The *Chara* genome: secondary complexity and implications for plant terrestrialization

Tomoaki Nishiyama[1,*], Hidetoshi Sakayama[2,*], Jan de Vries[4,5], Henrik Buschmann[3], Denis Saint-Marcoux[6,7], Kristian K. Ullrich[8,40], Fabian B. Haas[8], Lisa Vanderstraeten[9], Dirk Becker[10], Daniel Lang[38], Stanislav Vosolsobě[17], Stephane Rombauts[11], Per K.I. Wilhelmsson[8], Philipp Janitza[12], Ramona Kern[13], Alexander Heyl[14], Florian Rümpler[15], Luz Irina A. Calderón Villalobos[30], John M. Clay[16], Roman Skokan[17], Atsushi Toyoda[18], Yutaka Suzuki[19], Hiroshi Kagoshima[20], Elio Schijlen[39], Navindra Tajeshwar[14], Bruno Catarino[6], Alexander J Hetherington[6], Assia Saltykova[11,21,22], Clemence Bonnot[6,36], Holger Breuninger[6,23], Aikaterini Symeonidi[8], Guru V. Radhakrishnan[24], Filip Van Nieuwerburgh[37], Dieter Deforce[37], Caren Chang[16], Kenneth G. Karol[25], Rainer Hedrich[10], Peter Ulvskov[26], Gernot Glöckner[27], Charles F. Delwiche[16], Jan Petrášek[17], Yves Van de Peer[11,28], Jiri Friml[29], Mary Beilby[31], Liam Dolan[6], Yuji Kohara[20], Sumio Sugano[19], Asao Fujiyama[18], Pierre-Marc Delaux[32], Marcel Quint[12,30], Günter Theißen[15], Martin Hagemann[13], Jesper Harholt[33], Christophe Dunand[32], Sabine Zachgo[3], Jane Langdale[6], Florian Maumus[34], Dominique Van Der Straeten[9], Sven B. Gould[4], Stefan A. Rensing[8,35,*,+]

[1] Advanced Science Research Center, Kanazawa University, Kanazawa 920-0934, Japan

[2] Graduate School of Science, Kobe University, Kobe 657-8501, Japan

[3] Botany Department, School of Biology and Chemistry, Osnabrück University, 49076 Osnabrück, Germany.

[4] Institute for Molecular Evolution, Heinrich Heine University, 40225 Düsseldorf, Germany

[5] Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada

[6] Department of Plant Sciences, University of Oxford, Oxford, OX1 3RB, United Kingdom

[7] Université de Lyon, UJM-Saint-Étienne, CNRS, BVpam FRE3727, 42023 Saint-Étienne, France

[8] Plant Cell Biology, Faculty of Biology, University of Marburg, 35043 Marburg, Germany

[9] Laboratory of Functional Plant Biology, Department of Biology, Gent University, 9000 Gent, Belgium

[10] Molecular Plant Physiology & Biophysics, University of Wuerzburg, 97082 Wuerzburg, Germany

[11] Department of Plant Biotechnology and Bioinformatics, Gent University and VIB Center for Plant Systems Biology, 9052 Gent, Belgium

[12] Institute of Agricultural and Nutritional Sciences, Martin-Luther-University Halle-Wittenberg, 06120 Halle (Saale), Germany

[13] Plant Physiology, University Rostock, 18051 Rostock, Germany

[14] Department of Biology, Adelphi University, Garden City, NY 11530, USA

[15] Department of Genetics, Friedrich Schiller University Jena, 07743 Jena, Germany

[16] CBMG, University of Maryland, College Park, MD 20742, USA

[17] Department of Experimental Plant Biology, Faculty of Science, Charles University, 128 44 Prague 2, Czech Republic

[18] Comparative Genomics Laboratory and Advanced Genomics Center, National Institute of Genetics, Shizuoka 411-8540, Japan

[19] Department of Computational Biology and Medical Sciences, University of Tokyo, Kashiwa, Chiba 277-8562, Japan

[20] Genome Biology Laboratory, National Institute of Genetics, Shizuoka 411-8540, Japan

[21] Platform Biotechnology and Molecular Biology, Scientific Institute of Public Health (WIV-ISP), Brussels, Belgium

[22] Department of Information Technology, Gent University, IMinds, 9052 Gent, Belgium

[23] ZMBP, Entwicklungsgenetik, 72076 Tübingen, Germany

[24] Department of Cell and Developmental Biology, John Innes Centre, Norwich NR4 7UH, United Kingdom

[25] Lewis B. and Dorothy Cullman Program for Molecular Systematics, The New York Botanical Garden, Bronx, NY 10458, USA

[26] Department of Plant and Environmental Sciences, University of Copenhagen, DK-1871 Frederiksberg C, Denmark

[27] Biochemistry I, Medical Faculty, University of Cologne, 50931 Cologne, Germany

[28] Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, 0028, South Africa

[29] Institute of Science and Technology, 3400 Klosterneuburg, Austria

[30] Department of Molecular Signal Processing, Leibniz Institute of Plant Biochemistry, 06120 Halle (Saale), Germany

[31] School of Physics, University of NSW, Sydney, Kensington, 2052, NSW, Australia

[32] Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, CNRS, UPS, Auzeville, BP42617, 31326 Castanet Tolosan, France

[33] Carlsberg Research Laboratory, 1799 Copenhagen V, Denmark

[34] URGI, INRA, Université Paris-Saclay, 78026 Versailles, France

[35] BIOSS Centre for Biological Signalling Studies, Unigersity Freiburg, Germany

[36] Present address: Labex ARBRE, UMR 1136 INRA-Université de Lorraine (IAM), INRA-Grand Est-Nancy, Champenoux, France

[37] Laboratory of Pharmaceutical Biotechnology, Gent University, 9000, Gent, Belgium

[38] PGSB, Helmholtz Center Munich, 85764 Neuherberg, Germany

[39] Wageningen University, B.U. Bioscience, 6700 AA Wageningen, The Netherlands

[40] Present address: Max Planck Institute for Evolutionary Biology, 24306, Ploen, Germany.

+ Lead contact: Stefan A. Rensing

* Authors for correspondence:

tomoakin@staff.kanazawa-u.ac.jp

hsak@port.kobe-u.ac.jp

stefan.rensing@biologie.uni-marburg.de

Summary

Land plants evolved from charophytic algae, among which Charophyceae possess the most complex body plans. We present the genome of *Chara braunii*; comparison of the genome to those of land plants identified evolutionary novelties for plant terrestrialization and land plant heritage genes. *C. braunii* employs unique xylan synthases for cell wall biosynthesis, a phragmoplast (cell separation) mechanism similar to that of land plants, and many phytohormones. *C. braunii* plastids are controlled *via* land plant-like retrograde signaling, and transcriptional regulation is more elaborate than in other algae. The morphological complexity of this organism may result from expanded gene families, with three cases of particular note: genes effecting tolerance to reactive oxygen species (ROS), LysM receptor-like kinases, and transcription factors (TFs). Transcriptomic analysis of sexual reproductive structures reveals intricate control by TFs, activity of the ROS gene network, and the ancestral use of plant-like storage and stress protection proteins in the zygote.

Keywords: plant evolution, charophyte, phytohormones, transcriptional regulation, Phragmoplastophyta, Chara, streptophyte, reactive oxygen species, phragmoplast

Introduction

A pivotal event in the emergence of plant life was the mid-Paleozoic adaptation to land. While several algal lineages evolved to occupy terrestrial environments, only one represents the land plant ancestor; its terrestrialization event was fostered by a range of evolutionary novelties. The specific complement of traits that allowed a particular algal lineage to give rise to land plants and dominate the terrestrial environment remains under active study. Similarity of critical plant developmental, sensory, and regulatory pathways to homologous pathways in charophyte green algae has been demonstrated in several recent studies, emphasizing the close relationship among these lineages (reviewed in (Rensing, 2018)).

Charophytic algae are the closest living relatives of land plants (embryophytes), with both groups collectively referred to as streptophytes (Fig. 1). The Charophyceae, Coleochaetophyceae, and Zygnematophyceae together with the land plants represent the clade Phragmoplastophyta (Lecointre and Le Guyader, 2006), united by the presence of the phragmoplast (Pickett-Heaps, 1975), an array of microtubules perpendicular to the cell division plane that functions in the formation of the nascent cell wall. The Klebsormidiophyceae, Chlorokybophyceae, and Mesostigmatophyceae share fewer traits with land plants (Fig. 1). While Charophyceae were hypothesized to be most closely related to land plants on the basis of similar body plans (Pringsheim, 1862), recent studies indicated that the Zygnematophyceae are the land plant sister group (Wickett et al., 2014).

Extant Zygnematophyceae have simple body plans that seem to reflect secondary loss of morphological complexity. In contrast, the earlier diverging Charophyceae are morphologically more complex than all other charophytic algae: the haploid thallus body plan encompasses a shoot-like axis consisting of nodes with whorls, internodes, a simplex apical meristem, plus multicellular rhizoids (Fig. 2). Cells of the internode are large and complex, featuring endo- and ectoplasma, multiple plastids and nuclei, and communicate *via* electrical signals. The morphology of extant charophytic groups thus infers mosaic evolution and suggests that the genomes of Charophyceae, not Zygnematophyceae, will likely reveal the suite of traits that facilitated terrestrialization (Delwiche, 2016).

Here we present the genome sequence of the charophycean alga *Chara braunii*, one of the most morphologically complex extant Charophyta, shedding light on early embryophyte diversification and the colonization of land by plants.

Results and Discussion

The *Chara braunii* genome: assembly, annotation and comparison

C. braunii features a haplontic life cycle (Fig. 2), the draft sequence reported here represents a haploid genome. 1.75 Gbp of nuclear scaffold data were obtained, of which 1.43 Gbp were assembled into contigs, corresponding to ~74% of the *C. braunii* genome. RNA-seq of vegetative and reproductive stages was used together with full-length cDNA sequences to annotate the genome. 23,546 putative protein coding gene models were identified, of which 53% are supported by RNA-seq data (Table S4). At least 94% of several conserved core gene sets are encoded by the genome, indicating its suitability for genomic and comparative analyses (STAR Methods).

The observed chromosome number n=14 (Fig. S1) corresponds to the base chromosome number of *Chara* species. Indeed, synonymous substitution distance (Ks)-based analysis of *C. braunii* paralogs revealed no evidence of whole genome duplication (WGD) events (Fig. S3) and thus paralog acquisition and retention is probably due to small-scale duplications. Repetitive elements (Table S1F, S1G) collectively contribute approximately 1.1 Gbp (61%, or 75% if gaps are excluded). Unlike in most plants and green algae, there are no Copia-type long terminal repeat (LTR) retrotransposons (RT) detectable. We discovered a family of repeats with putative GIY-YIG homing endonuclease and reverse transcriptase domains, which are hallmarks of Penelope RTs and group II introns that are uncommon in plant genomes.

The density of LTR elements in the genome is intermediate between compact genomes like those of *Arabidopsis thaliana* or *Klebsormidium nitens*, and other large genomes such as maize and barley (Fig. 3). *C. braunii* introns are an order of magnitude longer than in any of the other genomes investigated here (Table S1L), although intron boundaries appear to be conserved. The high intron length coincides with a low number of introns per gene (3.82), similar to the value for the barley genome (3.89, Table S1L); intron length and number show negative correlation ($r = -0.42$). Repetitive elements represent 39% of the intron space (Fig. 3, Mendeley archive) which is strikingly enriched with Penelope-like elements and depleted of other types of repeats including Class 2 transposable elements (Helitrons and DNA transposons), suggesting differential integration bias and/or retention in introns as compared to intergenic space (Table S1L).

Evolutionary novelties enabling terrestrialization and land plant heritage genes (LPHG)

The lineage harboring *C. braunii* diverged from land plants 550–750 Ma (Morris et al., 2018). By identifying features that are shared between the *C. braunii* genome and extant land plants, putative ancestral traits can be identified that have been retained over several hundred Ma. Here we refer to the genes underlying these traits as land plant heritage genes (LPHG) and similarly deduce evolutionary novelties.

