confirmed by HMMER 3.1. The seed sequences were collected from the TAPScan transcription factor database (https://plantcode.online.uni-marburg.de/tapscan/).

(B) Gene identification and comparison of nine different categories of phytohormones and their first appearance in Archaeplastida (gray), Klebsormidophyceae (yellow), Charophyceae (purple), Coleocheatophyceae (orange), Zygnematophyceae (red), or Embryophyta (green). Different shapes of boxes refer to different elements: gene biosynthesis (rectangle), receptor (rounded rectangle), signal transduction components (hexagon), and transcription factors involved in signaling pathways (ellipse). Genes marked by # and $ indicate gene presence in *S. muscicola* and *M. endlicherianum*, respectively. Orthologs and paralogs were first searched by Blastp and confirmed by phylogenetic tree inference (see Method Details).

(C) Origin and diversification of the GRAS gene family. The GRAS domain (PF03514) was used to search GRAS gene sequences across the entire eukaryote (fungi, algae, protists, embryophytes, and animals) and bacteria databases by HMMER search. The phylogenetic tree was built based on the multiple sequence alignment of all GRAS protein domains (including the homologous bacteria-derived SAM-dependent methyltransferases as outgroup) by different programs (RAxML, FastTree, Phylobayes), different inference methods (ML, Bayesian inference), and different models of evolution (FastTree: JTT+CAT; RAxML: GTR+CAT models; Phylobayes: LG, CAT+GTR and CAT+GTR+Dayhoff recoding). The tree shown was obtained using Phylobayes with a CAT+GTR+Dayhoff recoding model based on Bayesian posterior predictive simulations for assessing model adequacy (Method Details).

(D) Phylogenetic tree of the ABA receptor PYL gene family. Seed PYL sequences and domains were collected from the PFAM database, and the same pipeline for gene identification and phylogenetic tree reconstruction as for the GRAS genes was used here. The tree shown was obtained using Phylobayes with a CAT+GTR model. The branch that the putative HGT events occurred is highlighted (arrow with HGT).

See also Figures S3 and S4.

genes linked to processes associated with a terrestrial lifestyle in embryophytes, such as AM symbiosis, have been lost several times independently upon reversal to an aquatic environment (Figure S4).

In conclusion, genome sequence analyses of two early diverging subaerial/terrestrial Zygnematophyceae identified a new lineage in Zygnematophyceae (subclass Spirogloeophycidae) and revealed that many genes essential for embryophytes were present in the common ancestor of Zygnematophyceae and embryophytes. We suggest that the common ancestor of Zygnematophyceae and embryophytes already lived in a subaerial/terrestrial environment and obtained genes from soil bacteria that, after diversification, played an important role in the evolution and radiation of embryophytes, regulating processes from growth and development to defense against biotic and abiotic stresses and to symbiotic interactions.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Sample preparation
  - Modifications of CTAB DNA extraction protocol for S. muscicola and M. endlicherianum
  - Data Quality Control and Genome Survey
  - Transcript Assembly, Unigenes and Gene Expression
  - Genome Annotation
  - Organellar Genome Assembly and Annotation
  - Whole Genome Triplication Detected in S. muscicola
  - Species Phylogeny Tree Reconstruction
  - Gene Family and Species Divergence Time
  - Identification of TFs and TRs
  - Genes and Phytohormones
  - Genes Involved in Cell-wall Related Genes
  - Symbiosis-related Genes
  - Meiosis-specific Genes
  - Flagella-related Genes
  - Resistance Genes
  - Gene Innovations and HGTs
  - PCR Validation on GRAS Gene Presence or Absence
  - Insertion Time Estimation of Bacteria GRAS Genes
- QUANTIFICATION AND STATISTICAL ANALYSES
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2019.10.019.

ACKNOWLEDGMENTS

Part of this work has been conducted at the LRSV laboratory, which belongs to the TULIP Laboratoire d'Excellence (LABEX) (ANR-10-LABX-41). M.M. wishes to thank Robert A. Andersen (Friday Harbor Laboratories, University of Washington, Seattle, WA) for helpful discussions concerning taxonomy and nomenclature of *S. muscicola* and Line Le Gall (Muséum national d'Histoire naturelle, Sorbonne Universités, Paris) for providing access to herbaria and the original watercolor drawings of de Brébisson. This work was supported by National Key R&D Program of China (2018YFA0903202 to S.C.); the Shenzhen Municipal Government of China (JCYJ20151015162041454 to S.C.); the Alberta Innovates AITF/iCORE Strategic Chair (RES0010334 to G.K.-S.W.); the National Natural Science Foundation of China (31700192 to Y.Z.); and the open research project of "Cross-Cooperative Team" of the Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences.

AUTHOR CONTRIBUTIONS

M.M. defined the species to sequence and conceived the project. M.M. and S.C. wrote the manuscript, supplementary texts, and methods. S.C. and G.K.-S.W. designed and coordinated genome sequencing, assembly, annotation, and comparative genomics analyses. S.C. and W.X. worked on the phylogenomics analyses of different genes/gene families and genetic traits of interest under the supervision of M.M. J.K. and P.-M.D. worked on gene phylogenies related to symbiosis. Y.F., X. Li, T.W., Y.Z., S. Wang, and L.L. performed genome assembly, annotation, and evaluation. W.S. worked on organellar genome assembly and annotation. Y.X. worked on transcriptome analysis. X. Liu, X.X., J.W., and H.Y. provided support in sequencing and the bioinformatics platform. S. Wittek and T.R. grew algae to quantity and isolated total RNA and DNA. T.W. performed the PCR validation experiment. G.G. took light micrographs of the two Zygnematophyceae and provided the video of plastid rotation in *S. muscicola*. B. Marin performed phylogenetic analyses using sequence comparisons of plastid- and nuclear-encoded rRNA genes. A.G. isolated strains of *S. muscicola* from natural samples. B. Melkonian provided algal strains from the CCAC (http://www.ccac.uni-koeln.de/) and made strain CCAC 0214 axenic.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 16, 2019
Revised: August 27, 2019
Accepted: October 21, 2019
Published: November 14, 2019

REFERENCES

Arratia, R., Martin, D., Reinert, G., and Waterman, M.S. (1996). Poisson process approximation for sequence repeats, and sequencing by hybridization. J. Comput. Biol. *3*, 425–463.

Bemm, F., Weiß, C.L., Schultz, J., and Förster, F. (2016). Genome of a tardigrade: Horizontal gene transfer or bacterial contamination? Proc. Natl. Acad. Sci. USA *113*, E3054–E3056.

Bergman, C.M., and Quesneville, H. (2007). Discovering and detecting transposable elements in genome sequences. Brief. Bioinform. *8*, 382–392.

Bowman, J.L., Kohchi, T., Yamato, K.T., Jenkins, J., Shu, S., Ishizaki, K., Yamaoka, S., Nishihama, R., Nakamura, Y., Berger, F., et al. (2017). Insights into land plant evolution garnered from the Marchantia polymorpha genome. Cell *171*, 287–304.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C., and Jaffe, D.B. (2008). ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res. *18*, 810–820.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Campbell, M.S., Law, M., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C.J., et al. (2014). MAKER-P:

a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. *164*, 513–524.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. *17*, 540–552.

Cenci, A., and Rouard, M. (2017). Evolutionary analyses of GRAS transcription factors in angiosperms. Front. Plant Sci. *8*, 273.

Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D., Cramer, C.L., and Huang, X. (2015). Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biol. *16*, 30.

Cooper, E., and Delwiche, E. (2016). Green algal transcriptomes for phylogenetics and comparative genomics, figshare. https://figshare.com/articles/Green_algal_transcriptomes_for_phylogenetics_and_comparative_genomics/1604778.

Csurös, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics *26*, 1910–1912.

de Bary, A. (1858). Untersuchungen über die Familie der Conjugaten (Zygnemeen und Desmidieen) (A. Förstnersche Buchhandlung).

de Vries, J., Curtis, B.A., Gould, S.B., and Archibald, J.M. (2018). Embryophyte stress signaling evolved in the algal progenitors of land plants. Proc. Natl. Acad. Sci. USA *115*, E3471–E3480.

Delaux, P.-M., Radhakrishnan, G.V., Jayaraman, D., Cheema, J., Malbreil, M., Volkening, J.D., Sekimoto, H., Nishiyama, T., Melkonian, M., Pokorny, L., et al. (2015). Algal ancestor of land plants was preadapted for symbiosis. Proc. Natl. Acad. Sci. USA *112*, 13390–13395.

Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. *45*, e18.

Domozych, D.S., Sørensen, I., Popper, Z.A., Ochs, J., Andreas, A., Fangel, J.U., Pielach, A., Sacks, C., Brechka, H., Ruisi-Besares, P., et al. (2014). Pectin metabolism and assembly in the cell wall of the charophyte green alga Penium margaritaceum. Plant Physiol. *165*, 105–118.

Doyle, J.A. (2013). Phylogenetic analyses and morphological innovations in land plants. In Annual Plant Reviews Volume 45: The Evolution of Plant Form, B.A. Ambrose and M. Purugganan, eds. (Blackwell Publishing), pp. 1–50.

Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. *29*, 1969–1973.

Emiliani, G., Fondi, M., Fani, R., and Gribaldo, S. (2009). A horizontal gene transfer at the origin of phenylpropanoid metabolism: a key adaptation of plants to land. Biol. Direct *4*, 7.

Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. *16*, 157.

Gitzendanner, M.A., Soltis, P.S., Wong, G.K.S., Ruhfel, B.R., and Soltis, D.E. (2018). Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. Am. J. Bot. *105*, 291–301.

Gontcharov, A.A., and Melkonian, M. (2004). Unusual position of the genus *Spirotaenia* (Zygnematophyceae) among streptophytes revealed by SSU rDNA and rbcL sequence comparisons. Phycologia *43*, 105–113.

Gonzalez, D.H. (2015). Plant Transcription Factors: Evolutionary, Structural and Functional aspects (Academic Press).

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. *27*, 221–224.

Graham, L.E. (1993). Origin of Land Plants (John Wiley & Sons, Inc.).

Guiry, M.D. (2013). Taxonomy and nomenclature of the Conjugatophyceae (= Zygnematophyceae). Algae *28*, 1–29.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. *31*, 5654–5666.

Hahn, C., Bachmann, L., and Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads–a baiting and iterative mapping approach. Nucleic Acids Res. *41*, e129.

Han, G.Z. (2019). Origin and evolution of the plant immune system. New Phytol. *222*, 70–83.

Harholt, J., Moestrup, Ø., and Ulvskov, P. (2016). Why plants were terrestrial from the beginning. Trends Plant Sci. *21*, 96–101.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. *35*, 518–522.

Hori, K., Maruyama, F., Fujisawa, T., Togashi, T., Yamamoto, N., Seo, M., Sato, S., Yamada, T., Mori, H., Tajima, N., et al. (2014). Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. Nat. Commun. *5*, 3978.

Huelsenbeck, J.P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics *17*, 754–755.

Husnik, F., and McCutcheon, J.P. (2018). Functional horizontal gene transfer from bacteria to eukaryotes. Nat. Rev. Microbiol. *16*, 67–79.

Jahan, A., Komatsu, K., Wakida-Sekiya, M., Hiraide, M., Tanaka, K., Ohtake, R., Umezawa, T., Toriyama, T., Shinozawa, A., Yotsui, I., et al. (2019). Archetypal roles of an abscisic acid receptor in drought and sugar responses in liverworts. Plant Physiol. *179*, 317–328.

Jensen, J.K., Busse-Wicher, M., Poulsen, C.P., Fangel, J.U., Smith, P.J., Yang, J.Y., Peña, M.J., Dinesen, M.H., Martens, H.J., Melkonian, M., et al. (2018). Identification of an algal xylan synthase indicates that there is functional orthology between algal and plant cell wall biosynthesis. New Phytol. *218*, 1049–1060.

Jia, Q., Li, G., Köllner, T.G., Fu, J., Chen, X., Xiong, W., Crandall-Stotler, B.J., Bowman, J.L., Weston, D.J., Zhang, Y., et al. (2016). Microbial-type terpene synthase genes occur widely in nonseed land plants, but not in seed plants. Proc. Natl. Acad. Sci. USA *113*, 12328–12333.

Johnson, M.T., Carpenter, E.J., Tian, Z., Bruskiewich, R., Burris, J.N., Carrigan, C.T., Chase, M.W., Clarke, N.D., Covshoff, S., Depamphilis, C.W., et al. (2012). Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. PLoS ONE *7*, e50226.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., et al. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. *24*, 1384–1395.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589.

Karol, K.G., McCourt, R.M., Cimino, M.T., and Delwiche, C.F. (2001). The closest living relatives of land plants. Science *294*, 2351–2353.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

Kent, W.J. (2002). BLAT–the BLAST-like alignment tool. Genome Res. *12*, 656–664.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.

Klümper, U., Riber, L., Dechesne, A., Sannazzarro, A., Hansen, L.H., Sørensen, S.J., and Smets, B.F. (2015). Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. ISME J. *9*, 934–945.

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

Ku, C., and Martin, W.F. (2016). A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70 % rule. BMC Biol. *14*, 89.

Ku, C., Nelson-Sathi, S., Roettger, M., Sousa, F.L., Lockhart, P.J., Bryant, D., Hazkani-Covo, E., McInerney, J.O., Landan, G., and Martin, W.F. (2015). Endosymbiotic origin and differential loss of eukaryotic genes. Nature *524*, 427–432.

Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst. Biol. *62*, 611–615.

Laslett, D., and Canbäck, B. (2008). ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics *24*, 172–175.

Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. *44* (W1), W242-5.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, arXiv:1303.3997v2.

Lohse, M., Drechsel, O., Kahlau, S., and Bock, R. (2013). OrganellarGenomeDRAW–a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res. *41*, W575-81.

Lowe, T.M., and Chan, P.P. (2016). tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. *44* (W1), W54-7.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience *1*, 18.

Lütkemüller, J. (1903). Über die Gattung Spirotaenia Bréb. II. Beschreibung neuer Arten und Bemerkungen über bekannte. Pl. Syst. Evol. *53*, 396–405, 483–488.

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764–770.

Marin, B. (2012). Nested in the Chlorellales or independent class? Phylogeny and classification of the Pedinophyceae (Viridiplantae) revealed by molecular phylogenetic analyses of complete nuclear and plastid-encoded rRNA operons. Protist *163*, 778–805.

Martin, W.F. (2017). Too much eukaryote LGT. BioEssays *39*, 1700115.

McFadden, G., and Melkonian, M. (1986). Use of Hepes buffer for microalgal culture media and fixation for electron microscopy. Phycologia *25*, 551–557.

Melkonian, M., and Weber, A. (1975). Der Einfluß von Kinetin auf das Wachstum von Fritschiella tuberosa Iyeng. (Chaetophorineae, Chlorophyceae) in axenischer Massenkultur. Z. Pflanzenphysiol. *76*, 120–129.

Mikkelsen, M.D., Harholt, J., Ulvskov, P., Johansen, I.E., Fangel, J.U., Doblin, M.S., Bacic, A., and Willats, W.G. (2014). Evidence for land plant cell wall biosynthetic mechanisms in charophyte green algae. Ann. Bot. *114*, 1217–1236.

Moreno-Hagelsieb, G., and Latimer, K. (2008). Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics *24*, 319–324.

Morris, J.L., Puttick, M.N., Clark, J.W., Edwards, D., Kenrick, P., Pressel, S., Wellman, C.H., Yang, Z., Schneider, H., and Donoghue, P.C.J. (2018). The timescale of early land plant evolution. Proc. Natl. Acad. Sci. USA *115*, E2274–E2283.

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

Nishiyama, T., Sakayama, H., de Vries, J., Buschmann, H., Saint-Marcoux, D., Ullrich, K.K., Haas, F.B., Vanderstraeten, L., Becker, D., Lang, D., et al. (2018). The Chara genome: secondary complexity and implications for plant terrestrialization. Cell *174*, 448–464.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. *26*, 1641–1650.

Roberts, A., and Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. Nat. Methods *10*, 71–73.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.

Rogers, S.O., and Bendich, A.J. (1985). Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. Plant Mol. Biol. *5*, 69–76.

Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J., and Walker, A.W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. *12*, 87.

Silva, P.C. (1952). A Review of Nomenclatural Conservation in the Algae from the Point of View of the Type Method (University of California Press).

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212.

Soucy, S.M., Huang, J., and Gogarten, J.P. (2015). Horizontal gene transfer: building the web of life. Nat. Rev. Genet. *16*, 472–482.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. *32*, W309-12.

Takenaka, M., Zehrmann, A., Verbitskiy, D., Härtel, B., and Brennicke, A. (2013). RNA editing in plants and its evolution. Annu. Rev. Genet. *47*, 335–352.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R., and Greiner, S. (2017). GeSeq - versatile and accurate annotation of organelle genomes. Nucleic Acids Res. *45* (W1), W6–W11.

Timme, R.E., Bachvaroff, T.R., and Delwiche, C.F. (2012). Broad phylogenomic sampling and the sister lineage of land plants. PLoS ONE *7*, e29696.

Turland, N.J., Wiersema, J.H., Barrie, F.R., Greuter, W., Hawksworth, D.L., Herendeen, P.S., Knapp, S., Kusber, W.-H., Li, D.-Z., Marhold, K., et al. (2018). International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017. Regnum Vegetabile 159 (Koeltz Botanical Books).

Veeckman, E., Ruttink, T., and Vandepoele, K. (2016). Are we there yet? Reliably estimating the completeness of plant genome sequences. Plant Cell *28*, 1759–1768.

Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., Gitzendanner, M.A., et al. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. USA *111*, E4859–E4868.

Wilhelmsson, P.K.I., Mühlich, C., Ullrich, K.K., and Rensing, S.A. (2017). Comprehensive genome-wide classification reveals that many plant-specific transcription factors evolved in streptophyte algae. Genome Biol. Evol. *9*, 3384–3397.

Wodniok, S., Brinkmann, H., Glöckner, G., Heidel, A.J., Philippe, H., Melkonian, M., and Becker, B. (2011). Origin of land plants: do conjugating green algae hold the key? BMC Evol. Biol. *11*, 104.

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

Yue, J., Hu, X., Sun, H., Yang, Y., and Huang, J. (2012). Widespread impact of horizontal gene transfer on plant colonization of land. Nat. Commun. *3*, 1152.

Zdobnov, E.M., and Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847–848.