Cell division and cell wall

C. braunii, like land plants, performs cytokinesis by assembling a cell plate using a phragmoplast microtubule array while *K. nitens* divides by an evolutionarily older cleavage (Fig. 1). Phragmoplast-mediated cell division is assumed to have enabled filament branching through a shift in the plane of cell division (Buschmann and Zachgo, 2016). Land plants also evolved another microtubule array, the preprophase band (PPB), which functions in phragmoplast and cell plate guidance. Focusing on genes for phragmoplast and PPB function, a list of 221 *A. thaliana* cytokinesis genes was compiled (Table S1C). Sequence comparisons showed that the genomes of *A. thaliana*, *P. patens*, *C. braunii* and *K. nitens* have a highly similar complement of cytokinesis-related genes while the unicellular chlorophyte *Chlamydomonas reinhardtii* is divergent. Interestingly, the *C. braunii* genome lacks the *TANGLED1* gene. In land plants, microtubule-associated TANGLED1 localizes to PPBs and is required for phragmoplast guidance (Walker et al., 2007). Since TANGLED1 homologs are found in several bryophytes, but none in any algae, this gene likely played an important role in PPB evolution (Fig. 1). To gain further insight into the evolution of the phragmoplast, we determined how many paralogs each of the cytokinesis genes has in *C. braunii* as compared to *K. nitens*. In this way we identified possible phragmoplast signature genes (Table S1C). Among others, we detected expansion of cyclins as well as EXOCYST and SNARE complex members (Table S1C, Data S1Q-S). The expansion of phragmoplast-related gene families in *C. braunii* / the Phragmoplastophyta, but not in Chlorophyta, *K. nitens* or *M. viride*, suggests their sub- and neofunctionalization to enable phragmoplast function.

Like land plant cell walls, those of *C. braunii* consist of cellulose embedded within a pectin and hemicellulose matrix (Sorensen et al., 2011), its synthesis is orchestrated by a repertoire of glycosyltransferases much like in land plants (Table S1H), with the exception of an apparently unique mechanism for xylan synthesis. The GT47 xylan synthase XYS1 has been identified in *K. nitens*, as well as IRX9 and IRX14 from GT43 (Data S1A), implicated in xylan biosynthesis despite no apparent requirement for being an active enzyme (Ren et al., 2014). Orthologs to neither XYS1 nor IRX9/14 could be identified in *C. braunii*, however, a deep branching, highly diverged form of GT43 was identified as the most likely *C. braunii* xylan synthase, providing the first hint that GT43 sequences are enzymatically involved in synthesizing xylan.

Phytohormones

Phytohormones enable the integration of environmental stimuli with developmental programs. As such, they are a key feature of land plants, with some apparently having origins in algae (Hori et al., 2014; Ju et al., 2015; Wang et al., 2015). Potential orthologs of phytohormone pathway genes were defined across *K. nitens*, *C. braunii*, *P. patens* and *A. thaliana* (Table 1, Fig. 4, Table S1J).

Auxin (AUX)

AUX is one of the major regulators of plant growth and development. Biosynthesis of AUX (Hori et al., 2014) as well as transcriptional and physiological response to high concentrations have been shown in *K. nitens* (Ohtaka et al., 2017). In contrast to *K. nitens*, genes enconding biosynthetic enzymes of the TAA and YUCCA families are absent from *C. braunii* (Table 1). In *C. australis* IAA, serotonin and melatonin accumulate in a synchronized manner during the

day/night cycle (Beilby et al., 2015). As the tryptamine IAA biosynthetic pathway intersects with the serotonin/melatonin pathway (Tivendale et al., 2014), *Chara* may synthesize and metabolize AUX *via* a different route than land plants.

Homologous genes for both *PIN*s and *ABCBs* are present in the *C. braunii* genome (Table 1, Table S1K), supporting previous data on polar AUX transport (PAT) in *K. nitens* (Hori et al., 2014) and Charales (Boot et al., 2012). Homologous sequences for AUX1/LAX-like influx carriers as well as the intracellular PIN-like (PILS) transporters, however, are absent from the *C. braunii* genome (Table 1), suggesting that AUX transport and homeostasis display an evolutionary history different from other streptophytes.

The land plant-type AUX signaling cascade, consisting of $SCF^{TIR1/AFB}$ and Aux/IAA co-receptors and ARF TFs, was suggested to be absent in *K. nitens* (Hori et al., 2014; Ohtaka et al., 2017). *K. nitens* encodes an Aux/IAA domain containing protein (Wang et al., 2015) that features an additional B3 domain, is not induced by IAA (Ohtaka et al., 2017) and thus not classified as canonical Aux/IAA (Table 1). In addition to all components of the ubiquitin-proteasome system (Table S1I), *C. braunii* features a single *ARF* (Data S1E) with land plant-like domain composition (Flores-Sandoval et al., 2018), and two Aux/IAA-like sequences (Table 1, Data S1F) clustering with the *A. thaliana* non-canonical IAA33 (lacking a TOPLESS interacting motif and degron for AUX-dependent $SCF^{TIR1/AFB}$-Aux/IAA interactions.

Commented [R1]: Marcel Quint: suggest to remove

Commented [R2R1]: Not so sure, because readers might want to compare with Wang et al.

C. braunii also encodes several F-box proteins (FBPs) with sequence similarity to land plant phytohormone co-receptors (Data S1P). None of the TIR1/AFB-like FBPs cluster with the land plant AUX co-receptor gene family (Data S1G). Structural modeling, however, reveals that the *C. braunii* sequences adopt a solenoid-fold architecture resembling TIR1 (Tan et al., 2007). Ligand binding modeling supported the potential ability to form an AUX binding pocket (Data S1P). The existence of only degron-less *C. braunii* Aux/IAAs, however, prompts to postulate that a land plant-like TIR1/AFB-Aux/IAA co-receptor pair is most likely not functional in *C. braunii*.

Consequently, while obvious candidates for canonical land plant AUX biosynthesis genes are lacking, there is a partial candidate gene set of the major land plant AUX signaling and PAT pathways in *C. braunii*. In conclusion, AUX biosynthesis, transport, and some form of signaling were already present in the last common ancestor of *C. braunii* and *K. nitens*, but AUX signaling *via* ARFs was apparently gained in the common ancestor of Phragmoplastophyta, as was ARF repression by Aux/IAAs (Table S1Q, Fig. 4).

Cytokinin (CK)

The CK signaling pathway consists of four protein families: the receptor, the histidine-containing phosphotransfer protein, and the type A and B response regulators (RRA and RRB) (Heyl et al., 2013). The *C. braunii* genome encodes members of the first three, but no RRBs (Table 1, Fig. 4). This is contrast to their presence in all chlorophytes and charophytes analysed (Hori et al., 2014; Wang et al., 2015). Several RR domains closely related to RRBs were found, but none contained the Myb domain essential for RRB function (Table S1J). Given the complexity of the *C. braunii* genome, it is possible that not all genes were correctly or completely predicted, but neither genome nor transcriptome data (Data S1H) provide evidence

for RRB genes. Their loss suggests either the rewiring of CK signaling or substitution of RRB function by other genes.

Ethylene (ETH)

The *C. braunii* genome possesses one or more potential homologs of all of the core components associated with ETH signaling (Table 1, Fig. 4 and Table S1J). *Chara* exhibits ETH-binding activity (Wang et al., 2006), and *C. braunii* encodes several ETH receptor homologs. Notably, *C. braunii* possesses a full-length homolog of *EIN2*, a central regulator in ETH signaling. This is in contrast to both the *K. nitens* genome, which lacks *EIN2* (Hori et al., 2014), and the *Spirogyra pratensis* transcriptome, which shows only a partial *EIN2* sequence (Ju et al., 2015). Except for *EIN2*, *S. pratensis* possesses an ETH signaling pathway that is functionally conserved with the pathway known in land plants (Ju et al., 2015). These findings indicate that the land plant-like ETH signaling pathway was established in the common ancestor of the Phragmoplastophyta, after its divergence from the lineage leading to *Klebsormidium*.

Abscisic acid (ABA)

Orthologs of the core ABA signaling components are present in bryophytes and it has been suggested that all ABA biosynthesis/signaling components were gained in the common ancestor of Charophyta (Ju et al., 2015; Wang et al., 2015), with the exception of PYR/PYL receptors that were probably gained in the common ancestor of Zygnematophyceae and land plants (de Vries et al., 2018). The *C. braunii* genome does not contain homologs of the co-receptors ABI/HAB, nor the PYR/RCAR family of receptors (Park et al., 2009), but possesses homologs of genes encoding enzymes that act early in the ABA synthesis pathway (from carotenoid synthesis to violaxanthin; Table 1, Fig. 4 and Table S1J). Given that the presence of ABA has been confirmed in *C. braunii* (Hackenberg and Pandey, 2014), it is likely that the biosynthetic pathway differs from that found in land plants, with ABA possibly being synthesized directly from farnesyl pyrophosphate.

Strigolactones (SL)

Orthologs of all the core SL signaling components have been identified exclusively in the genomes of seed plants; however, D14-like receptor homologs are found encoded by bryophytes and charophytes (Bythell-Douglas et al., 2017; Wang et al., 2015). Two SL-related homologs were identified in *C. braunii*, one encoding beta-carotene isomerase D27, and one encoding the candidate SL/karrikin receptor D14-like (Table 1, Fig. 4 and Table S1J). Given the presence of SL in several Charales species, and the activity of the synthetic SL GR24 on *Chara corallina* rhizoid growth (Delaux et al., 2012), it is likely that SL synthesis and signaling differ in charophyceans and in seed plants (Bythell-Douglas et al., 2017). It has been suggested that D14-like proteins might act as (the) SL receptor(s) in this group.

In summary, although the phytohormones AUX and CK seem to be ancestral features of streptophytes, and SL and ABA of Phragmoplastophyta (Fig. 1), the respective biosynthesis and/or signaling pathways differ between seed plants and *C. braunii.* Some features of these

four phytohormone networks, and of ETH signaling, first appeared in the Phragmoplastophyta as evident by their presence in *C. braunii*. Others were either not present in the ancestor or have since diverged.

Plastid evolution: photorespiration and retrograde signaling

Photorespiration, which recycles the 2-carbon compound formed when ribulose bisphosphate carboxylase/oxygenase reacts with oxygen instead of CO_2, is crucial to photosynthesis in an oxygen-rich atmosphere. The *C. braunii* genome encodes proteins necessary to carry out a plant-like photorespiratory cycle, including a plant-type glycolate oxidase (GOX) (Table S1M) with structural features preferring glycolate over lactate (Hackenberg et al., 2011). Plant-type GOX is also present in *K. nitens*, while *C. reinhardtii* uses a mitochondrial glycolate dehydrogenase for photorespiratory glycolate metabolism (Nakamura et al., 2005). Apparently, plant-like photorespiration was present in the common ancestor of Streptophyta, the pathway being a feature that might have aided terrestrialization.

The plastid to nucleus signaling network optimizes plastid function in land plants. All Chloroplastida (Fig. 1) share EXECUTOR-transduced singlet oxygen and a rudimentary tetrapyrrole-derived retrograde signaling, to which streptophytes recruited GUN2/3 (Fig. 5A). Our data show that *C. braunii*, but not *K. nitens*, encodes GUN1, at which multiple retrograde signals converge in land plants (reviewed by (Chan et al., 2016)). The only GUN1 candidate protein in *K. nitens* (kfl00096_0090) clustered with streptophyte algae- and bryophyte-specific PPRs, but not GUN1 (Data S1I). Hence, retrograde signaling featuring GUN1 might represent an evolutionary novelty of the Phragmoplastophyta (Fig. 1).

Plastid-encoded RNA-polymerase (PEP) is the ancestral and for most Archaeplastida the only PEP. In land plants, PEP activity is controlled through PEP-associated proteins (PAPs) (Pfalz and Pfannschmidt, 2013). We detected 5, 8, 10 and 11 PAP orthologs in *C. reinhardtii*, *K. nitens*, *C. braunii*, and *P. patens*, respectively. PAPs were thus already present in streptophyte algae (Fig. 5A) and underwent expansion in land plants. Most of the detected PAPs are predicted to be targeted to the chloroplast, the mitochondrion or both (Table S1N); dual-localization of PAPs to both organelles might be an ancient and conserved character state.

Transcriptional regulation

Within the Chloroplastida, morphological complexity correlates with the number of TF (acting in a sequence-specific manner, typically by binding to *cis*-regulatory elements) and transcriptional regulator (TR, acting on chromatin or *via* protein-protein interaction) genes (Lang et al., 2010). We identified 730 TF/TR genes in the *C. braunii* genome (Table S1Q), the complement of such proteins thus being larger than in *K. nitens* (627) or *C. reinhardtii* (542), coinciding with morphological complexity. *C. braunii* encodes several TFs that are not present in other algae, including *K. nitens*. Based on the available data, these families first appear in the Phragmoplastophyta, although they were previously thought to have been gained in the common ancestor of Coleochaetophyceae, Zygnematophyceae and land plants (Wilhelmsson et al., 2017). They include the single canonical ARF mentioned before, as well as TCP, HRT and

Zn cluster TFs (Fig. 1). The *C. braunii* genome encodes two TCP genes, which belong to TCP-P (class I) and TCP-C (class II). The two TCP subgroups are known to exert antagonistic functions in *A. thaliana* with regard to growth proliferation of organs and tissues (Nicolas and Cubas, 2016), implying that the appearance of two different TCP genes might have contributed to regulation of proliferation in the Phragmoplastophyta.