Zhang, D., Iyer, L.M., and Aravind, L. (2012). Bacterial GRAS domain proteins throw new light on gibberellic acid response mechanisms. Bioinformatics *28*, 2407–2411.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
S. muscicola genomic DNA and RNA sequencing data	This study	BioProject: PRJNA541068
M. endlicherianum genomic DNA and RNA sequencing data	This study	BioProject: PRJNA541331
S. muscicola, nuclear-encoded SSU rDNA	This Study	Accession MN585752
Experimental Models: Organisms/Strains		
S. muscicola CCAC 0214	Rur-Valley (Eifel, Germany) near Dreistegen/Monschau (coordinates: 50.550693, 6.222094), scraped from the surface of a rock, 2006. This study	Maintained at Central Collection of Algal Cultures (http://www.ccac.uni-koeln.de/).
M. endlicherianum SAG 12.97	Portugal, Lagoa das Bracas (coordinates: 40.243191 / −8.80488) (http://sagdb.uni-goettingen.de/detailedList.php?str_number=12.97)	Maintained at Sammlung von Algenkulturen Göttingen (http://www.uni-goettingen.de/en/184982.html)
Software and Algorithms		
SOAPdenovo2	Luo et al., 2012	https://sourceforge.net/projects/soapdenovo2/
SOAPfilter	Luo et al., 2012	https://sourceforge.net/projects/soapdenovo2/
ALLPATH-LG	Butler et al., 2008	http://software.broadinstitute.org/allpaths-lg/blog/
Platanus	Kajitani et al., 2014	http://platanus.bio.titech.ac.jp
BUSCO	Simao et al., 2015	https://busco.ezlab.org/
Gapcloser	Luo et al., 2012	https://sourceforge.net/projects/soapdenovo2/
Bridger	Chang et al., 2015	https://github.com/fmaguire/Bridger_Assembler
eXpress	Roberts and Pachter, 2013	https://github.com/adarob/eXpress
RepeatMasker	Bergman and Quesneville, 2007	http://www.repeatmasker.org/
MAKER-P	Campbell et al., 2014	http://www.yandell-lab.org/software/maker-p.html
PASA	Haas et al., 2003	https://github.com/PASApipeline/PASApipeline
Augustus	Stanke et al., 2004	http://augustus.gobics.de/
SNAP	Korf, 2004	https://github.com/KorfLab/SNAP
InterproScan	Zdobnov and Apweiler, 2001	https://www.ebi.ac.uk/interpro/search/sequence-search
NOVOPlasty 2.7	Dierckxsens et al., 2017	https://github.com/ndierckx/NOVOPlasty
MITObin v1.8	Hahn et al., 2013	https://github.com/chrishah/MITObim
GeSeq	Tillich et al., 2017	https://chlorobox.mpimp-golm.mpg.de/geseq.html
Blat	Kent, 2002	https://genome.ucsc.edu/FAQ/FAQblat.html
ARWEN v1.2.3	Laslett and Canbäck, 2008	http://www.mybiosoftware.com/arwen-1-2-3-trna-detection-in-metazoan-mitochondrial-sequences.html
tRNAscan-SE v2.0	Lowe and Chan, 2016	http://lowelab.ucsc.edu/tRNAscan-SE/
OGDRAW	Lohse et al., 2013	http://ogdraw.mpimp-golm.mpg.de/
Orthofinder v2.2.6	Emms and Kelly, 2015	https://github.com/davidemms/OrthoFinder
MAFFT v7.3.10	Katoh and Standley, 2013	https://mafft.cbrc.jp/alignment/software/
Gblocks v0.91	Castresana, 2000	http://molevol.cmima.csic.es/castresana/Gblocks_server.html
RAxML version 8	Stamatakis, 2014	https://cme.h-its.org/exelixis/software.html
SeaView 4.3.0	Gouy et al., 2010	http://doua.prabi.fr/software/seaview

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MrBayes v3.2.6	Huelsenbeck and Ronquist, 2001	http://nbisweden.github.io/MrBayes/download.html
Count	Csurös, 2010	http://www.iro.umontreal.ca/~csuros/gene_content/count.html
PAML 4	Yang, 2007	http://abacus.gene.ucl.ac.uk/software.html
Fasttree 2.1	Price et al., 2009	http://www.microbesonline.org/fasttree/
IQ-TREE	Nguyen et al., 2015	http://www.iqtree.org/
Phylobayes-MPI	Lartillot et al., 2013	https://github.com/bayesiancook/pbmpi
tBLASTn	Camacho et al., 2009	http://nebc.nox.ac.uk/bioinformatics/docs/tblastn.html
ModelFinder2	Kalyaanamoorthy et al., 2017	http://www.iqtree.org/ModelFinder/
UltraFast Bootstraps	Hoang et al., 2018	http://www.iqtree.org/
iTOL v4.2.3	Letunic and Bork, 2016	https://itol.embl.de/
BWA 0.7.12-r1039	Li, 2013	http://bio-bwa.sourceforge.net/
IGV v2.4.18	Robinson et al., 2011	https://software.broadinstitute.org/software/igv/
HISAT2	Kim et al., 2015	https://ccb.jhu.edu/software/hisat2/index.shtml
BEAST	Drummond et al., 2012	http://beast.community/
Other		
S. muscicola and *M. endlicherianum* genome assembly and annotation for human curation	This study	Figshare: https://figshare.com/articles/Genomes_of_subaerial_Zygnematophyceae_provide_insights_into_land_plant_evolution/9911876/1
Data S1 and S2 for human curation	This study	Mendeley Datasets http://dx.doi.org/10.17632/pvf47s35xy.1

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Michael Melkonian (michael.melkonian@uni-koeln.de).

This study did not generate new unique reagents. For availability of algal strains see EXPERIMENTAL MODEL AND SUBJECT DETAILS.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

A natural sample containing *Spirogloea muscicola* was collected in September 2006 by K.-H. Linne von Berg (Cologne) in the Rur-Valley (Eifel, Germany) near Dreistegen/Monschau (coordinates: 50.550693, 6.222094), scraped from the surface of a rock. Samples were spread on 1.5% (w/w) agar plates supplemented with culture medium WarisH (McFadden and Melkonian, 1986) and incubated for several weeks at 15°C in a walk-in, temperature-controlled chamber at low light (10 μmol photons m^{-2} s^{-1}) in a 14:10 h L/D cycle. Single cell-derived algal colonies were removed with a micropipette, transferred to glass tubes with sterile culture medium and grown under the same culture conditions. Cultures were made axenic: after low-intensity ultrasonication and washing by centrifugation to remove bacteria and mucilage, cells were sprayed onto an agar plate (1% agar; WarisH:BSM [McFadden and Melkonian, 1986; Melkonian and Weber, 1975] 100:1). Single axenic cells were transferred to a new agar plate. *S. muscicola* (strain CCAC 0214) is available from the Central Collection of Algal Cultures (http://www.ccac.uni-koeln.de/). Axenic cultures of *Mesotaenium endlicherianum* (strain SAG 12.97) were obtained from the Sammlung von Algenkulturen Göttingen (http://www.uni-goettingen.de/en/184982.html). Algae were grown to quantity in aerated (ambient) 10 L glass bottles at 40 μmol photons m^{-2} s^{-1} (fluorescent light tubes: L36W/640i energy saver cool white and L58W/956 BioLux, Osram, Munich, Germany) in a 14/10 hr L/D cycle. Algae were harvested near the end of log-phase growth by low-speed centrifugation (300 x *g*). During all steps of culture scale-up until nucleic acid extraction, axenicity was monitored by sterility tests as well as light microscopy. Total RNA was extracted from *S. muscicola* and *M. endlicherianum* using the CTAB-PVP Method as described in Johnson et al. (2012). Total DNA was extracted using a modified CTAB protocol (Rogers and Bendich, 1985). Light microscopy was performed with a Leica DMLB light microscope using a PL-APO 100/1.40 objective, an immersed condenser N.A. 1.4 and a Metz Mecablitz 32 Ct3 flash system. Time lapse video microscopy of chloroplast rotation in *S. muscicola* (see Video S1) was done with the same microscope using a Blackmagic Micro Studio Camera 4K with Atomus Shogun video recorder. The time lapse was 6x (3 min reduced to 30 s).

METHOD DETAILS

Sample preparation

Modified CTAB protocol for DNA extraction from Spirogloea muscicola and Mesotaenium endlicherianum

Pre-treatment of biomass:

1) Removal of external polysaccharides by sonication and washing
2) Centrifugation of cells to a dense pellet
3) Freezing in liquid nitrogen and opening of cells in the frozen state with a tissue lyser and steel capsules (containing a large steel) 2 times at 28/s for 2 min
4) Maximum 3 mL frozen powder per 15 mL Falcon tube
5) Storage of biomass at −80°C until CTAB-extraction

It is important that once the biomass is frozen it never melts until extraction.

CTAB-extraction protocol:

DNA extraction via CTAB method reagents

TE buffer: 10 mM Tris-HCl (pH 8), 1mM EDTA, sterile bidistilled water 5% SDS solution

proteinase K solution: 10 mg/ ml

RNase: 10 mg/ml

5M NaCl

CTAB buffer (pre-warmed at 65°C): 1.4 M NaCl, 0.1 M Tris-HCl (pH 8), 25 mM EDTA, 2% CTAB (w/v), sterile bidistilled water

24:1 chloroform: isoamyl alcohol

Isopropanol/ 2-propanol (pre-cooled at −20°C)

80% ethanol (pre-cooled at −20°C)

Protocol:

1, harvest algae and wash pellet several times with respective medium (centrifugation at lowest-possible speed)
2, homogenize pellet (≈1 ml) with liquid nitrogen
3, transfer homogenized powder very quickly into a sterile falcon tube containing a mixture of 900 μL TE buffer, 700 μL 5% SDS and 25 μL proteinase K solution and vortex immediately
4, incubate in water bath at 60°C, 20‘ (vortex occasionally)
5, add 500 μL 5 M NaCl, 25 μL RNase A and 5 mL CTAB buffer (pre-warmed at 65°C) and vortex
6, incubate in water bath at 60°C, 10‘(vortex occasionally)
7, centrifuge the sample (3,000 *g*, 15’) and transfer supernatant into a new sterile Falcon tube
8, add equal volume of 24:1 chloroform: isoamyl alcohol and vortex rigorously
9, incubate at RT, 10’
10, separate phases by centrifugation (3,000 *g*, 15’)
11, transfer and portion the upper, aqueous phase carefully into sterile 1.5 mL Eppendorf tubes (if the aqueous phase is not clear, the chloroform-isoamylalcohol extraction has to be repeated)
12, maintain DNA extracts on ice and perpetuate this condition
13, add 2/3 volume of isopropanol (pre-cooled at −20°C) and vortex
14, incubate at −20°C, at least 1 h (or overnight)
15, centrifuge (17,000 *g*, 15’, 4°C) and discard the supernatant carefully
16, wash pellet with 1 mL 80% ethanol (pre-cooled at −20°C) and vortex
17, centrifuge (17,000 *g*, 5′, 4°C) and discard the supernatant carefully
18, repeat the washing step
19, air-dry the pellet and dissolve the pellet in 50 - 200 μL TE buffer
20, store at −20°C

Modifications of CTAB DNA extraction protocol for S. muscicola and M. endlicherianum

3, mix the following components per ml frozen cell powder in a 15 mL Falcon tube: 1.5 mL TE-buffer, 0.75 mL SDS (5%), 25 μL proteinase K (10 mg/ml), 0.5 mL sodium acetate (3M) and incubate at 55°C for 10 min. Add the frozen biomass-powder to the heated extraction mix and immediately shake to prevent clumping of frozen biomass that have no contact to the extraction buffer (will otherwise cause degradation of DNA)
4, incubation at 55°C for 20 min and shaking of the tubes every 5 min
5, mix the following components per ml frozen cell powder in a separate Falcon tube: 1 mL NaCl (5M), 25 μL RNase A (10 mg/ml), 5 mL CTAB-buffer and incubate at 60°C for 10 min. Add biomass-extraction buffer mix of step 3 to heated extraction mix of step 5 and incubate at 60°C for 10 min with shaking every 2 min.