Two separate clades of MADS-box genes exist (Type I and II), with land plant Type II genes further subdivided into so called MIKC^{C} and MIKC*-type genes (Gramzow and Theißen, 2010). No Type I genes were identified in the *C. braunii* genome, but three Type II genes, of which only *CbMADS1* shows a canonical MIKC-type domain structure. Phylogeny reconstructions together with exon-intron structure analysis (Fig. 5B; Fig. S5, Data S1K) suggest that MIKC^{C} and MIKC*-type genes evolved from the duplication of an ancestral Type II gene followed by different exon duplications in both gene lineages. As such, *CbMADS1* may serve as a model for the ancestral MIKC-type gene that gave rise to MIKC^{C}- and MIKC*-type genes of land plants.

C. braunii encodes 11 bHLH TFs in 5 subfamilies. The Va(2) subfamily is present in chlorophytic and charophytic genomes and not present in land plants, suggesting that this subfamily was lost in the lineage leading to land plants (Table S1O, S1P). The *C. braunii* genome encodes 11 homeodomain (HD) TFs grouped into 9 subfamilies (Table S1O, S1P). Consistent with previous analyses (Catarino et al., 2016), *C. braunii* contains members of the KNOX, BEL, DDT and PINTOX subfamilies that are conserved in chlorophytes.

Zygotes and spores as analogs to seeds

Dormant haploid spores of mosses share features of regulation and coat biosynthesis with diploid seeds of flowering plants (Daku et al., 2016; Vesty et al., 2016). The diploid zygotes of *Chara* are dormant diaspores that presumably undergo meiosis and germinate upon suitable environmental cues (Delwiche and Cooper, 2015). Differential expression analysis shows that a number of transcripts related to seed storage proteins (cupin superfamily, oleosins) and to stress tolerance proteins found in seeds (e.g. late embryogenesis abundant), accumulate to high levels in zygotes (Fig. S4). These proteins probably enable the *C. braunii* zygotes to withstand harsh environmental conditions and represent a reservoir of nutrients to facilitate germination and growth. Homologs of these genes have apparently been adopted during land plant evolution to enable dormancy in other diaspores, namely spores and seeds.

Evolutionary novelties of the *Chara* lineage

Trihelix TFs

The number of TFs per family is lower in *C. braunii* than in land plants for most families, with the trihelix family being an exception: 302 members are encoded, while land plant genomes typically encode ca. 30 copies (Table S1Q). Trihelix TFs are involved in the regulation of development (e.g. embryogenesis, flower development), as well as responses to abiotic and biotic factors. Based on RNA-seq data, at least 28 of the *C. braunii* genes are expressed (Table S4, Fig. S4); 19 in vegetative tissue (of which 6 are expressed exclusively in vegetative tissue)

and 22 in reproductive tissues (antheridia, oogonia, zygotes; Fig. S4). Phylogenetic analysis shows that the vast majority of *C. braunii* trihelix paralogs groups outside of the four clades previously defined (Kaplan-Levy et al., 2012) (Data S1J). Similar to secondary expansion of TF families in other lineages the expansion of trihelix TFs in *C. braunii* might be connected to the independent evolution of morphological complexity.

Phytohormones: PINs

There are six PIN AUX transporter proteins potentially encoded by the *C. braunii* genome (Table S1J). In land plants, the evolution of morphological complexity in the gametophytic generation, and later in the sporophytic generation, coincides with independent radiations within the *PIN* gene family (Bennett, 2015). Given its high morphological complexity, the same might have occurred in *C. braunii*.

Commented [R3]: Marcel Quint: suggest to remove or integrate with auxin part

Commented [R4R3]: But it is mentioned here are one of the expansions, I would keep it as is. Is also mentioned in the cladogram Figure

Motor network

The evolution of land plants is accompanied by increased abundance of myosin and kinesin domain proteins. Because *K. nitens* has slightly more predicted kinesins than *C. braunii* (Table S1S), it appears that phragmoplast evolution did not depend on the neofunctionalization of kinesin paralogs. However, myosin motors use filamentous actin as tracks. The expansion of the actin family in *C. braunii* (*K. nitens* and *C. reinhardtii* encode 7 actin genes, whereas *C. braunii* has 16; Data S1T, U), with each gene encoding a slightly different protein, hints at varying functions among the cytoskeleton. Land plants have 9 actin genes (*Marchantia polymorpha*) to often 12 (*A. thaliana*, papaya, *Amborella trichocarpa*), and up to 34 in the polyploid maize, while transcriptomic data of other Charales suggests high numbers of underlying genes, e.g. 27 transcripts in *Nitella mirabilis*, 101 in *N. hyalina* (and 46 in the desmid *Penium margaritaceum*). The high numbers of actin genes detected in the amoebal protist *Naegleria gruberi* (86), and the slime mold *Dictyostelium discoideum* (39) (Joseph et al., 2008), can to a large part be explained by their involvement in cell movement. Thus, the additional actin genes of *Chara*, *Nitella* and *Penium* may serve the enhanced cytoplasmic streaming observed in these organisms.

Electrical excitability

Inspired by the work of (Hodgkin and Huxley, 1952) on the squid axon, the large internodal cells of *Chara* emerged as an excellent experimental system for electrophysiological studies on plant excitability - the “Green Axon” (Beilby, 2007). On a slower time scale (1000x), the internodal cells fire action potentials (APs) in response to such as depolarization, light, heat shock, injury or touch. The *C. braunii* genome encodes several putative Touch/Mechano-Sensitive (MS) channels: two members of the MscS-like (MSL) family, as well as an ortholog of the eukaryote specific Piezo-type channel. The negative resting potential (up to -250 mV) across the plasma membrane is generated by the P-type H^+-ATPases, encoded in the *C. braunii* genome (Table S1R). Ca^{2+} and Cl^- contribute to the depolarizing phase of the *Chara* AP, while K^+ efflux shapes the AP repolarization phase as in animals. No animal-like voltage-gated Na^+

or Ca^{2+} channels were identified, but a single ALMT-type anion channel gene is present in *C. braunii*. The anion channel in *Chara* is Ca^{2+}-activated and voltage sensitive, so an Anoctamin-like channel poses another possibility. A Shaker-type, voltage-gated K^+ channel in *C. braunii* genome probably mediates the depolarization-activated potassium efflux of the AP repolarization phase. The *C. braunii* habit of long internodal cells might require long distance electrical signaling (Beilby, 2015) enabled by its peculiar set of ion channels. The similarities or differences of *C. braunii* AP, as compared to flowering plants, are yet to be established.

Sensing of biotic interaction and microbiome

Land plants harbor a large number of LysM receptor-like kinases (RLK) involved in the perception of chitin-based signals produced by pathogenic and beneficial microorganisms. One member of this family has been described in charophytic algae suggesting either an inability to discriminate microorganisms or an alternative system to do so (Delaux et al., 2015). The *C. braunii* genome revealed the presence of seven LysM-RLKs (Fig. 5C; Data S1N) that expanded independently of land plant LysM-RLKs. This expansion may reflect an adaptation of *C. braunii* to an extended range of interacting microorganisms (co-cultured bacteria: Table S1T, S1U). This is noteworthy given that many have failed to axenically cultivate Charophyceae, raising the possibility that growth may be dependent on microbiotic commensalism or mutualism.

Sexual reproduction and the ROS network

To analyze reproductive mechanisms, transcriptomes of antheridia, oogonia and zygotes were generated (Fig. 5/6, Fig. S6, Table S2 & S3). For antheridia, the data demonstrate that cell motility is up-regulated as expected (Fig. 5D; Fig. S6A). Of 949 differentially expressed genes (DEGs) upregulated in antheridia, 49 encode proteins harboring dynein heavy chains. Dynein-mediated transport is employed in flagellate cells such as spermatozoids and was lost during land plant evolution, concomitant with the loss of motile cells (Rensing et al., 2008). 22 of 302 trihelix TFs are expressed in reproductive tissues. Of those, 9 are expressed in all three tissues, 5 specifically in antheridia, 7 in oogonia and antheridia, and 1 specifically in the zygote (Fig. S4B). This expression profile may suggest a possible role for these genes in sexual reproduction, in particular in antheridia. Transcripts of a HMG TR and a RWP-RK TF also specifically accumulated in antheridia. Members of these families were shown to be involved in mating in fungi (Barve et al., 2003) and gamete differentiation in *C. reinhardtii* (Lin and Goodenough, 2007), and the single RWP-RK TF in *M. polymorpha* keeps egg cells quiescent in the absence of fertilization (Rovekamp et al., 2016).

Zygote transcriptome profiles are characterized by transcription, microtubule-based movement and protein kinase activities (Fig. S6D), processes that might be hallmarks of the diploid zygote maturing and entering dormancy. 87 TFs/TRs are differentially expressed between zygotes and oogonia, among them families typically linked to the regulation of development (e.g. bHLH, HD, AP2/EREBP; Fig. S4C), supporting the hypothesis that transcription undergoes a switch after fertilization. One of the seven LysM RLKs (g44510) is strongly induced in zygotes. In line with potential commensalism mentioned above, this protein might detect the presence of

beneficial microbes as a putative factor triggering meiosis and germination of the dormant zygote.

Of particular interest is the up-regulation of oxidation reduction processes in oogonia as compared to antheridia or zygotes (Fig. 5E; Fig. S6B/C). Like all living organisms, *C. braunii* needs to deal with constitutive production of reactive oxygen species (ROS) using the ROS gene network (Fig. S7, Table S1X). In contrast to land plants, aquatic plants have the option to let ROS diffuse into the water. *C. braunii* encodes all families responsible for ROS scavenging, but with lower gene copy number in comparison to land plants. In contrast, CC-type glutaredoxins (GRX) (ROXYs in *A. thaliana*), which exert crucial functions during angiosperm reproductive development (Gutsche et al., 2015), could not be detected (Table S1X). Among redox-associated genes (Table S1X) the class III peroxidases (Prx), thioredoxins and respiratory burst oxidase homologs expanded greatly during land plant evolution. However, only Prx expanded in *C. braunii* compared to *K. nitens* (Data S1O). With both peroxidative and hydroxylic catalytic cycles, these enzymes can regulate ROS and polymerize cell wall compounds (Francoz et al., 2015). Most of the *C. braunii* Prx are predicted to be secreted, as such, they may contribute to the formation of the strikingly elaborate reproductive structures, e.g. the thick zygote wall (Fig. 2).

7 out of 12 Prx are 2-8 fold higher expressed in oogonia than in antheridia or zygotes (Fig. 6). The higher expression of the ROS gene network could be related to the ROS homeostasis regulation necessary for an optimum fecundation. Flowering plant stigmas exhibit high levels of peroxidase activity when receptive to pollen (McInnis et al., 2006) and have been discussed to be involved in pollen-pistil interaction or pollen tube growth/penetration (Beltramo et al., 2012). For *A. thaliana* root and shoot apical meristems it was shown that stem cell-specific Prx fine tune the balance between superoxide anions ($O_{2}^{\cdot-}$) and hydrogen peroxide (H_2O_2) and thereby affect the switch between cell maintenance and differentiation (Zeng et al., 2017). Differential regulation of ROS levels by Prx might control sexual reproduction in *C. braunii*. Potentially, this mechanism arose in the common ancestor of Phragmoplastophyta and has been recruited from the gametophyte to the sporophyte during land plant evolution.

Conclusions

The *C. braunii* genome encodes more proteins than other algae, but less than most land plants. Both, specific gains / expansions and losses, can be attributed to the *Chara* lineage (Fig. 1). In absence of a WGD gene family expansions resulted from gene duplication and differential loss. Many of these events likely represent secondary gains in *Chara* complexity *via* sub- and neofunctionalization. We hypothesize that many gene family expansions detected in the *C. braunii* genome underpin its strikingly complex morphology.

Comparative genome analysis clearly reflects the phylogenetic placement of *C. braunii* as a close relative of land plants, with both striking similarities and important differences. It demonstrates the substantial insights into fundamental aspects of plant biology that can be gained by comparing diverse relatives. Molecular signatures across genomes reveal that AUX transport *via* PINs, trihelix TFs, MIKC-type MADS genes as well as photorespiration and

diaspore storage proteins were present prior to the divergence of *K. nitens* (Fig. 1). Other features, such as the non-motile vegetative phase and filamentous growth, evolved later.

Hence, much of what was previously considered land plant-like features clearly evolved in the common ancestor of the Phragmoplastophyta (Fig. 1). These features include polyplastidy, branching, cellulose synthase rosettes, apical cell growth, several features of phytohormone networks, potential involvement of ROS in sexual reproduction and the phragmoplast. Some features evolved after the split of Charophyceae or Coleochaetophyceae such as GRAS TFs and the PPB-like isthmus band of microtubules. Life on land meant increased exposure to UV light. RNA editing repairs UV-B induced mutations in land plants (Maier et al., 2008). Editing evolved after the divergence of Charophyceae from the lineage leading to Zygnematophyceae and land plants (Cahoon et al., 2017). Key editing factors (PPR proteins) are much less abundant in *C. braunii* (57) than in the *Spirogyra* (379) or *P. patens* (100) genomes (Table S1Y). Other features, such as the multicellular sporophyte and embryogenesis, the synthesis of a complex cuticle and the ability to associate with arbuscular mycorrhizal fungi evolved at the base of the land plants, and further during land plant evolution (Fig. 1). Among the latter features are hallmarks of plants' adaptations to land. Yet, before any of these adaptations evolved, LPHGs enabled the first steps of terrestrialization. The key to their identification lies in comparative genomics studies using streptophyte algae, as exemplified here for *C. braunii*.

Acknowledgements

We thank K. Yamada, M. Göttig, M. Schallenberg-Rüdinger and F. Donges for technical assistance and S. Kato for kind assistance with strain isolation. Financial support was provided MEXT & JSPS KAKENHI (17020008 to YK, YS, SS, 20017013, 22128008, 24370095 to TN, 22770083, 24570100, 15K07185 to HS, 221S0002 to AT, AF, YS, SS); Hyogo Science and Technology Association grant to HS; DFG (GO1825/4-1 & CRC1208 to SBG, VR 132/1-1 to JdV, SFB 944 to HB and SZ, FOR964 to DB and RH, SFB 924 for DL); MEYS CR project LO1417 to SV, RS and JP; Carlsberg Foundation and the Villum Foundation's Young Investigator Programme to JH; LRSV laboratory (ANR-10-LABX-41) to PMD; Gent University to DVDS; Research Foundation Flanders (G.0317.17N to DVDS and PhD fellowship 1S17917N to LV); ERC Advanced Grants (EVO500 to LD, ETAP to JF and EDIP to JAL); Leibniz Association to MQ; NSF (DEB-1020660 and DEB-1036466 to KGK, MCB1714993 to CC, DEB 1036506 to CFD). Computation was partially performed at NIG and NIBB, Japan & High Performance and Cloud Computing University Tübingen, Baden-Württemberg bwHPC, Germany.

Authors' contributions

AF, AT, ES, HK, HS, JF, JAL, LD, MB, MQ, SAR, SR, SS, TN, YK, YS, YVP provided resources and materials.

AF, AiS, AT, SAR, SR, TN generated the draft genome.

AH, AiS, AsS, BC, CC, CD, CFD, DB, DL, DS-M, DVDS, FBH, FM, FR, GG, GT, GVR, HB, HS, JdV, JH, JMC, JP, KGK, KKU, LIACV, LV, MH, NT, PJ, PKIW, PMD, PU, RH, RK, RS, SAR, SG, SH, SR, SV, SZ, TN analyzed data.

JdV, JAL, LD, SAR, SG, TN wrote the paper.

All authors helped discuss the results and write the paper.

Declaration of Interests

The authors declare no competing interests.

References

Abouelhoda, M.I., Kurtz, S., and Ohlebusch, E. (2004). Replacing suffix trees with enhanced suffix arrays. Journal of Discrete Algorithms *2*, 53-86.

Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. Nat Methods *11*, 1144-1146.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Barve, M.P., Arie, T., Salimath, S.S., Muehlbauer, F.J., and Peever, T.L. (2003). Cloning and characterization of the mating type (MAT) locus from Ascochyta rabiei (teleomorph: Didymella rabiei) and a MAT phylogeny of legume-associated Ascochyta spp. Fungal Genet Biol *39*, 151-167.

Bauwe, H., Hagemann, M., and Fernie, A.R. (2010). Photorespiration: players, partners and origin. Trends Plant Sci *15*, 330-336.
Beilby, M.J. (2007). Action potential in charophytes. Int Rev Cytol *257*, 43-82.
Beilby, M.J. (2015). Salt tolerance at single cell level in giant-celled Characeae. Front Plant Sci *6*, 226.
Beilby, M.J., Turi, C.E., Baker, T.C., Tymm, F.J., and Murch, S.J. (2015). Circadian changes in endogenous concentrations of indole-3-acetic acid, melatonin, serotonin, abscisic acid and jasmonic acid in Characeae (Chara australis Brown). Plant Signal Behav *10*, e1082697.
Beltramo, C., Torello Marinoni, D., Perrone, I., and Botta, R. (2012). Isolation of a gene encoding for a class III peroxidase in female flower of Corylus avellana L. Mol Biol Rep *39*, 4997-5008.
Bennett, T. (2015). PIN proteins and the evolution of plant development. Trends Plant Sci *20*, 498-507.
Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120.
Boot, K.J., Libbenga, K.R., Hille, S.C., Offringa, R., and van Duijn, B. (2012). Polar auxin transport: an early invention. J Exp Bot *63*, 4213-4218.
Buschmann, H., and Zachgo, S. (2016). The Evolution of Cell Division: From Streptophyte Algae to Land Plants. Trends Plant Sci *21*, 872-883.
Bythell-Douglas, R., Rothfels, C.J., Stevenson, D.W.D., Graham, S.W., Wong, G.K., Nelson, D.C., and Bennett, T. (2017). Evolution of strigolactone receptors by gradual neo-functionalization of KAI2 paralogues. BMC Biol *15*, 52.
Cahoon, A.B., Nauss, J.A., Stanley, C.D., and Qureshi, A. (2017). Deep Transcriptome Sequencing of Two Green Algae, Chara vulgaris and Chlamydomonas reinhardtii, Provides No Evidence of Organellar RNA Editing. Genes (Basel) *8*.
Catarino, B., Hetherington, A.J., Emms, D.M., Kelly, S., and Dolan, L. (2016). The Stepwise Increase in the Number of Transcription Factor Families in the Precambrian Predated the Diversification of Plants On Land. Mol Biol Evol *33*, 2815-2819.
Chan, K.X., Phua, S.Y., Crisp, P., McQuinn, R., and Pogson, B.J. (2016). Learning the Languages of the Chloroplast: Retrograde Signaling and Beyond. Annu Rev Plant Biol *67*, 25-53.
Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E.*, et al.* (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods *10*, 563-569.
Daku, R.M., Rabbi, F., Buttigieg, J., Coulson, I.M., Horne, D., Martens, G., Ashton, N.W., and Suh, D.Y. (2016). PpASCL, the *Physcomitrella patens* Anther-Specific Chalcone Synthase-Like Enzyme Implicated in Sporopollenin Biosynthesis, Is Needed for Integrity of the Moss Spore Wall and Spore Viability. PLoS One *11*, e0146817.
Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics *27*, 1164-1165.
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 U S A.
Delaux, P.-M., Xie, X., Timme, R.E., Puech-Pages, V., Dunand, C., Lecompte, E., Delwiche, C.F., Yoneyama, K., Bécard, G., and Séjalon-Delmas, N. (2012). Origin of strigolactones in the green lineage. New Phytologist *195*, 857-871.
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 U S A *112*, 13390-13395.
Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O., and Salzberg, S.L. (1999). Alignment of whole genomes. Nucleic Acids Res *27*, 2369-2376.
Delwiche, C.F. (2016). The genomes of charophyte green algae. Adv Bot Res *78*, 255-270.
Delwiche, C.F., and Cooper, E.D. (2015). The Evolutionary Origin of a Terrestrial Flora. Current biology : CB *25*, R899-910.
Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res *45*, e18.

530 Duong, T., Cowling, A., Koch, I., and Wand, M.P. (2008). Feature significance for multivariate kernel
531 density estimation. Computational Statistics & Data Analysis *52*, 4225-4242.
532 Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space
533 complexity. BMC Bioinformatics *5*, 113.
534 Flores-Sandoval, E., Eklund, D.M., Hong, S.F., Alvarez, J.P., Fisher, T.J., Lampugnani, E.R., Golz, J.F.,
535 Vazquez-Lobo, A., Dierschke, T., Lin, S.S.*, et al.* (2018). Class C ARFs evolved before the origin of land
536 plants and antagonize differentiation and developmental transitions in Marchantia polymorpha. New
537 Phytol *218*, 1612-1630.
538 Francoz, E., Ranocha, P., Nguyen-Kim, H., Jamet, E., Burlat, V., and Dunand, C. (2015). Roles of cell wall
539 peroxidases in plant development. Phytochemistry *112*, 15-21.
540 Gao, X.H., Huang, X.Z., Xiao, S.L., and Fu, X.D. (2008). Evolutionarily conserved DELLA-mediated
541 gibberellin signaling in plants. J Integr Plant Biol *50*, 825-834.
542 Garcia, M., Myouga, F., Takechi, K., Sato, H., Nabeshima, K., Nagata, N., Takio, S., Shinozaki, K., and
543 Takano, H. (2008). An Arabidopsis homolog of the bacterial peptidoglycan synthesis enzyme MurE has
544 an essential role in chloroplast development. Plant J *53*, 924-934.
545 Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea,
546 T.P., Sykes, S.*, et al.* (2011). High-quality draft assemblies of mammalian genomes from massively
547 parallel sequence data. Proc Natl Acad Sci U S A *108*, 1513-1518.
548 Gramzow, L., and Theißen, G. (2010). A hitchhiker's guide to the MADS world of plants. Genome Biology
549 *11*, 214.
550 Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New
551 algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of
552 PhyML 3.0. Syst Biol *59*, 307-321.
553 Gutsche, N., Thurow, C., Zachgo, S., and Gatz, C. (2015). Plant-specific CC-type glutaredoxins: functions
554 in developmental processes and stress responses. Biol Chem *396*, 495-509.
555 Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D.,
556 Li, B., Lieber, M.*, et al.* (2013). De novo transcript sequence reconstruction from RNA-seq using the
557 Trinity platform for reference generation and analysis. Nature protocols *8*, 1494-1512.
558 Hackenberg, C., Kern, R., Huge, J., Stal, L.J., Tsuji, Y., Kopka, J., Shiraiwa, Y., Bauwe, H., and Hagemann,
559 M. (2011). Cyanobacterial lactate oxidases serve as essential partners in N2 fixation and evolved into
560 photorespiratory glycolate oxidases in plants. Plant Cell *23*, 2978-2990.
561 Hackenberg, D., and Pandey, S. (2014). Heterotrimeric G-proteins in green algae. An early innovation in
562 the evolution of the plant lineage. Plant Signal Behav *9*, e28457.
563 Han, G.Z. (2017). Evolution of jasmonate biosynthesis and signaling mechanisms. J Exp Bot *68*, 1323-
564 1331.
565 Heyl, A., Brault, M., Frugier, F., Kuderova, A., Lindner, A.C., Motyka, V., Rashotte, A.M., Schwartzenberg,
566 K.V., Vankova, R., and Schaller, G.E. (2013). Nomenclature for members of the two-component signaling
567 pathway of plants. Plant Physiol *161*, 1063-1065.
568 Hodgkin, A.L., and Huxley, A.F. (1952). A quantitative description of membrane current and its
569 application to conduction and excitation in nerve. J Physiol *117*, 500-544.
570 Hori, K., Maruyama, F., Fujisawa, T., Togashi, T., Yamamoto, N., Seo, M., Sato, S., Yamada, T., Mori, H.,
571 Tajima, N.*, et al.* (2014). *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial
572 adaptation. Nature Communications *5*, 3978.
573 Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res *9*, 868-877.
574 Huelsenbeck, J.P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees.
575 Bioinformatics *17*, 754-755.
576 Huson, D.H., Beier, S., Flade, I., Gorska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.J., and Tappu, R.
577 (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome
578 Sequencing Data. PLoS Comput Biol *12*, e1004957.
579 Inupakutika, M.A., Sengupta, S., Devireddy, A.R., Azad, R.K., and Mittler, R. (2016). The evolution of
580 reactive oxygen species metabolism. J Exp Bot *67*, 5933-5943.