7, Centrifugation of Falcon tubes at room temperature at 4,500 rpm for 1h. Decant clear extract into a new Falcon tube and add equal volume of 24:1 chloroform:isoamylalcohol. Mix regularly and incubate for 20 min.
10, Centrifugation at room temperature for 1h at 4,500 rpm, transfer clear supernatant into a new Falcon tube and repeat washing step with chloroform:isoamylalcohol until no interphase is visible.
15, Centrifuge at 4°C for 2h at 4,500 rpm (a DNA pellet must be visible afterward, if not, add again the same volume of isopropanol, incubate and centrifuge again).
19, it is extremely important that DNA pellets are completely dry before the addition of water (otherwise the 260/230 nm absorption ratio will be very low). Appearance of DNA changes from white to clear.
Library Construction and Sequencing

A strategy of hierarchical DNA libraries with different insert sizes was applied, which typically includes a combination and complementation of multiple pair-end libraries with insert sizes of 170bp∼800bp and mate-pair libraries of 2∼20kb. Each library was constructed for whole genome shotgun sequencing with the Illumina platform (Hiseq2000 and Hiseq4000) following the company's protocol. Large DNA fragments are required for the preparation of mate-pair libraries, these sequencing data is designed for scaffolding to connect assemble contigs, basically utilizing varied long insert-size fragments that would able to extend over different kinds of repeat regions. For *S. muscicola*, 100.52 Gb raw data were generated, resulting in 577X sequencing depth and were further reduced to 354X clean data (61Gb) after data quality filtering (see below) for genome assembly. For *M. endlicherianum*, 162.51 Gb raw data was generated and the sequencing depth is about 993.20 X. The sequencing depth after quality filtering is 412.52 X, approximate 67.50 Gb clean data was taken as input for genome assembly. All information on library construction and sequencing is summarized in Table S1.

Data Quality Control and Genome Survey

To minimize the sequencing error rate and avoid assembly artifacts, a strict quality control was performed following the "reads filtering" protocol by SOAPfilter (Luo et al., 2012), including filtering out N-rich reads (removing those reads with $\geq$ 10% of "N" bases); low-quality reads (those reads with $\geq$ 40% of their bases as low-quality which are defined as base quality score $\leq$ 7); PCR duplicates; and Read ends trimming process (low-quality bases at the end of reads were trimmed out).

The genome size was estimated using 17-mer analysis. A k-mer refers to a continuous sequence with k base pairs, typically extracted from the reads (he k size is usually smaller than the length of a read, e.g., 17 bases per k-mer). For a typical K-mer analysis, it assumes that during a randomly whole-genome sequencing process, the start position of sequencing reads along the whole genome will follows a Poisson distribution (Arratia et al., 1996; Marçais and Kingsford, 2011; Veeckman et al., 2016) if ideally no sequencing errors or sequencing coverage bias from the sequencing dataset. Based on this, we can simulate and measure the occurrence and frequency of all kinds of continuous sequence with k base pairs (k-mers) from the sequencing reads generated and observe some patterns (distribution with peaks of the sequencing depth) to estimate the basic genome characters from the computation perspective using statistics of the Poisson distribution model. Basically, we can estimate that: Genome size = k-mer number/average sequencing depth. For *S. muscicola*, the peak depth is about 43 and the total K-mer count is 7,493,127,226. The genome size was estimated as 174.26 Mb. For *M. endlicherianum*, the peak depth is about 70 and the total K-mer count is 1,145,371,4030. The genome size was estimated as 163.62 Mb. There is a small peak at the two-fold position of the *S. muscicola* K-mer curve, indicating a rich repeat content or a gene/genome duplication burst. On the other hand, no other peak besides the main peak in the *M. endlicherianum* K-mer curve, suggests that *M. endlicherianum* is a simple genome.

Multiple genome assemblers were applied for the whole genome assembly of both Zygnematophyceae genomes, including SOAPdenovo2 (Luo et al., 2012), ALLPATH-LG (Butler et al., 2008), and Platanus (Kajitani et al., 2014). After several rounds of genome assembly and evaluation according to assembly contiguity (N50) and genome completeness (BUSCO and RNA mapping) of the assembled contigs (Simao et al., 2015), the best assemblies were selected for the downstream gapfilling step by Gapcloser (Luo et al., 2012) (version 1.2).

Transcript Assembly, Unigenes and Gene Expression

To facilitate gene model annotation, gene expression analysis, and organellar gene prediction, two different libraries were constructed both for *S. muscicola* and *M. endlicherianum*, respectively; one library is by poly-A selection, to sequence and measure transcripts expressed in the nucleus; the other type of RNA library is by rRNA depletion (https://www.illumina.com/products/by-type/molecular-biology-reagents/ribo-zerorrna-removal-plant.html), which is designed to qualify and quantify transcripts encoded and expressed both from organellar and nuclear genes. The rRNA depletion library was also used to explore RNA editing events in the organellar transcripts by mapping RNaseq reads against their own (organellar) genomes (Takenaka et al., 2013). Sequencing reads of each RNA library were used as input into Bridger (Chang et al., 2015) for transcript assembly. Unigenes were generated from the transcript annotations taking the Plant Refseq (https://www.ncbi.nlm.nih.gov/refseq/) database as reference. We used express 1.5.1 (Roberts and Pachter, 2013) to estimate gene expression of the RNA poly-A library and RNA rRNA libray, respectively.

Genome Annotation

Repetitive elements were identified and analyzed by RepeatMasker (Bergman and Quesneville, 2007) (version 4-0-5). Self-training custom repeat libraries, including LTR, MITE, and TRIM repeat libraries *de novo* predicted by RepeatModeler (version 1-0-8), were generated independently and then integrated into a nonredundant custom library. This custom library was prepared as the input for RepeaMasker to predict each type of transposable element, resulting in 27.5% and 33.2% of the two genomes, *Spirogloea muscicola* and *Mesotaenium endlicherianum*, being defined as repeat regions. The predominant composition of the repeat content is the long terminal retrotransposon.

MAKER-P (Campbell et al., 2014) pipeline was implemented for gene annotation in two rounds of interactions, by integrating multiple annotation resources. First, RNA-aided gene model building is a pivotal step, during which a set of representative assembled transcripts were mapped back against the masked genome assembly using PASA (Haas et al., 2003), producing a set of complete protein-coding gene models. Parameter training was performed on Augustus (Stanke et al., 2004) for such a set of complete gene models. Homolog-based gene prediction was completed by selecting closely-related green algal genomes as well as those of well-annotated model species, such as *Chara braunii*, *Klebsormidium nitens*, *Chlamydomonas reinhardtii*, *Marchantia polymorpha*, *Physcomitrella patens*, and *Arabidopsis thaliana*. Gene models and gene characteristics derived from RNA-aided and protein-based pipelines were also used to train the SNAP (Korf, 2004) pipeline, a *de novo* prediction software. Finally, MAKER-P combined all of these sources to annotate genes for both Zygnematophyceae genomes. Gene annotation revealed a markedly different protein-coding gene number, with 27,137 for *S. muscicola* and 11,080 for *M. endlicherianum*, and both were supported by the assembled transcripts (with 78.3% and 73.5% of the predicted genes), and BUSCO (with 87.8% and 80.2% of the predicted genes), respectively. Gene functional prediction and assignment was carried out by the HMMER-based InterproScan (Zdobnov and Apweiler, 2001) (v5.11). Domains and motifs were searched against the PFAM database (https://pfam.xfam.org/). Protein functional annotation was based on the Swissprot functional database (https://www.ebi.ac.uk/uniprot). Pathway discovery and gene components involved was predicted and assigned by the KEGG database (https://www.genome.jp/kegg/pathway.html).

Organellar Genome Assembly and Annotation

One pair-end library (PE100) with insert size 170bp (with sequencing overlap) was used for both chloroplast and mitochondria genome assembly. The sequencing depth for the entire nuclear genome (170Mb) was > 100X, which indicated that the data were sufficient for organellar genome assembly because of the much smaller genome size (~100kb). A strict data filtering process was carried out first by CLC (https://www.qiagenbioinformatics.com/products/clc-assembly-cell/), and SOAPfilter v2.2, to filter out low-quality reads. Parameters for CLC were: clc_adapter_trim -c 10 -m 50; clc_remove_duplicates -r -s -o; clc_quality_trim -m 75 -p -r. Parameters for the SOAPfilter were: -y -i -z -g 1 -p -M 2 -o. NOVOPlasty 2.7 (Dierckxsens et al., 2017) (https://github.com/ndierckx/NOVOPlasty) was implemented for chloroplast and mitochondria genome assemblies of both *S. muscicola* and *M. endlicherianum*. However, for the mitochondria genome assembly of *M. endlicherianum*, an improved version was obtained by MITObin (Hahn et al., 2013) v1.8 An online annotator GeSeq (Tillich et al., 2017) was used for organellar gene annotation (https://chlorobox.mpimp-golm.mpg.de/geseq.html), in which Blat (Kent, 2002) was implemented to search for coding sequencing (protein-coding and rRNA genes) and build HMM gene models. The results predicted from ARWEN v1.2.3 (Laslett and Canbäck, 2008) and tRNAscan-SE (Lowe and Chan, 2016) v2.0 were integrated for rRNA identification. The circular annotation figures for both organellar genomes were plotted by OGDRAW (Lohse et al., 2013) (http://ogdraw.mpimp-golm.mpg.de/).