581 Iseli, C., Jongeneel, C.V., and Bucher, P. (1999). ESTScan: a program for detecting, evaluating, and
582 reconstructing potential coding regions in EST sequences. In Proc Int Conf Intell Syst Mol Biol (Menlo
583 Park, CA, USA: American Association for Artificial Intelligence), pp. 138-148.
584 Joseph, J.M., Fey, P., Ramalingam, N., Liu, X.I., Rohlfs, M., Noegel, A.A., Muller-Taubenberger, A.,
585 Glockner, G., and Schleicher, M. (2008). The actinome of Dictyostelium discoideum in comparison to
586 actins and actin-related proteins from other organisms. PLoS One *3*, e2654.
587 Jouffroy, O., Saha, S., Mueller, L., Quesneville, H., and Maumus, F. (2016). Comprehensive repeatome
588 annotation reveals strong potential impact of repetitive elements on tomato ripening. BMC Genomics
589 *17*, 624.
590 Ju, C., Van de Poel, B., Cooper, E.D., Thierer, J.H., Gibbons, T.R., Delwiche, C.F., and Chang, C. (2015).
591 Conservation of ethylene as a plant hormone over 450 million years of evolution. Nat Plants *1*, 14004.
592 Kaplan-Levy, R.N., Brewer, P.B., Quon, T., and Smyth, D.R. (2012). The trihelix family of transcription
593 factors--light, stress and development. Trends Plant Sci *17*, 163-171.
594 Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7:
595 Improvements in Performance and Usability. Molecular Biology and Evolution *30*, 772-780.
596 Keller, O., Kollmar, M., Stanke, M., and Waack, S. (2011). A novel hybrid gene prediction method
597 employing protein multiple sequence alignments. Bioinformatics *27*, 757-763.
598 Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J. (2015). The Phyre2 web portal for
599 protein modeling, prediction and analysis. Nat Protoc *10*, 845-858.
600 Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.
601 Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate
602 alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol
603 *14*, R36.
604 Köster, J., and Rahmann, S. (2012). Snakemake--a scalable bioinformatics workflow engine.
605 Bioinformatics *28*, 2520-2522.
606 Kwantes, M., Liebsch, D., and Verelst, W. (2012). How MIKC* MADS-box genes originated and evidence
607 for their conserved function throughout the evolution of vascular plant gametophytes. Mol Biol Evol
608 *29*, 293-302.
609 Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., and Ussery, D.W. (2007). RNAmmer:
610 consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res *35*, 3100-3108.
611 Lang, D., Ullrich, K.K., Murat, F., Fuchs, J., Jenkins, J., Haas, F.B., Piednoel, M., Gundlach, H., Van Bel, M.,
612 Meyberg, R*., et al.* (2018). The Physcomitrella patens chromosome-scale assembly reveals moss
613 genome structure and evolution. Plant J *93*, 515-533.
614 Lang, D., Weiche, B., Timmerhaus, G., Richardt, S., Riano-Pachon, D.M., Correa, L.G., Reski, R., Mueller-
615 Roeber, B., and Rensing, S.A. (2010). Genome-wide phylogenetic comparative analysis of plant
616 transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. Genome
617 Biol Evol *2*, 488-503.
618 Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey,
619 V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput Biol *9*, e1003118.
620 Lecointre, G., and Le Guyader, H. (2006). The Tree of Life: A Phylogenetic Classification (Harvard
621 University Press).
622 Lee, E., Helt, G.A., Reese, J.T., Munoz-Torres, M.C., Childers, C.P., Buels, R.M., Stein, L., Holmes, I.H.,
623 Elsik, C.G., and Lewis, S.E. (2013). Web Apollo: a web-based genomic annotation editing platform.
624 Genome Biol *14*, R93.
625 Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database, C. (2011).
626 The sequence read archive. Nucleic Acids Res *39*, D19-21.
627 Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or
628 without a reference genome. BMC Bioinformatics *12*, 323.
629 Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform.
630 Bioinformatics *26*, 589-595.
631 Lin, H., and Goodenough, U.W. (2007). Gametogenesis in the Chlamydomonas reinhardtii minus mating
632 type is controlled by two genes, MID and MTD1. Genetics *176*, 913-925.

633 Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for
634 RNA-seq data with DESeq2. Genome Biol *15*, 550.
635 Maier, U.G., Bozarth, A., Funk, H.T., Zauner, S., Rensing, S.A., Schmitz-Linneweber, C., Borner, T., and
636 Tillich, M. (2008). Complex chloroplast RNA metabolism: just debugging the genetic programme? BMC
637 Biol *6*, 36.
638 Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of
639 occurrences of k-mers. Bioinformatics *27*, 764-770.
640 McInnis, S.M., Desikan, R., Hancock, J.T., and Hiscock, S.J. (2006). Production of reactive oxygen species
641 and reactive nitrogen species by angiosperm stigmas and pollen: potential signalling crosstalk? New
642 Phytol *172*, 221-228.
643 Morris, J.L., Puttick, M.N., Clark, J.W., Edwards, D., Kenrick, P., Pressel, S., Wellman, C.H., Yang, Z.,
644 Schneider, H., and Donoghue, P.C.J. (2018). The timescale of early land plant evolution. Proc Natl Acad
645 Sci U S A.
646 Nakamura, Y., Kanakagiri, S., Van, K., He, W., and Spalding, M.H. (2005). Disruption of the glycolate
647 dehydrogenase gene in the high-CO2-requiring mutant HCR89 of *Chlamydomonas reinhardtii*.
648 Canadian Journal of Botany *83*, 820-833.
649 Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective
650 stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol *32*, 268-274.
651 Nicolas, M., and Cubas, P. (2016). TCP factors: new kids on the signaling block. Curr Opin Plant Biol *33*,
652 33-41.
653 Ohtaka, K., Hori, K., Kanno, Y., Seo, M., and Ohta, H. (2017). Primitive Auxin Response without TIR1 and
654 Aux/IAA in the Charophyte Alga Klebsormidium nitens. Plant Physiol *174*, 1621-1632.
655 Park, S.Y., Fung, P., Nishimura, N., Jensen, D.R., Fujii, H., Zhao, Y., Lumba, S., Santiago, J., Rodrigues, A.,
656 Chow, T.F*., et al.* (2009). Abscisic acid inhibits type 2C protein phosphatases via the PYR/PYL family of
657 START proteins. Science *324*, 1068-1071.
658 Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in
659 eukaryotic genomes. Bioinformatics *23*, 1061-1067.
660 Pfalz, J., and Pfannschmidt, T. (2013). Essential nucleoid proteins in early chloroplast development.
661 Trends Plant Sci *18*, 186-194.
662 Pickett-Heaps, J.D. (1975). Green Algae: Structure. Reproduction and Evolution in Selected Genera
663 (Sinauer).
664 Pringsheim, M. (1862). On the Pro-Embryos of the Charae. The Annals and Magazine of Natural History
665 *59*, 321-326.
666 Pruesse, E., Peplies, J., and Glockner, F.O. (2012). SINA: accurate high-throughput multiple sequence
667 alignment of ribosomal RNA genes. Bioinformatics *28*, 1823-1829.
668 Puranik, S., Acajjaoui, S., Conn, S., Costa, L., Conn, V., Vial, A., Marcellin, R., Melzer, R., Brown, E., Hart,
669 D*., et al.* (2014). Structural basis for the oligomerization of the MADS domain transcription factor
670 SEPALLATA3 in Arabidopsis. Plant Cell *26*, 3603-3615.
671 Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc
672 Bioinformatics *47*, 11 12 11-34.
673 Ren, Y., Hansen, S.F., Ebert, B., Lau, J., and Scheller, H.V. (2014). Site-directed mutagenesis of IRX9, IRX9L
674 and IRX14 proteins involved in xylan biosynthesis: glycosyltransferase activity is not required for IRX9
675 function in Arabidopsis. PLoS One *9*, e105014.
676 Rensing, S.A. (2018). Great moments in evolution: the conquest of land by plants. Curr Opin Plant Biol
677 *42*, 49-54.
678 Rensing, S.A., Ick, J., Fawcett, J.A., Lang, D., Zimmer, A., Van de Peer, Y., and Reski, R. (2007). An ancient
679 genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella*
680 *patens*. BMC Evol Biol *7*, 130.
681 Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.-F.,
682 Lindquist, E.A., Kamisugi, Y*., et al.* (2008). The *Physcomitrella* genome reveals evolutionary insights into
683 the conquest of land by plants. Science *319*, 64-69.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012). MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst Biol *61*, 539-542.
Rost, B. (1999). Twilight zone of protein sequence alignments. Protein Eng *12*, 85-94.
Rovekamp, M., Bowman, J.L., and Grossniklaus, U. (2016). *Marchantia* MpRKD Regulates the Gametophyte-Sporophyte Transition by Keeping Egg Cells Quiescent in the Absence of Fertilization. Curr Biol *26*, 1782-1789.
Saier, M.H., Jr., Reddy, V.S., Tsu, B.V., Ahmed, M.S., Li, C., and Moreno-Hagelsieb, G. (2016). The Transporter Classification Database (TCDB): recent advances. Nucleic Acids Res *44*, D372-379.
Sakayama, H., Kasai, F., Nozaki, H., Watanabe, M.M., Kawachi, M., Shigyo, M., Nishihiro, J., Washitani, I., Krienitz, L., and Ito, M. (2009). Taxonomic reexamination of *Chara globularis* (Charales, Charophyceae) from Japan based on oospore morphology and rbcL gene sequences, and the description of *C. leptospora* sp. nov. J Phycol *45*, 917-927.
Schwacke, R., Schneider, A., van der Graaff, E., Fischer, K., Catoni, E., Desimone, M., Frommer, W.B., Flugge, U.I., and Kunze, R. (2003). ARAMEMNON, a novel database for Arabidopsis integral membrane proteins. Plant Physiol *131*, 16-26.
Scrucca, L., Fop, M., Murphy, T.B., and Raftery, A.E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. R J *8*, 289-317.
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.
Sorensen, I., Pettolino, F.A., Bacic, A., Ralph, J., Lu, F., O'Neill, M.A., Fei, Z., Rose, J.K., Domozych, D.S., and Willats, W.G. (2011). The charophycean green algae provide insights into the early origins of plant cell walls. Plant J *68*, 201-211.
Steinbiss, S., Willhoeft, U., Gremme, G., and Kurtz, S. (2009). Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Res *37*, 7002-7013.
Tan, X., Calderon-Villalobos, L.I., Sharon, M., Zheng, C., Robinson, C.V., Estelle, M., and Zheng, N. (2007). Mechanism of auxin perception by the TIR1 ubiquitin ligase. Nature *446*, 640-645.
Theißen, G., Melzer, R., and Rümpler, F. (2016). MADS-domain transcription factors and the floral quartet model of flower development: linking plant development and evolution. Development *143*, 3259-3271.
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.
Tivendale, N.D., Ross, J.J., and Cohen, J.D. (2014). The shifting paradigms of auxin biosynthesis. Trends Plant Sci *19*, 44-51.
Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.
Van Bel, M., Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y., and Vandepoele, K. (2012). Dissecting plant genomes with the PLAZA comparative genomics platform. Plant Physiol *158*, 590-600.
Vesty, E.F., Saidi, Y., Moody, L.A., Holloway, D., Whitbread, A., Needs, S., Choudhary, A., Burns, B., McLeod, D., Bradshaw, S.J*., et al.* (2016). The decision to germinate is regulated by divergent molecular networks in spores and seeds. New Phytol *211*, 952-966.
Vriet, C., Lemmens, K., Vandepoele, K., Reuzeau, C., and Russinova, E. (2015). Evolutionary trails of plant steroid genes. Trends Plant Sci *20*, 301-308.
Walker, K.L., Muller, S., Moss, D., Ehrhardt, D.W., and Smith, L.G. (2007). Arabidopsis TANGLED identifies the division plane throughout mitosis and cytokinesis. Curr Biol *17*, 1827-1836.
Wang, C., Liu, Y., Li, S.S., and Han, G.Z. (2015). Insights into the origin and evolution of the plant hormone signaling machinery. Plant Physiol *167*, 872-886.

Wang, W., Esch, J.J., Shiu, S.H., Agula, H., Binder, B.M., Chang, C., Patterson, S.E., and Bleecker, A.B. (2006). Identification of important regions for ethylene binding and signaling in the transmembrane domain of the ETR1 ethylene receptor of Arabidopsis. Plant Cell *18*, 3429-3442.
Wass, M.N., Kelley, L.A., and Sternberg, M.J. (2010). 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Res *38*, W469-473.
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 U S A *111*, E4859-4868.
Wilhelmsson, P.K.I., Muhlich, 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.
Xiong, W., He, L., Lai, J., Dooner, H.K., and Du, C. (2014). HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proc Natl Acad Sci U S A *111*, 10263-10268.
Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol *24*, 1586-1591.
Zeng, J., Dong, Z., Wu, H., Tian, Z., and Zhao, Z. (2017). Redox regulation of plant stem cell fate. EMBO J *36*, 2844-2855.

Figure titles and legends

Figure 1: Evolution of charophytic algae and land plant features

Cladogram symbolizing streptophytic evolution shows gain/expansion (green lines) and loss (red lines) of features; topology as in (Wickett et al., 2014) with phytohormone-related terms in blue and transcription factors (TF) and transcriptional regulators (TR) in brown. Expansions (and gains/losses) detected in the *Chara* lineage are shown by asterisk. See text for abbreviations. Modes of cytokinesis: a cleavage furrow with persistent telophase spindle as seen in *Klebsormidium*, and a phragmoplast seen in *Chara* that differs from that of land plants as the cell plate in *Chara* shows little centrifugal growth but is formed simultaneously across the cell's equator.

Figure 2: Life cycle and habit of *Chara braunii*

Meiosis occurs just prior to germination. At germination, a positively gravitropic rhizoid and a protonema that develops into the thallus are formed. The shoot-like thallus (phototropic and negatively gravitropic) comprises stem-like structures (axes) and whorls of branchlets (lateral organs appended to the main axes having adaxial-abaxial differentiation) at axial nodes. Growth of the axis/stem is axial from the terminal (apical) cell. Internodal cells, up to 5 cm long, are multinucleate. Internodal cells and branchlets are connected *via* specialized nodes or central cells connecting the internodes. Nodal cells can serve asexual propagation as they can form apical cells *de novo*. Female (oogonia) and male (antheridia) gametangia are borne on branchlet nodes of the monoicous thalli and generate female (egg cell) and male (sperm cell) gametes. The oogonial complex is comprised of egg cell and associated corona, jacket (five spiral tube cells), and basal cells. Sperm cells arise from filaments produced on the inner surfaces of antheridial shield cells. Upon fertilization the only diploid cell of the life cycle, the dormant zygote or oospore, is formed. Charasomes are plasmamembrane invaginations that allow carbon concentration *via* local acidification. Cells are connected by plasmodesmata. Actin-myosin based cytoplasmic streaming provides a fast transport mechanism. *C. braunii* is ecorticate, other species develop cortical cells (filaments with spine cells) from the nodes that cover the axis and branchlet internodal cells. LS: longitudinal section.

Figure 3: Gene and transposon length and density in selected plant and algal genomes.

Comparative box and whisker plots depicting distributions of feature lengths (A) and densities in 100 kbp windows (B). Organisms are ordered top-down by decreasing genome size; x-axes are logarithmic scale. Features are color-coded (legend on the right) and comprise predicted genes, helitrons, intact full-length long terminal repeat elements (flLTRE) and potentially fragmented copies (LTREs).

Figure 4: Overview of predicted presence of factors in phytohormone biosynthesis and signaling pathways of *C. braunii*.

Shown are biosynthesis enzymes (rectangles), receptors (pentagons), signal transduction components (hexagons), and TFs (ovals). Elements for which no orthologs were found (light green dashed boxes) and for which putative orthologs were identified (dark green boxes) (*cf.* Table 1, S10/11). Abbreviations as in Table 1.

Figure 5: Land plant heritage genes present in the *C. braunii* genome

(A) Growing repertoire of retrograde signaling components as well as PAPs along the streptophyte trajectory. Potential retrograde signaling orthologs are marked with colored dots (see species key). PEP-associated proteins (PAPs) are shown in the bottom inset. XRN2/XRN3 were not distinguished due to paralogy; faded dots mark the paralogy of *Chlamydomonas* FSD2 and the detection of *P. patens* PTAC7 ortholog with $E<10^{-4}$; mosses encode the cyanobacterial (i.e. non-PAP) version of MurE (Garcia et al., 2008), potentially applying for algal MurEs, too. (B) Bayesian inference phylogenetic tree of plant MADS-box genes. Posterior probabilities (≥0.6) of main branches are depicted next to the tree. Insert shows the exon-intron structures of representatives of $MIKC^C$-type genes together with the *Chara* MIKC-type genes. (C) Condensed ML tree of the LysM-RLK family. The Charales sequences form a single clade (blue branches) encompassing 7 *C. braunii* sequences. Duplication (red circle) leading to the LYK (orange) and LYR (green) subclades occurred at the base of the embryophytes. The moss and liverwort clades are clustered. (D, E) GO enrichment word clouds (biological process). Word clouds of genes down-regulated (D) or up-regulated (E) in oogonia as compared to antheridia. Font size correlates with significance; red terms are depleted, green terms enriched; top three terms each are shown. See also Fig. S5, S6, related to Table S2-S4.

Figure 6: Expression of the ROS gene network during sexual reproduction.

ROS-related gene abundance expressed in transcripts per million (TPM) was transformed to log scale and represented as heatmap in zygotes, oogonia and antheridia. Gene distance was calculated using the Euclidean method and genes were clustered using complete linkage. DEGs ($p < 0.01$) between zygotes and oogonia / oogonia and antheridia are depicted: green up arrow, log2(fold-change) > 0; red down arrow, log2(fold-change) < 0. The expanded family of class III peroxidases is shown in bold. See also Fig. S4, S6, S7, related to Table S2-S4.

823 **Tables**

Gene/ Gene family	*K. nitens*	*C. braunii*	*P. patens*	*A. thaliana*
AUX biosynthesis				
Tryptophan aminotransferase-related proteins (TAA/TAR)	1	0	6	5
YUCCA (YUC)	1	0	8	11
AUX signaling				
Transport inhibitor response 1/AUX signaling F-box (TIR1/AFB)	0	0	5	5
AUX response factor (ARF)	0	1	15	22
Indole-3-acetic acid inducible (Aux/IAA)	1/0[a]	2	4	29
AUX metabolism				
Gretchenhagen (GH)	4	1	2	20
AUX transport				
ATP-binding cassette B (ABCB)	7	5	10	22
AUX resistance 1 (AUX1/LAX)	1	0	9	4
PIN-formed 1 (PIN)	1	6	4	8
PIN-likes 1 (PILS)	3	0	3	7
CK Signaling				
CHASE domain containing histidine kinase (CHK)	6	2	11	3
Histidine-containing phosphotransfer proteins (HPT)	1	1	2	5
Response regulator type B (RRB)	1	0	5	11
Response regulator type A (RRA)	1	2	7	10
ETH biosynthesis				
1-aminocyclopropane-1-carboxylate synthase (ACS)	1	2	2	12
1-aminocyclopropane-1-carboxylate oxidase (ACO)	0	0	0	5
ETH signaling				
ETH response/ETH response sensor (ETR/ERS)	5	4	8	5
Constitutive triple response1 (CTR1)	1	2	1	1
ETH insensitive2 (EIN2)	0	1	2	1
ETH insensitive3 (EIN3)	1	4	2	6
EIN3 binding F-box protein (EBF1)	1	1	2	2

ABA biosynthesis				
Phytoene synthase1 (PSY1)	1	1	3	1
Phytoene desaturase (PDS)	2	1	2	1
Lutein deficient (LUT)	1	1	1	3
Zeaxanthin epoxidase (ZEP/ABA1)	1	1	1	1
9-Cis-epoxycarotenoid dioxygenase (NCED)	0	0	2	5
Abscisic aldehyde oxidase3 (AAO3)	1	0	2	1
ABA signaling				
Pyrabactin resistance (PYR)	0	0	4	14
Protein phosphatase 2C (PP2C - Group A)	1	0	2	9
SNF related kinase (SnRK)	1	1	4	5
CBL-interacting protein kinase (CIPK)	1	0	7	25
Calcium-dependent protein kinase (CPK)	1	2	30	34
SL synthesis				
Beta-carotene isomerase (D27)	2	1	1	1
Carotenoid cleavage dioxygenase (CCD7)	2	0	1	1
Carotenoid cleavage dioxygenase (CCD8)	2	0	1	1
SL signaling				
Alfa beta hydrolase (D14)	0	0	0	1
D14-like/ Karrikin insensitive2 (KAI2)	2	1	11	2
More axillary branching 2 (MAX2)	0	0	1	1

Table 1: Comparison of gene families operating in the biosynthesis and signaling networks of phytohormones.

A specific set of individual genes or gene families encoding steps in the phytohormone biosynthesis/signaling/metabolism/transport networks have been analysed in *K. nitens*, *C. braunii*, *P. patens* and *A. thaliana* (Table S1J).

a, kfl00094_0070 features Aux/IAA domains but also a B3 domain (see text for details).

Commented [SR5]: Character count from start to here; must be <= 60,000

Supplemental Figure Titles and Legends

Figure S1, related to STAR methods: Chromosomes in an antheridial filament of *C. braunii* (n=14, strain S276).

The chromosomes during cell division in young antheridial filaments of strain S276 were observed after Feulgen staining. The chromosome number n=14 was confirmed by counts made on chromosomes during metaphase or anaphase. Most Chara species have either n=14 or n=28 chromosomes, Nitella and the other genera have different base numbers. There are numerous examples of monoecious/dioecious species pairs in the family, with the dioecious species always displaying half the number of chromosomes than their monoecious counterpart. For Chara typically dioecious=14, monoecious=28 (or other multiples of 14). *C. braunii* is monoecious, but is unique in having the dioecious chromosome number of 14. There are no known dioecious sister taxa to *C. braunii*, perhaps due to the already reduced genome. Scale bar = 2 µm.

Figure S2, related to STAR methods: Assembly characteristics and decontamination

A) k-mer frequency analysis of the S276 paired end read data with k=25. Number of 25-mers at frequency 3 to 200 are shown with the solid line. Circles shows the points from 16 to 80 as what was recognized the major peak, presumably representing the single copy region in *C. braunii*. B) Scatter plot of mapped reads of two *C. braunii* strains on each scaffold. Blue and light blue points are scaffolds with GC content of at least 55% and less than 55%, respectively. C) Frequency distribution of scaffold wise GC content compared between putative *C. braunii* derived scaffolds (blue) and other scaffolds (green).

Figure S3, related to STAR methods: Ks-based analysis of *C. braunii* paralogs

Paranome-based WGD signature prediction. (A) Ks frequency plot highlighting mixture model components mean and standard-deviation (top: #component, bottom: mean Ks) based on raw Ks value classification. (B) Ks frequency plot highlighting mixture model components mean and standard-deviation (top: #component, bottom: mean Ks) based on log-transformed Ks value classification. (C) Ks group assignment for raw Ks classification. (D) Ks group assignment for log-transformed Ks classification. (E) Significant zero crossing (SiZer) plot. (F) Significant convexity (SiCon) plot. (G-J) Significant features of kernel density estimates using indicated bandwidths, highlighting significant gradient regions in blue and significant curvature regions in green using a significance level of 0.05. Red vertical lines represent Ks value of 0.1 and 2.0, dotted red vertical line represents Ks value of 0.235 corresponding to 12.5 Ma ago (these events might be no WGDs but only more or less recent local duplication events). For *C. braunii* no single predicted WGD signature was supported by three different bandwidth kernel densities (cf. STAR Methods).

Figure S4, related to Figure 6: Expression profiles during sexual reproduction.

Expression profile of trihelix TF genes based on RNA-seq evidence (Table S4) was visualized as A) a Venn diagram using venny (http://bioinfogp.cnb.csic.es/tools/venny/) and B) as a heatmap showing gene expression and DEGs from reproductive organs with RPKM > 1 in minimum two samples. C) Shows expression of differentially expressed TFs/TRs during sexual reproduction. D) Expression of DEGs associated with seeds during sexual reproduction. Transcripts per million (TPM) were transformed to log2 scale and clustered using the euclidean distance method and the complete clustering method (B, C, D).

Figure S5, related to Figure 5: Exon-intron structure comparison of MIKC^C-type, MIKC*-type and charophyte MIKC-type genes.

(A) Exon-intron structures of representatives of MIKC^C-type and MIKC*-type genes together with the charophyte MIKC-type genes *CbMADS1*, *CbMADS2* and *KnMADS1*. The exons encoding MADS-, I- , K- and C-domains are color coded in black, red, blue and green, respectively. Among the three Type II genes that were identified in the *C. braunii* genome only *CbMADS1* shows a canonical MIKC-type gene sequence. In contrast *CbMADS2* lacks most (but not all) introns and thus probably evolved via a retrotransposition and recombination event. *CbMADS3* lacks the conserved K-box that encodes for the protein-protein interacting K-domain (data not shown). (B and C) Analysis of exon-intron structures suggest that *CbMADS1* directly descends from an ancestral MIKC-type gene that was a common ancestor of MIKC^C- and MIKC*-type genes. (B) It was previously suggested that the N-terminal part of the K-domain of MIKC*-type proteins evolved through a duplication of two K-domain exons of an ancestral MIKC-type gene (Kwantes et al., 2012). The aligned amino acid sequences encoded by exon 2 of *CbMADS1*, and by the first K-domain exons of *KnMADS1*, *MpMADS1*, *PPM3*, *SmMADS4* and *AGL30* indeed strongly support this hypothesis. (C) In addition, striking similarities between the aligned amino acid sequences encoded by exon 5 of *CbMADS1*, exon 6 of *KnMADS1* and exons 5 and 6 of *MpMADS2*, *PPM1*, *SmMADS3* and *SEP3*, respectively, suggest that also the K-domain of MIKC^C-type proteins evolved through an exon duplication of an ancestral MIKC-type gene. This is especially intriguing considering the fact that, based on structural data, the last two K-domain exons of most if not all MIKC^C-type genes encode for a protein-protein interaction interface that facilitates tetramer formation of MIKC^C-type proteins (Puranik et al., 2014). It has already been suggested that the ability of MIKC^C-type proteins to tetramerize was an important precondition to evolve and diversify efficient developmental switches that facilitated the transition to land and the evolution of complex body plans of land plants (Theißen et al., 2016). Thus it is tempting to speculate that an exon duplication of an ancestral MIKC^C-type gene in the MRCA of extant land plants created the molecular prerequisites for this evolutionary novelty.