Whole Genome Triplication Detected in S. muscicola

Comparison of the two Zygnematophyceae genomes revealed that most (> 99%) clustered orthogroups are shared, with only 30 genes from 9 orthogroups and 58 genes from 16 orthogroups uniquely presented in *M. endlicherianum* and *S. muscicola*, respectively (Data S1N). Meanwhile, while shared orthologous gene pairs in a ratio of 1:1 between *S. muscicola* and *M. endlicherianum* are obvious, we observed that the predominant pattern with orthologous gene ratio between *S. muscicola* and *M. endlicherianum* is 3:1 (3,293 clusters) (Data S1L). These pairwise comparisons unambiguously indicate massive gene duplications in *S. muscicola* genome since their divergence from the last common ancestor, which is consistent with the fact that the total number of the annotated genes in *S. muscicola* is almost tripled of that in *M. endlicherianum*, and consistent with the observation that 75.9% of *S. muscicola* annotated genes are duplicated in BUSCO evaluation while for *M. endlicherianum* only 1.3% was identified (Table S1F). Furthermore, we performed whole genome alignment within and between the two Zygnematophyceae genomes, unraveling large-scale triplicated syntenic blocks within *S. muscicola* (Figure 2A); we also recovered hundreds of 3:1 triplicated orthologs between *S. muscicola* and *M. endlicherianum*, but no or only a few of syntenic blocks within *M. endlicherianum* were detected. Age distribution of paralogous gene pairs were presented (Figure 2C), confirming a recent whole genome triplication event exclusively occurred in *S. muscicola*.

Furthermore, a large fraction set of homologs, which were clustered from orthogroups that maintain triplicated genes in *S. muscicola* while only corresponding to one copy in *M. endlicherianum*, were further analyzed. Syntenic and collinear blocks within and between genomes were generated by McscanX, which were further analyzed by an in-house customized pipeline to cluster and categorize duplicated segments within *S. muscicola* genome. Pseudo-subgenomes were made manually by combing the self-alignments within *S. muscicola* and inter-alignment between *S. muscicola* and *M. endlicherianum* (Data S1M), for which, seed scaffolds (basically three copies) were extended based on the overlap of gene anchors with other groups of triplicated segments. During the

extension process, a random selection was made if no overlap connecting different groups of triplicated segments; singleton scaffolds without duplicated segments were abandoned for the Pseudo-subgenomes construction.

Species Phylogeny Tree Reconstruction

To reconstruct the species phylogeny tree, we applied two approaches: one is based on the concatenated super-genes constructed from the low-copy orthologs clustered from the published whole genomes; the other is based on the bait sequences from the rRNA operons encoded by the nucleus and the plastid, which were annotated from the extended sampling from 1KP dataset (1,000 plant transcriptomes, https://sites.google.com/a/ualberta.ca/onekp/), retrieved from Genbank or newly sequenced for this study. First, we built a multigene phylogenetic tree (as shown in Figure 1B). We selected 16 taxa that represented the major green lineages across Viridiplantate genomes and 3 genomes from Glaucophyta and Rhodophyta. These species include *Arabidopsis thaliana*, *Oryza sativa* Japonica Group, *Ginkgo biloba* L., *Azolla filiculoides*, *Selaginella moellendorffii*, *Physcomitrella patens*, *Marchantia polymorpha*, *Mesotaenium endlicherianum*, *Spirogloea muscicola*, *Chara braunii*, *Klebsormidium nitens*, *Ostreococcus tauri*, *Micromonas pusilla* CCMP1545, *Chlamydomonas reinhardtii*, *Volvox carteri*, *Chlorella* NC64A, and the 3 outgroups: *Cyanidioschyzon merolae*, *Chondrus crispus*, and *Cyanophora paradoxa*. Orthofinder (Emms and Kelly, 2015) (version 2.2.6) was used for gene family clustering. Finally, 85 low-copy orthogroups (one or two gene copies of each orthogroup for each genome) were selected to construct a phylogenetic tree. Orthologous sequences were aligned using multiple sequence alignment by MAFFT (Katoh and Standley, 2013) (version 7.3.10) across lineages and were concatenated into one super-sequence for each species. Poor quality alignments were filtered out by Gblocks (Castresana, 2000) (version 0.91b), and only conserved regions were retained. RAxML (Stamatakis, 2014) (version 8) was implemented for tree building, using the LGX4 amino acid substitution model.

In addition, we built a phylogenetic tree based on the rRNA operons from an extended species sampling (as shown in Figure 1C). 9 strains of streptophyte algae and one hornwort (*Phaeoceros* sp.) were used for determination of new rRNA sequence data in this study (accession numbers in bold in Table S1Y). Origins of algae were: CCAC = Central Collection of Algal Cultures (http://www.ccac.uni-koeln.de/) (from 2019: University of Duisburg-Essen); CCAP = Culture Collection of Algae and Protozoa, UK (https://www.ccap.ac.uk/); CCMP = The Provasoli-Guillard National Center for Culture of Marine Phytoplankton (https://ncma.bigelow.org); NIES = Microbial Culture Collection at National Institute for Environmental Studies, Tsukuba, Japan (http://www.nies.go.jp/biology/mcc/home.htm); SAG = Sammlung von Algenkulturen, University of Göttingen, Germany (https://www.epsag.uni-goettingen.de/html/sag.html); SCCAP = The Scandinavian Culture Collection of Algae and Protozoa at the University of Copenhagen (http://www.sccap.dk/search/), now hosted by the NORCCA (https://niva-cca.no/); UTEX = University of Texas Culture Collection of Algae, USA (http://www.bio.utexas.edu/research/utex/).

New sequence data of rRNA operons were generated as previously described (Marin, 2012). In addition to annotated rRNA sequences from GenBank (https://www.ncbi.nlm.nih.gov/genbank/), complete rRNA operons were assembled from non-annotated transcriptome sequence data (JGI, ONE_KP; see Table S1Y). Sequences were aligned manually on the basis of rRNA/ tRNA secondary structures using SeaView 4.3.0 (Gouy et al., 2010) (http://doua.prabi.fr/software/seaview). For phylogenetic analyses, 9,618 unambiguously aligned nucleotide (nt) positions were used, defined as 4 large divisions: 18S rDNA (1,785 nt), 5.8S and 28S rDNA (3,437 nt), 16S rDNA and two tRNA genes (1,584 nt), and 23S rDNA (2,812 nt). Tree reconstructions were performed by Maximum Likelihood (RAxML v8.2.10; 100 bootstrap replicates with 100 starting trees respectively, using the GTRCAT model), and a Bayesian analysis (MrBayes [Huelsenbeck and Ronquist, 2001] v3.2.6; 1,000,000 generations using the GTR+I+G model). Bootstrap percentages below 50% and Bayesian posterior probabilities below 0.9 were regarded as 'unsupported'.

Gene Family and Species Divergence Time

Orthogroup analysis was performed based on the clustered orthogroups generated by Orthofinder from the selected 16 represented green lineages as described above, results were summarized in Tables S1K and S1L. The Count (Csurös, 2010) software was implemented (Wagner parsimony algorithm, with a weighted gene gain penalty of 1.2) to infer ancestor- and lineage-specific changes in the key nodes on orthogroup gains, losses, expansions, and contractions, basically along the evolutionary phylogenetic tree (derived from Figure 1B) by ancestral reconstruction. The definitions for orthogroup gains, losses, expansions and contractions were described using a posterior probability in the User's Guide of Count software: ''COUNT: Evolutionary Analysis of Phylogenetic Profiles and Other Numerical Characters User's Guide,'' in which, the ancestral state was constructed and compared with the closest outgroup using a phylogenetic birth-and-death statistic model. Note that orthogroups clustered by Orthofinder do not overlap completely with the functional gene families or subfamilies for each species. Therefore, further manual verification and confirmation analysis was performed for each orthogroup (Table S1M) indicative of gene innovation in the common ancestor of Zygnematophyceae and embryophytes, by Blast (< 1e-5) and HMMER search (< 1e-5); the latter was based on either a customized alignment matrix by HMMERbuild from a carefully-filtered multiple sequence alignment, or a matrix built using the known Pfam domains reported elsewhere, to retrieve false negatives and filter out false positives. For a selected set of gene families, that were addressed in the main text, a detailed phylogenetic analysis was performed to further verify ortholog/paralog relationships (see below). Finally, for the putative gained orthogroups from the common ancestor of Zygnematophyceae and embryophytes, a criterion must be defined with > = 1 gene present in embryophytes and > = 1 gene present in Zygnematophyceae and 0 gene present in other early-diverging algal genomes. On the other hand, for the expanded orthogroups from the common ancestor of Zygnematophyceae and embryophytes (Table S1N), all lineages in Zygnematophyceae and embryophytes should contain gene numbers larger than that in

the species that diverged earlier (if both *Chara braunii* and *Klebsormidium nitens* have genes < 2 copies), or ≥ 2-fold than that of the early-diverging lineages (if *Chara braunii or Klebsormidium nitens* have genes > 2 copies).

To estimate the divergence time between the two Zygnematophyceae species (*M. endlicherianum* and *S. muscicola*), and between Zygnematophyceae and embryophytes, we used mcmctree-4.5 (http://abacus.gene.ucl.ac.uk/software/paml.html) implemented in the PAML (Yang, 2007) package, a Markov Chain Monte Carlo process (MCMC) was run for 1,000,000 generations. 85 low-copy orthogroups (one or two gene members for each selected genome) were used and aligned first by protein sequences and then translated into nucleotide coding sequence alignments, based on which the four-degenerate sites of the genes were used as input for MCMCTREE. Convergence was checked by two independent runs. Multiple fossil times were used for time calibrations from Timetree (http://www.timetree.org/): 1) the *Arabidopsis thaliana* and *Oryza sativa Japonica* divergence time (148∼173 million years ago); 2), the *Ginkgo biloba*_L and *Arabidopsis thaliana* divergence time (330∼365 million years ago); 3), *Selaginella moellendorffi* and *Azolla filiculoides* divergence time (392∼432 million years ago). The fourth constraint used for time calibration is the estimated time of the origin of embryophytes (473∼514 million years ago), cited from Morris et al. (2018).

Identification of TFs and TRs

For the identification and classification of transcription factors and transcriptional regulators, we generally used the TAPscan (https://plantcode.online.uni-marburg.de/tapscan/) database as protein sequence reference. First, we selected the most representative sequences from several model land plants (like *Arabidopsis thaliana*, *Oryza sativa, Physcomitrella patens*), and then performed multiple sequence alignments to define the conserved domains or selected the corresponding conserved functional domains in the PFAM database for each category of gene family to establish a domain matrix for HMMER search (http://hmmer.org/) across different lineages. For example, for the GRAS gene family, the conserved PFAM GRAS domains were used to search against the genome databases from bacteria, viruses, fungi, protists, algae, embryophytes and animals. GRAS domain sequences were extracted for multiple sequence alignment by MAFFT, and detailed maximum likelihood phylogenetic trees were built by different software (FastTree [Price et al., 2009], IQtree [Nguyen et al., 2015], RAxML, and Phylobayes-MPI [Lartillot et al., 2013]), with different evolutionary models (models: LG4X, CAT + GTR, and CAT + GTR + Dayhoff recoding). For the HD-KNOX gene family, a phylogeny-based approach was used to distinguish HD-KNOX1 and HD-KNOX2 (these two subfamilies share high sequence similarity). Three conserved domains, Homeobox (PF00046), KNOX1 (PF03790), and KNOX2 (PF03791), were prerequisite to define the HD-KNOX gene family, while the separation of HD-KNOX1 and HD-KNOX2 subfamilies was made by introducing other HD subfamily sequences as ''outgroups'' in a phylogenetic tree by the IQtree software (maximum likelihood algorithm). For the BBR/BPC gene family, the PFAM domain GAGA-binding (PF06217) was used by HMMER search across different genome databases, followed by multiple sequence alignment and phylogenetic analysis as described above. For more details, see Table S1O.

Therefore, for TFs or TRs that have been extensively studied elsewhere, with clear signature domains that could be used to define gene homology, e.g., reported in functional experiments whenever a conserved domain is known to define the gene in question, a HMM-based domain search was the first step. This was followed by a phylogenetic inference, to further evaluate homology and help distinguish orthologs from paralogs. We acknowledge that this approach can miss genes that either lost domains or have not yet evolved domains, but sequence-wise may still be homologous to the query. Therefore, our conclusions regarding the presence/absence of these genes should be regarded as a conservative estimate.

For many other TFs and TRs, with no conserved domains detected, a homolog-based search by Blastp (e-value < 1e-5) was first used, followed by a detailed phylogenetic analysis to infer ortholog/paralog relationships, or to confirm duplication/speciation events where necessary.

This ''HMMER + Phylogeny'' strategy was carefully applied here and for genes involved in phytohormones (see below), in cell-wall biosynthesis and signaling (see below). The ''Blast + Phylogeny'' strategy was also used for hundreds of other candidate genes to survey a consistent homologous gene set, such as genes involved in symbiosis, in which a careful evaluation of the presence/absence of specific domains was also examined where needed.

Genes and Phytohormones

Multiple approaches were combined to identify and confirm presence, absence, and gene copy number variation, for genes involved in phytohormone biosynthesis, signaling and metabolism. The methodology is similar to that described in the section above.

AUXIN: seed sequences (queries) were collected from *Arabidopsis thaliana*, and a primary Blastp search (e-value < 1e-10, similarity ≥ 60%, aligned coverage ≥ 60%) were performed to survey algal and land plant genomes. The query of TAR was AT1G23320; the query of YUC was AT4G32540; the query of TIR was AT3G62980; the query of AUX/IAA was AT2G38120; the query of ARF was AT1G59750. Conserved domains were defined either through multiple sequence alignment of the homolog candidate, or from PFAM domain recovery by functional annotation. HMMER search was used to confirm each of the homologous genes. For example, F-box (PF00646) or F-box-like (PF12937) was a prerequisite domain defined by HMMER to identify TIR genes. For AUX/IAA, the domain AUX_IAA PF02309 must be confirmed by HMMER but domain Auxin_resp PF06507 must be excluded (absence) by HMMER. For ARF, both the Auxin_resp (PF06507) domain and the AUX_IAA (PF02309) domain should be found by HMMER. Furthermore, a detailed phylogenetic approach (FastTtree, maximum likelihood algorithm) was implemented for most of the gene families (TIR, AUX/IAA, ARF) to confirm the genes PAV (gene presence/absence variation) and CNV (gene copy number variation). A similar procedure was implemented for genes involved in ABA, CK, ETH, GA, JA, SL, BR, and SA. Many of these genes were further

analyzed by multiple sequence alignment, a HMMER approach, functional domain annotation, and detailed phylogenetic analysis. For details see Table S1P, and Data S1R–S1AE.

For some gene families, it is difficult to definitively establish gene homology and retrieve all homologous gene copies over long evolutionary distances, mostly because of functional domain recruitment, shuffling or loss that led to neo-functionalization or sub-functionalization of the gene in question. In this sense, functional experiments would be required to verify the possible (conserved) gene function.

Genes Involved in Cell-wall Related Genes

The methodology here is similar to that described above. Blastp was the first step to search for homologous sequence signals. For ambiguous homologs, a characteristic conserved domain (if detected) was defined by multiple sequence alignment and built by HMMbuild, which was used as input for HMMER search to survey further the algal and land plant genome databases. The query sequences were collected from *Arabidopsis thaliana*, for which, CSLC (GT2), queries of XXT (GT34), MUR3 (GT47), FUT1 (GT37), FXG1, AXY8 (GH95), β-Gal10 (GH35), β-G (GH1), XYL1 (GH31), and XTH (GH16) were AT3G28180.1, AT3G62720.1, AT2G20370.1, AT2G03220.1, AT1G67830.1, AT4G34260.1,AT5G63810.1, AT1G52400.3, AT1G68560.1, and AT3G44990.1, respectively. A similar procedure was implemented for genes involved in xylan, and pectin metabolism. Many of these genes were further analyzed by multiple sequence alignment, domain verification by PFAM or HMMER search, and detailed phylogenetic analysis as illustrated in the supplemental Data figures. For details see Table S1Q, and Data S1AF–S1AM.

Symbiosis-related Genes

Genomes of *M. endlicherianum* and *S. muscicola* were screened for 15 genes involved in arbuscular mycorrhizal symbiosis (AMS). Genes were searched using the tBLASTn (Camacho et al., 2009) 2.7.1+ with the default parameters and an e-value threshold of 1e-10. For large gene families, such as *STR*, *MLD-RLK*, the threshold was set to 1e-30. Blasts were performed against a database composed of representative genomes of vascular plants, Zygnematophyceae and close relatives (*Chara braunii* and *Klebsormidium nitens*) as well as the moss *Physcomitrella patens*, the liverwort *Marchantia polymorpha* and 23 transcriptomes of liverworts from the 1KP project. Coding sequences were predicted from the transcriptomes using the TransDecoder suite (http://transdecoder.github.io). The retained sequences were aligned using MAFFT v7.380 and alignments cleaned using GBlocks with parameters set for a less stringent selection than default and allowing smaller final blocks, gap positions within final blocks and less strict flanking positions. The alignments obtained were subjected to phylogenetic analysis using Maximum Likelihood. The best-fitting evolutionary model was determine using ModelFinder2 (Kalyaanamoorthy et al., 2017), and trees were reconstructed using IQ-TREE v1.6.7. Branch supports were tested with 10,000 replicates of UltraFast Bootstraps (Hoang et al., 2018). Trees were visualized using the iTOL v4.2.3 (Letunic and Bork, 2016) platform. The phylogenetic tree for each of the major symbiosis-related genes was summarized in Data S1AQ–S1AX. Sequences of symbiosis-related genes were summarized in Table S1U.

Meiosis-specific Genes

To study the presence and absence of meiosis-related genes, we retrieved a set of query sequences from land plants, and identified 11 meiosis-specific genes (Table S1R). For comparison, we surveyed algal and embryophyte genome databases, for homologous signals, and further used RBH blast (Reciprocal best hit) (Moreno-Hagelsieb and Latimer, 2008) to confirm the presence or absence of each gene.

Flagella-related Genes

To study putative genes involved in the biosynthesis and regulation of flagella in the two Zygnematophyceae, we collected 62 core genes classifying them into 7 categories (Radial Spoke protein, central pair protein, Outer Dynein protein, Inner Dynein protein, Intraflagellar Transport protein, Dynein protein, Basal Body protein) from *Chlamydomonas reinhardtii*, and identified homologs from a wide-range of species among algae, protists, and embryophytes for comparison. The species used here (with genomes available) including dinoflagellates, Chlorarachniophyta, Cryptophyta, Ochrophyta, Glaucophyta, Rhodophyta, Chlorophyta, streptophyte algae and embryophytes. Reciprocal Best Hit (e-value < 1e-5) and HMMER searching was combined to identify and confirm the presence/absence of each gene sequence across different lineages.

Resistance Genes

Two kinds of R genes were surveyed: TIR-NBS-LRR proteins (TNLs) and non-TIR-NBS-LRR proteins (nTNLs). Pfam domains were collected as queries. The NB-ARC (PF00931) domain was used as a domain profile by HMMER search to identify conserved domains across different algal and land plant genomes. For LRR-RLK, both the LRR domain (PF00560) and the KD domain (PF00069) must be defined by HMMERsearch, which were then confirmed by TMHMMv.2.0 (http://www.cbs.dtu.dk/services/TMHMM/) to check for the presence of transmembrane domains (TMs). The result on LRR-RKL/R genes was summarized in Tables S1S and S1T.