Figure S6, related to Figures 5/6: Transcriptome analyses of reproduction and early development.

GO enrichment word clouds (category biological process); genes down-regulated (A) or up-regulated (B) in oogonia as compared to antheridia, genes down-regulated (C) or up-regulated (D) in zygotes as compared to oogonia. Antheridia are strongly enriched with the GO category GO:0015074 “DNA integration” (A). 349 gene models expressed in antheridia were classified in this category; of these, 324 genes were found to be overlapping with a TE to at least 50 % (Table S4). Most of these genes were annotated as “integrase”, “ribonuclease H-like”, “reverse transcriptase”, and “aspartyl protease” by homology-based approach, terms typical of Ty3/Gypsy pol gene composition (Havecker et al., 2004). Ty3/Gypsy elements represent 20 % of the *C. braunii* genome. These results might indicate mobilization of retrotransposons and other mobile elements during male gametogenesis. This could be a consequence of genome rearrangement during male gamete formation. One could also imagine that mobilization and integration of retrotransposons might enhance genomic diversity during sexual reproduction.

Figure S7, related to Figure 6: Major reactive oxygen species scavenging pathway in plants.

Proteins associated with ROS scavenging are in bold. Number of genes found for *A. thaliana* and *C. braunii* (in green) are indicated in brackets. APx: Ascorbate peroxidase, Asn: ascorbate, DHA: Dehydroascorbate, DHAR: Dehydroascorbate reductase, GPx: Plant glutathione peroxidase, GR: Glutathione reductase, Grx: Glutaredoxins superfamily, GSH: reduced glutathione, GSSH: oxidized glutathione. Kat: Catalase, MDAR: Monodehydroascorbate reductase, PrxR: Peroxiredoxins family, RBOH: Respiratory burst oxidase homolog also called NADPH oxidase, SOD: Superoxide dismutase, Trx: Thioredoxins, MDA: Monodehydroascorbate, adapted from (Inupakutika et al., 2016).

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact Stefan A. Rensing (stefan.rensing@biologie.uni-marburg.de).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Two strains of *C. braunii* (S276 and S277) were used. The strain S276 was isolated from the thallus, which germinated from the bottom soil of Lake Kasumigaura (Ibaraki, Japan) and was maintained at Kobe University. The unialgal isolation of this strain was achieved as follows. First, collected oospores were surface sterilized for 5 to 8 min in 20% (v/v) NaClO (aq) with 0.05% (v/v) Tween20. The sterilized oospores were then transferred into autoclaved soil-water medium for the Charales (SWC-3), containing distilled water and two layers of substrate: a mixture of black soil and river sand on top of a layer of leaf mould. In the present study, strain S277 was newly collected from a pond at Saijo (Ehime, Japan) on October 18, 2011. Newly collected specimens of *C. braunii* were identified based on their *rbc*L DNA sequences. The methods employed for field collection and DNA barcoding followed (Sakayama et al., 2009). The thalli were collected using a handmade anchor. Total DNA was extracted from field-collected samples using the Qiagen DNeasy Plant Mini Kit. Partial *rbc*L DNA sequences were amplified using the primers CHAR-RF-1 (5'-ATGTCACCACAGACAGAAACTAA-3') and CHAR-RR-4 (5'-GCTCCTGGAGCATTTCCCCAAG-3'). PCR conditions were 95 °C for 5min; 32 cycles at 95 °C for 40s, 55 °C for 40s, and 72 °C for 1.5min; and 72 °C for 7 min using Ex Taq (Takara Bio). PCR products were sequenced using the primers CHAR-RF-1, CHAR-RR-4, CHAR-RF-2 (5'-GAGCTGTATATGAATGTCTTCG-3') and CHAR-RR-3 (5'-GTTTCTGCTTGAGATTTATA-3'). The sequences obtained were aligned with published *rbc*L DNA sequences of the genus *Chara* downloaded from GenBank. Sequence alignment was performed using MUSCLE (Edgar, 2004) with default options. The aligned dataset of the *rbc*L DNA sequences was subjected to the Neighbour-Joining (NJ) method with Jukes-Cantor distances and 1,000 bootstrap replicates, using MEGA 6.0. Based on NJ trees, field-collected samples were identified at the species level. The unialgal culture of S277 was established following the same procedure as outlined for S276. The pressed specimens of S276 and S277 (TNS-AL 209137 and 209138) were deposited at the Herbarium, Department of Botany, National Science Museum (TNS), Tsukuba, Japan. Routine culture was essentially performed at 23 °C with a 16-h light: 8-h dark cycle with 24.5 μmol photons m^{-2} s^{-1} illumination provided by fluorescent lamps using soil-water medium for the Charales (SWC-3).

METHOD DETAILS

DNA extraction

Thalli of strain S276 were harvested in SWC-3 medium, washed with distilled water, frozen in liquid nitrogen, and stored at -80 °C until DNA extraction. High molecular weight DNA was prepared by the CTAB method followed by purification with Qiagen Genomic Tip. The frozen powder was weighed and poured on 6 volumes of 2X CTAB buffer (2%

hexadecyltrimethylammonium bromide [CTAB], 1.4M NaCl, 100 mM Tris-Cl pH 8, 20 mM EDTA, 1% Polyvinylpyrrolidone, 1% 2-mercaptoethanol) on a hotplate stirrer at 60 °C. After two rounds of Chloroform:IAA 25:1 extraction, the supernatant was mixed with 3 col of CTAB precipitation buffer (1% CTAB, 50 mM Tris-Cl pH 8, 10 mM EDTA). The precipitate was recovered by centrifugation and dissolved in NaCl solution (1 M NaCl, 10 mM Tris-Cl pH 8, 1 mM EDTA), then precipitated with 0.6 vol of 2-propanol. The precipitate was dissolved in TE and further purified with a Qiagen Genomic Tip according to the manufacturer's instruction. The integrity of the DNA was confirmed with pulsed field electrophoresis using CHEF DR-II (Bio-Rad). Alternatively, genomic DNA from harvested thalli was isolated by grinding the flash frozen material, adding 15 mL extraction buffer (100mM Tris, 50mM EDTA, 500mM NaCl, 10mM 2-mercaptoethanol; pH8) and 2 mL 10% SDS, and incubating for 10 min at 65 °C with mild agitation. Subsequently, 5.4 mL 5M potassium acetate were added and incubated 20 min on ice. After centrifugation at 13,000 g for 20 min at 4 °C the DNA is precipitated by adding 14 mL 2-propanol, incubation for 30 min at -20 °C and centrifugation at 13,000 g for 15 min at 4 °C. After the isopropanol precipitation the air dryed pellet was dissolved in 700 µl 1x TE buffer (pH 8), 1-3µl RNaseA (10mg/ml) was added and incubated for 10 min at 37 °C. To purify the DNA 600 µl phenol/chloroform 1:1 were added, mixed, centrifuged at 10,000 g for one minute and the aqueous phase extracted. To this phase 600 µl chloroform/isoamylalcohol 24:1 were added, mixed, centrifuged at 10,000 g for one minute and the aqueous phase extracted. To precipitate the DNA 70 µl 3M Na-acetate and 500 µl isopropanol were added, mixed and centrifuged at 10,000 g for ten minutes. The pellet was washed with one ml 70% ethanol, dried and afterwards was dissolved in deionized water. Quality was controlled using Nanodrop, Qubit measurement and agarose gel electrophoresis.

Chromosome observation

The thalli with young antheridia were collected within the first hour of the dark period and fixed in ethanol:acetic acid (3:1). Fixed material was stored at 4 °C until used. Chromosome preparations were made using the Feulgen squash method (Fig. S1). Fixed samples were rehydrated by passing through a graded series of ethanols and rinsed gently in distilled water. The samples were treated with 1N HCl for 5 min at room temperature, then treated with 1N HCl for 8 min in a water bath at 60 °C, and rinsed gently in distilled water. Afterwards, the samples were transferred into Schiff's reagent (Merck Millipore) for 60 min at room temperature. After rinsing the samples in distilled water, they were squashed and observed.

Genome sequencing and assembly

Genomic DNA of the uni-algal strain S276 isolated from Lake Kasumigaura (Ibaraki, Japan) was sequenced as the reference genome using Illumina technology and sequences were compared with those of the strain S277 that was isolated from the pond at Ehime (Japan). Approximately 0.25 Gbp of scaffolds were present in only one of the datasets and found to be of bacterial origin. After removal of these prokaryotic sequences, 1.75 Gbp of scaffold data (N50 size of 2.26 Mbp at #234) were obtained, of which 1.43 Gbp were assembled into contigs. This corresponds to ~74% of the *C. braunii* genome as measured by flow cytometry (1.89-1.96 Gbp) and to ~61% of the 2.35 Gbp estimated by k-mer analysis. The plastid and mitochondrial genome were assembled separately to recover 187,771 and 67,059 bp circular genomes, respectively.

Genome sequencing of *C. braunii* strain S276

A paired-end library with insert size of 250 bp was constructed using an S2 ultrasonicator (Covaris) and a TruSeq DNA PCR-Free LT Sample Prep Kit (Illumina) according to the manufacturer's protocols. The products were size-selected on an agarose gel and purified using the Qiagen MinElute Gel Extraction Kit. Nucleotide sequences were determined for 150 bp from both ends with an Illumina HiSeq 2500. Sixteen Mate-pair libraries were constructed using a Nextera Mate-pair library construction kit with standard and modified input DNA of 5.6, 8, and 20 μg in the reaction. The first set, four libraries were constructed using the standard protocol, a gel-free method starting with 1 μg DNA (one library), and gel-excision starting with 4 μg DNA (three libraries). In the Gel-free protocol tagmented DNA was purified with AMPure XP resulting in a broad size with a peak at 2.7 kbp. In the Gel (+) protocol, the size range was 3-5 kbp, 5-8 kbp, and larger, resulting in a peak of 4.5, 5.8, and 9 kbp, respectively, as measured with a Bioanalyzer after purification with a Zymoclean Large Fragment DNA Recovery Kit. After circularization, fragmentation with Covaris S2, end-repair, A-tailing and adapter ligation, gel-free and 4.5 kbp library were amplified for 10 cycles, whereas 5.8 kbp and 9 kbp libraries were amplified for 15 cycles. After purification and quantification, the libraries were further subjected to 8, 6, 6, and 8 cycles of PCR, for gel-free, 4.5, 5.8 and 9 kbp libraries, respectively (Table S1A).

In the second set, two libraries were constructed using 20 μg DNA instead of the standard 4 μg DNA to obtain larger fragment size distribution after tagmentation. In this sample, though the large molecules were not well separated on the agarose gel, three fractions, thick band at high molecular weight above all marker bands, below the band to 12 kbp, and a 8-12kbp fraction were recovered. The size of the recovered DNA could not be measured accurately using a Bioanalyzer, though the peak was around the 17kbp marker. The final amplification was done for 21 cycles and additional 8 cycles. The lowest 8-12 kbp fraction did not amplify well and was not used in further analysis.

In the third set, five libraries were constructed using 5.6 μg of starting DNA (1.4-fold of standard) and an additional five libraries using 8.0 μg of starting DNA (2-fold of standard); pulsed field electrophoresis on a CHEF-DRII (Bio-Rad) was used for the separation after the tagmentation. The electrophoretic conditions were 6 V/cm, 11 hours, switch time 1-6 s, on 1% agarose gel, in 0.5 X TBE buffer. The gel was stained with SYBR Gold and the gel slices were recovered in five fractions each. The lower limit of each slice was 5.0, 7.5, 10.0, 15.0, and 23.5 kbp. After purification, the DNA was immediately subjected to circularization without measuring its size. The final amplification was conducted for 15 cycles. Of these (Table S1A), 15 had good insert size distribution when mapped to a preliminary version of the assembly, but one (S276MP3 xk) had not and thus excluded for further analysis.

Another two mate pair libraries were constructed by GATC (3-4 kbp fragment size) and sequenced on an Illumina HiSeq 2000. One library was constructed using Crelox with an insert size of 3 kbp. DNA was fragmented using the Covaris S2 AFA instrument and sequencing was performed on an Illumina HiSeq 2000 at 2 x 100 bp.

K-mer frequency analysis

K-mer frequency with $k = 25$ in the paired end reads were counted with JELLYFISH (Marcais and Kingsford, 2011), applying the min-quality=20 option. A clear peak at 51 was observed with a valley at 16 (Fig. S2A). The peak at 51 was interpreted as the single copy genomic sequence and those less than 16 were mostly k-mers containing sequencing errors. The cumulative *k*-mer count from 16 upto 10,000 (which was the default upper limit of JELLYFISH) divided by 51 suggested the genome size be 2.355 Gbp. Note that this number includes *k*-mers derived from organellar and bacterial sequences and supposed to be overestimate for the nuclear genome size. With the peak at 51, the amount of paired-end reads are supposed to be sufficient for the assembly. The region from 16 to 80 as the putative single copy region comprised 0.95 Gbp.