Gene Innovations and HGTs

Combining the gene clustering analysis by Orthofinder, we further focused on all annotated protein sequences from the two Zygnematophyceae species (*M. endlicherianum* and *S. muscicola*) to confirm manually a high-confidence set of gene innovations in the

common ancestor of Zygnematophyceae and embryophytes. In total, 38,217 gene models from the two newly sequenced genomes in this study were collected as queries to search against genome databases (with genomes sequenced and published for algae and embryophytes), as well as the NCBI Refseq database (https://www.ncbi.nlm.nih.gov/refseq/) for bacteria, viruses, fungi, protists, archaea, and the 1KP transcriptome database (https://sites.google.com/a/ualberta.ca/onekp/). Based on these, we defined innovative genes of Zygnematophyceae by the following criteria:

1, we excluded genes from Zygnematophyceae that have any homologous signal with "pre-Zygnematophyceae" genomes of Viridiplantae (based on the Viridiplantae phylogeny), i.e., Chlorophyta, *Mesostigma viride* and *Chlorokybus atmophyticus* (Wang et al., unpublished data), *Chara braunii*, *Klebsormidium nitens*, Coleochaetophyceae (unpublished genome sequence from *C. scutata*) using both Blastp (e-value < 1e-5) and HMMER (e-value < 1e-5) for confirmation.
2, genes retained from step1, were compared with genomes from the published embryophyte genomes. We excluded genes that have no homologous signal in any of the embryophyte genomes and retained those for which at least one homologous gene was found in the embryophyte genomes. This is to retain genes that are commonly shared between Zygnematophyceae and embryophytes but excluding Zygnematophyceae lineage-specific genes. Both Blastp (e-value < 1e-10) and HMMER (e-value < 1e-10) with a strict criterion were used for analysis process.
3, for genes retained from step2, we expanded the dataset collected from the 1KP database (1,000 Plants Transcriptome Project), and further confirmed the absence of genes from the "pre-Zygnematophyceae" genomes, and the presence of at least one gene shared with embryophytes.

This gene list was finalized as putative gains in the common ancestor of Zygnematophyceae and embryophytes for downstream analysis.

Then, we blasted (e-value < 1e-5) this set of genes against the genome databases of bacteria, viruses, archaea, fungi and protists from the published genome reference and the NCBI Refseq database. We then defined a set of putative HGT candidate genes according the following criteria:

1, the genes must have a homologous signal with at least one prokaryotic sequence (Bacteria, Archaea, fungi), which was further confirmed by HMMER search and phylogenetic analysis. Based on this, all the homologous sequences identified from bacteria, and eukaryotes (i.e., Zygnematophyceae and embryophytes) were collected for detailed gene family phylogenetic analyses. Multiple sequence alignment was performed for each by MAFFT, the alignments were further processed by G-Block (removing sites if > 50% of sequences/species showed a gap "N" along the aligned corresponding orthologous site).

A maximum likelihood (ML-based) RAxML approach with the evolutionary model LG4X was applied to reconstruct each of the gene family trees.

Specifically, a series of phylogenetic trees were built based on multiple sequence alignment of all GRAS protein domains (including the homologous bacteria-derived SAM-dependent methyltransferase as outgroup) by RAxML (1,000 bootstrap replicates), IQ-tree (1,000 generations of Ultrafast bootstrap) and FastTree (computes local support values with the Shimodaira-Hasegawa test), respectively, implementing the Maximum Likelihood method (ML-based) with different evolutionary models (RAxML: LG4X, and GTR+CAT models; IQtree: LG4X model; FastTree: JTT + CAT model).

Furthermore, to identify the evolutionary model that fits the observed data best, we used Phylobayes-MPI (models: CAT+GTR, and CAT+GTR+Dayhoff recoding) to perform posterior predictive simulations through converged runs to evaluate model fit. Basically, the readpb_mpi implemented in the PhyloBayes-MPI package is used to perform tests, both for the across-site and across-branch tests to compare the (CAT + GTR + Dayhoff recoding), (CAT + GTR) and LG models. Two chains were run in parallel and bpcomp and tracecomp implemented in Phylobayes-MPI were used to evaluate convergence, which was defined by the summary statistics that all dropped to < 0.3 for the two chains as well as for the maximum discrepancies in bipartition frequencies (bpcomp), while the effective sample size of each parameter > 100 as defined in the PhyloBayes manual. A particular test failed (the model doesn't fit the observed data) if the test statistic calculated from the real data doesn't fall in the central 95% of the simulated distribution. From our posterior predictive simulations, the (CAT + GTR + Dayhoff recoding) model fitted best for the data of GRAS gene family, while for PYL gene family, the (CAT + GTR) model fitted the data best.

2, the gene must be located in a "correct" continuous assembled contig/scaffold, for which the contiguity was validated by pair-end and mate-pair mapping reads (Data S2A and S2B). Clean reads from the pair-end and mate-pair libraries were mapped against the two Zygnematophyceae genomes by BWA (Li, 2013) (version 0.7.12-r1039). The alignment of mate-pair DNA library/reads to the HGT regions along scaffold (target HGT gene and its up-/down-stream 10kb) is to support the co-assembly in regions that contain the HGT candidate genes and the flanking genes. The alignment of pair-end DNA library/reads against the HGT regions supports a HGT event if the distribution of the read depth of the HGT candidate gene is similar to the target HGT and surrounding genes (up-/down-stream genes). IGV (Robinson et al., 2011) (version 2.4.18) was used to visualize pair-end mapping regions around the target HGT genes (GRAS and PYL). ClustersPloter (https://github.com/orangeSi/ClustersPloter) was used to visualize the mate-pair mapping regions around the target HGT genes (GRAS and PYL).

3, the gene must be expressed evaluated by RNA-seq mapping, flanked by other eukaryotic genes (Data S2C). Aligner Hisat2 (Kim et al., 2015) was used to map the clean RNA-seq reads from the Poly-A selection library against the genome, IGV was used to visualize the RNA-seq read mapping regions around the HGT candidate genes to show gene expression and gene structure.

PCR Validation on GRAS Gene Presence or Absence

We designed and synthesized 4 pairs of primers based on the transcriptome sequence of the "putative GRAS gene" (comp30290_c0_seq1) reported in the accession *Chaetosphaeridium globosum*, figshare. https://figshare.com/articles/Green_algal_transcriptomes_for_phylogenetics_and_comparative_genomics/1604778 (Cooper and Delwiche, 2016). Green algal transcriptomes for phylogenetics and comparative genomics). The primer sequences were summarized in Table S1Z. We blasted these primers against the genome of *Spirogloea muscicola*, and calculated the PCR primer length using the software geneious both for the target sequences in *Spirogloea muscicola*, as well as for those in *Chaetophaeridium globosum* (oligo synthesis). The PCR reaction components and the running step and parameters for the PCR program were summarized in Supplementary Table S1Z. The electrophoresis result was shown in Data S1BD.

Insertion Time Estimation of Bacteria GRAS Genes

To estimate the potential divergence time of the GRAS gene family, we implemented BEAST (Drummond et al., 2012) to infer the ancient HGT insertion interval from bacteria into Zygnematophyceae. GRAS genes from Bryophyta, two Zygnematophyceae and bacteria (the homologous bacteria-derived SAM-dependent methyltransferases were excluded) were used to reconstructed the phylogeny tree by BEAST for time inference. An uncorrelated relaxed-clock model with rates drawn from a lognormal distribution was used to run MCMC (10,000,000 replicates, Log parameter is 1,000). Fossil time on the origin of ancient embryophytes was used as speciation prior for time calibration, 473∼514 million years ago from the literature.

QUANTIFICATION AND STATISTICAL ANALYSES

All details of the statistics applied (e.g., for the kmer-based analysis to estimate genome complexity) are provided alongside the respective analysis in the Method Details section.

DATA AND CODE AVAILABILITY

Genome sequences, whole-genome assemblies, and genome annotations of *S. muscicola* and *M. endlicherianum* have been submitted to the National Center for Biotechnology Information (NCBI) database under BioProject PRJNA543679, PRJNA543678; the raw data of DNA and RNA sequencing reads have been submitted to NCBI under BioProject PRJNA541068, PRJNA541331. Those data are also available in the CNGB Nucleotide Sequence Archive (CNSA: http://db.cngb.org/cnsa; accession number CNP0000746). The genome and annotation files are also available in figshare (https://figshare.com/articles/Genomes_of_subaerial_Zygnematophyceae_provide_insights_into_land_plant_evolution/9911876/1). Data S1 and S2 have been deposited as Mendeley Datasets (http://dx.doi.org/10.17632/pvf47s35xy.1) and figshare Datasets (https://doi.org/10.6084/m9.figshare.10251038.v1). The nuclear-encoded SSU rDNA sequence of *S. muscicola* has been deposited in GenBank under the accession MN585752. All other data and materials are available from the corresponding authors upon reasonable request.

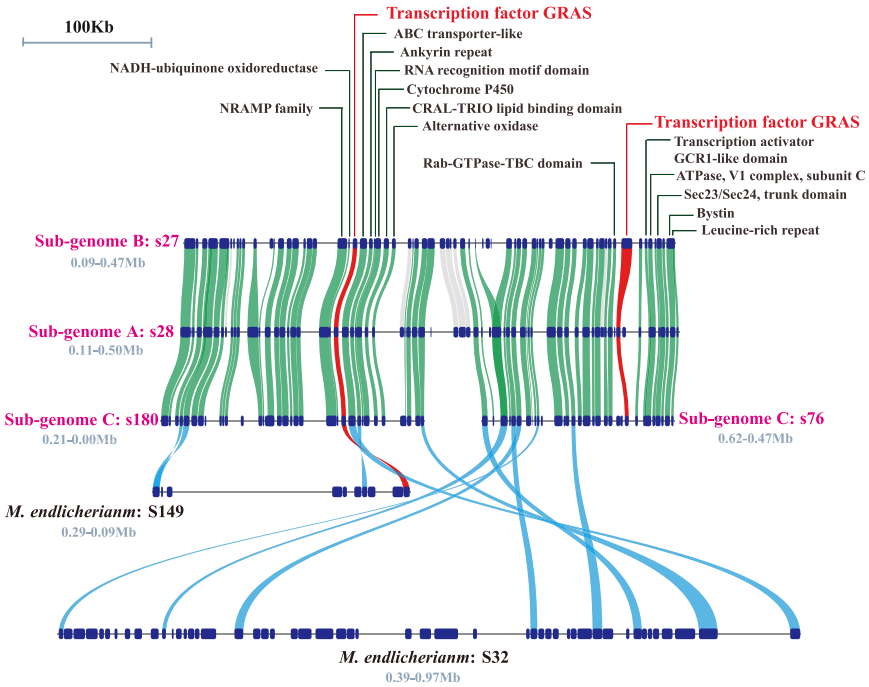


Figure S1. An Example of the Triplicated Syntenic Blocks, where GRAS Genes Are Located, Related to Figure 2

GRAS genes and their links across syntenic blocks are marked in red. Links connecting triplicated genes within *Spirogloea muscicola* scaffolds are marked purple, links connecting orthologous genes between *Mesotaenium endlicherianum* and *Spirogloea muscicola* are marked in cyan. Functional annotations for the up-/down-stream flanking genes were given.

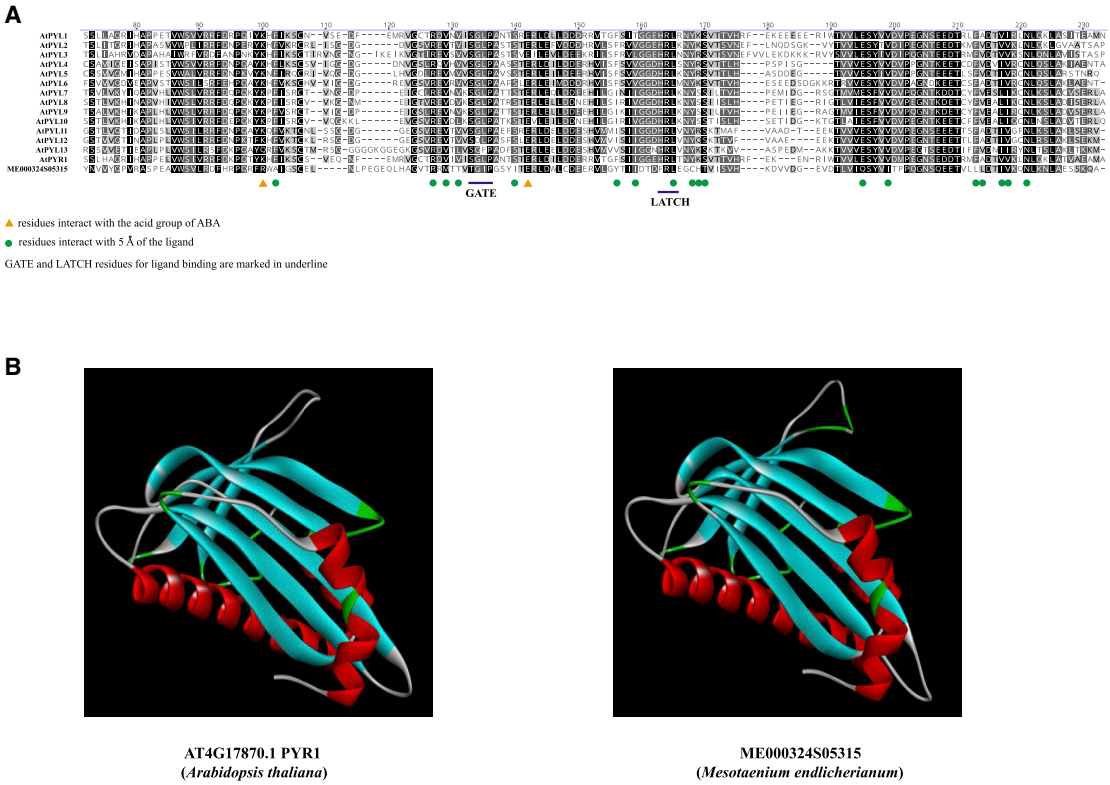


Figure S2. The Alignment and Predicted 3D Protein Structure of PYR/PYL/RCAR Protein Sequences, Related to Figure 3

A, Multiple sequence alignment was performed by MAFTT, and visualized by Geneious. Green arrow indicates that the residues interact with the acid group of ABA, Orange arrow indicates that the residues are amino acids 5 A of the ligand and the underline indicates that the residues are GATE and LATCH for ligand binding. Good sequence conservation of the GATE and LATCH regions of the PYR/PYL/RCAR genes was observed between *Mesotaenium endlicherianum* and *Arabidopsis thaliana*. (B), Illustration of the conservation of the 3D protein structure for the PYL genes, between *Mesotaenium endlicherianum* (ME000324S05315) and *Arabidopsis thaliana* (AT4G17870).

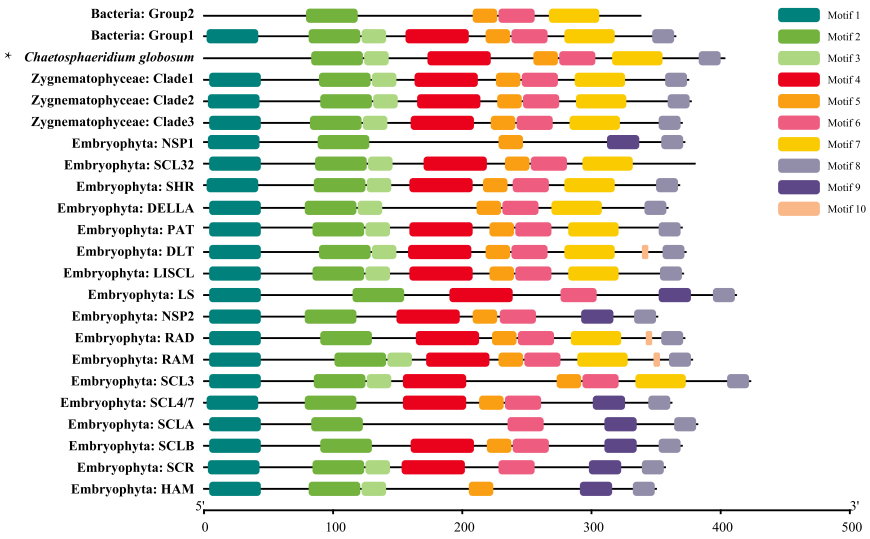


Figure S3. Visualization of GRAS-Domain Motifs, Related to Figure 4

A consensus sequence was made by Geneious (https://www.geneious.com) for each GRAS subfamily of bacteria, *Chaetosphaeridium globosum*, Zygnematophyceae and embryophytes, respectively. The accession number for the *Chaetosphaeridium* homolog is 'comp30290_c0_seq1' (from Cooper and Delwiche, 2016). Green algal transcriptomes for phylogenetics and comparative genomics. figshare. https://figshare.com/articles/Green_algal_transcriptomes_for_phylogenetics_and_comparative_genomics/1604778). Further evidence is needed to verify whether the *Chaetosphaeridium* GRAS sequence is genuine or just a contamination from bacteria. Each module represents a GRAS motif. The line at the bottom indicates amino acid sequence length (aa). For each motif, the conserved alignment was obtained from the multiple sequence alignment and was predicted by MEME suit 5.0.3 (http://meme-suite.org/doc/download.html).

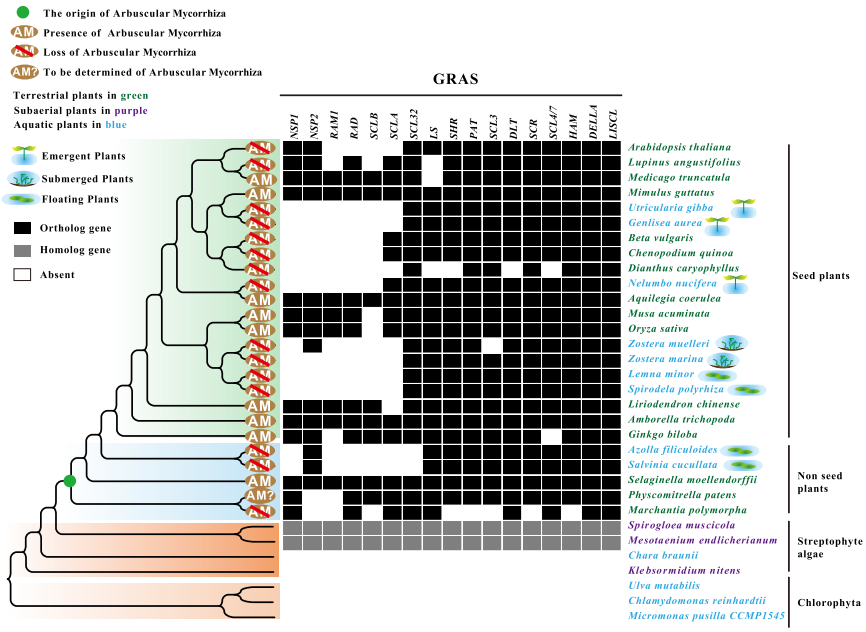


Figure S4. Phylogenetic Distribution of GRAS Gene Subfamilies, Related to Figure 4

Each of the representative GRAS subfamilies were studied. Phylogeny analysis, blastp and HMMER search were used to confirm the presence or absence of genes. The published genomes of aquatic plants were added for the purpose of this analysis. The illustrations of the origin of the arbuscular mycorrhizal symbiosis (AM fungi), loss of AM-symbiosis, and the status of aquatic habitats were derived from the literature. Absence or contraction of several GRAS subfamilies were consistently correlated with the loss of AM symbiosis, which is highly related with the transition of "going back to water" of these aquatic plants.