Assembly

The raw sequences were assembled with ALLPATHS-LG (Gnerre et al., 2011). Initially the assembly started with R48517 on a machine having 768 GB of memory and 32 CPU cores. After running a month this process stopped at UnipathPatcher phase. Continuation was tried with the settings: PATCH_UNIPATHS=False FIX_LOCAL=False PATCH_SCAFFOLDS=False FIX_SOME_INDELS=False; unfortunately this failed again. The run directory was copied to a machine having 2 TB of memory and 80 cores and the assembly was continued with R48777 and completed after another twenty days (with 48 slots=threads), with reported peak memory usage of 1,756 GB. The assembly resulted in 28,091 scaffolds with a total length of 1.99 Gbp, comprised of 250,979 contigs with a total length of 1.65 Gbp. The library information is summarized in Table S1B.

Genome sequencing of *C. braunii* strain S277

Thalli of strain S277 were harvested in SWC-3 medium, washed with distilled water, frozen in liquid nitrogen, and stored at -80 °C until DNA extraction. Total DNA was extracted as described above. A paired end library was constructed using a TruSeq DNA PCR-free library preparation kit (Illumina) and sequenced with HiSEQ (DRA accession: DRR054048). 1.1 µg of DNA was fragmented with Covaris S2, using micro tube, duty cycle 10%, intensity 4, 200 cycles/burst and total time of 80 s. The fragments were size selected using a bead-based method following the 350-bp protocol.

PacBio sequencing of fosmid clones for quality control

C. braunii S276 genomic DNA was cloned into the pNGS fosmid vector using the aNxSeq 40 kbp Mate-Pair Cloning Kit (Lucigen). Six fosmid clones with verified end sequence and one 96 well plate of undetermined clones were pooled and shotgun sequenced on a PacBio SMRT cell (608 Mbp, 63,768 reads post-filtering). The resulting reads were assembled into contigs using HGAP (Chin et al., 2013) in smrtanalysis (PacificBiosciences). The contig sequences were further polished with two rounds of Quiver. Bacterial contamination was removed using MEGAN, and comparative mapping of S276 and S277 reads, resulting in 22 probable *C. braunii* contigs. All but one of those could be BLAST-mapped to the assembly. One clone appeared to be chimeric based on mapping Illumina mate-pair library data on the clone. Of the remaining 20, 14 were mapping to single scaffolds, the other 6 to 2-4 scaffolds. 10 of the 22 contigs were found to map with >=95% identity and >= 90% coverage to the assembly, the remaining 12 did not meet these parameters, probably due to assembly gaps. In summary, 45%

of the assembled fosmid clones had high quality representations in the assembly, and 91% could be mapped, demonstrating the good quality of the assembly.

Distinction of bacterial sequences

Paired end sequences of S276 and S277 were mapped to the assembly with bwa mem (Burrows-Wheeler Aligner) (Li and Durbin, 2010) and the number of mapped sequences were counted on each scaffold (Fig. S2). Number of tags of both samples on each scaffold was plotted and we found two groups. The two groups were separated by a line in which S277 had 1/100 of S276 tags (Fig. S2B). The GC content of each scaffold was calculated and compared between the two groups. The group showing less tags in S277 had a higher GC content distribution (Fig. S2C). Thus, these scaffolds were presumed to be derived of different organisms, which were probably bacteria that survived autoclaving. In addition, scaffold_64 was found to be of bacterial origin in manual inspection during gene prediction. Further, the genomic scaffolds were split into 1 kbp fragments. Using tera-BLASTn 9.0.0 on DeCypher 9.0.0.25 (http://www.timelogic.com/catalog/757/tera-blast) each fragment was BLASTed against the NCBI nt database. The BLAST output was analysed by MEGAN 6 (Huson et al., 2016) and bacterial hits assigned to the 1 kbp fragments. All scaffolds containing more than 50 % of bacterial hit fragments were extracted. If no non-bacterial hits were contained on the scaffold and the bit score of the bacterial contamination exceeded 50 per hit the scaffold was removed as contamination. This affected 153 scaffolds with a total length of 312 kbp (Table S5), containing 120 gene models (marked in Table S4). Thus, 11,655 scaffolds totaling 1,751,225,565 bp, comprised of 234,221 contigs totaling 1,429,911,168 bp were recovered as representing the *C. braunii* nuclear genome. N50 scaffold size, and N50 contig size were 2,261,426 bp (at #234) and 10,124 bp (at #41,610), respectively.

Microbiome analysis

The diversity of microorganisms is expected to be low due to lab-culturing conditions and DNA sequence extraction protocols. To isolate the microorganisms remaining in the bulk of data, we mapped reads to the eukaryotic genome and only analyzed reads left unmapped. Given that S276 and S277 were reared at different geographical locations, analyzes were done on both sets separately. The two separate sets of remaining reads were assembled into contigs and analyzed from a meta-genomics point of view. Two separate assemblies have been generated using CLC-assembly cell using the larger word-size (kmer) of 50 nt to force more specificity (CLC bio, Aarhus, Denmark). These assemblies resulted in respectively 322685 contigs with a total size of 76.7 Mbp (N50 242 bp, max size 167358 bp, min size 100 bp) and 325720 contigs with a total size of 90.1 Mbp (N50 373 bp, max size 172440 bp, min size 100 bp). The obtained contigs represent a mixture of microorganisms that where clustered using CONCOCT (Alneberg et al., 2014) according to the manual, using BEDtools (Quinlan, 2014), Picard-tools and R, to create and format the needed input files. Several runs were done, aiming at providing the minimal number of differentiated clusters. In some cases large clusters were isolated and submitted again for a new round of clustering. The clusters (or bins) were calculated based on read coverage and sequence tetramer composition of the contigs following an iterative fitting of mixture-of-Gaussian models on the available data; each group is supposed to represent an organism that was further characterized to establish the species. Taxonomic assignment of the bins was performed using a similarity-based labeling of the fragments with MEGAN5. A first assessment

of the quality and completeness of the bins was done by monitoring the presence of 36 COG single copy genes. 16S rRNA genes were isolated from the sequences using online RNAmmer 1.2 Server (Lagesen et al., 2007) and provided to SINA Alignment Service within Silva database for classification (Pruesse et al., 2012). Not all clusters could be identified up to species level, but for those for which we could find a reference genome, we show also a level of completeness by comparing to the respective reference genomes using nucmer from the MUMmer (Delcher et al., 1999) v3.23 package (Table S1T, S1U).

Transcriptome sequencing

Thalli of strain S276 were harvested in SWC-3 medium under controlled laboratory conditions at 23 °C with a 16-h light: 8-h dark cycle with 24.5 µmol photons m^{-2} s^{-1} illumination provided by fluorescent lamps. Two and seven different samples, for full-length cDNA and RNA-seq analyses, respectively, were collected, frozen in liquid nitrogen, and stored at −80 °C until further processing. Frozen samples were ground in liquid nitrogen. Total RNAs were then extracted with ISOGEN (Nippon Gene, Tokyo, Japan), and purified using the Qiagen RNeasy Plant Mini Kit. For the extraction of total RNA in oospores and rhizoids, Fruit-mate (Takara Bio, Shiga, Japan) was used prior to the extraction by ISOGEN. Full-length cDNA libraries were constructed using the oligo-capping method. Total RNA was treated with bacterial alkaline phosphatase (BAP; Takara) at 37°C for 40 min with RNasin (Promega). After extraction with phenol:chloroform (1:1) twice and ethanol precipitation, the RNA was treated with tobacco acid pyrophosphatase (TAP; in house purified) with RNasin at 37°C for 45 min. The BAP-TAP treated RNA were ligated with 5′-oligo (5′-AGC AUC GAG UCG GCC UUG UUG GCC UAC UGG-3′) using T4 RNA ligase (Takara).The first strand cDNAs were amplified using 5′ (5′-AGC ATC GAG TCG GCC TTG TTG-3′) and 3′ (5′-GCG GCT GAA GAC GGC CTA TGT-3′) PCR primers. The amplified cDNAs were digested with SfiI and cloned into DraIII-digested pME18S-FL3-3 (AB009864). Clones were picked and sequenced with ABI sequencers at National Institute of Genetics, Japan. After filtering for vector, synthetic oligonucleotides, and low-quality sequences 73,388 reads were left in total (Table S1D). RNA-seq libraries were constructed via the Illumina mRNA-Seq Sample Prep Kit using RNA extracted from various tissues (Table S1E). 76 or 101 bp paired end sequencing was performed on an Illumina HiSEQ 2000. Additionally, a late reproductive phase thalli (harvested 2-3 weeks after appearing of the gametangial primordia) library was constructed as RNA-ligation based stranded library using the combined method of mRNA-Seq Sample Prep Kit and Small RNA Sample Preparation Kit (Illumina), following the manufacturer's instructions. This library was sequenced by 76 bp single end sequencing performed on a GAIIx (Illumina).

Quantitative transcriptome comparison of antheridia, oogonia, and zygotes

Antheridia and oogonia were hand-dissected in Qiagen RNA*later* from *C. braunii* thalli (strain S276) grown under a 14:10 hours light:dark cycle at 22 °C. Zygotes were collected once detached from mother plants grown in identical conditions. Samples were flash frozen in liquid nitrogen then kept at -80 °C until further processing. Approximately 20 mg of starting material was ground in liquid nitrogen then total RNA was extracted using Ambion mirVana kit following manufacturer’s recommendations. DNA was digested from RNA extracts using Promega RQ1 DNase and RNA was cleaned using a Qiagen RNeasy MinElute Cleanup Kit. RNA was then amplified using an Ovation RNA-Seq System V2 (NuGEN) amplification kit

following manufacturer's protocol. Final amplified cDNAs were cleaned using the Qiagen PCR cleanup kit. Three biological replicates were obtained for antheridia, oogonia and zygotes. One sample containing vegetative and reproductive tissues was similarly prepared, except for the amplification step. 20 µg of RNA from each replicate was paired-end sequenced on an Illumina HiSeq 2000 platform at the Beijing Genomics Institute in China; at least 2 x 10 million reads were obtained per sample. Reads were processed to remove low quality sequences, PCR adapters, foreign sequences introduced by the amplification procedure and any detectable bias using Trimmomatic v0.36 (Bolger et al., 2014) and Perl scripts. Transcript were inferred from the reads pooled and aligned to the *C. braunii* genome sequence using Tophat v2.1.0 (Kim et al., 2013) and Cufflinks v2.0.2 (Trapnell et al., 2010). Both programs were given the *C. braunii* genomic structure as a guide. A custom Perl script was then used to clean Cufflinks predictions from spurious gene fusions and other detectable problems. Unaligned reads were further normalised, assembled and scaffolded into transcripts. Both reference guided and *de novo* assemblies were merged. Coding sequences were predicted, and sequence annotation and GO terms were obtained from transcripts using a pipeline based on BLAST v2.2.29 (Altschul et al., 1997) and TransDecoder v2.0.1 (Haas et al., 2013). A summary of assembly and read mapping statistics is presented in Table S1W. Read counts were obtained by mapping reads onto the inferred transcriptome with RSEM v1.2.11 (Li and Dewey, 2011). Differential expression was tested between zygotes and oogonia samples and between oogonia and antheridia samples and was conducted in R using DESeq2 v1.14.1 (Love et al., 2014). Genes were considered differentially expressed between two conditions with an adjusted p-value < 0.01 and a log2 fold-change (logFC) > 2. Differentially expressed genes are listed in Table S2. GO terms enrichment analysis was conducted in R using topGO v2.22.0. Enriched GO terms and associated genes are listed in Table S3. Heatmaps were generated using R and the package pheatmap v1.0.8. Visualization of the GO terms was implemented using word clouds via the http://www.wordle.net application. The weight of the given terms was defined as the -log10(q-values) and the color scheme used for the visualization was red for down-regulated GO terms and green for those up-regulated. See Table S2 for DEGs and Table S3 for GO analyses.

Identification of repeat sequences with RepeatModeler/RepeatMasker

A species-specific repeat model was constructed using RepeatModeler Version open-1.0.7 with ncbi engine. Repeats were identified using RepeatMasker version open-4.0.5 with Search Engine: NCBI/RMBLAST [2.2.27+] and RepeatMaskerLib.embl (Complete Database: 20140131), resulting in masking 46% of the genome. The breakdown is shown in Table S1F.

Gene prediction

High throughput cDNA sequencing (RNA-seq) was conducted on several libraries representing vegetative and reproductive stages, including thallus, gametangia and zygotes. These data were used together with full-length cDNA sequences to annotate the genome with AUGUSTUS. 35,445 putatively protein-coding genes were identified, of which 63% could be annotated using similarity-based approaches. A total of 13,331 gene models overlap to at least 50% with TE evidence and thus might not represent canonical protein-coding genes, bringing the number of protein-encoding genes down to 23,546. In total, the expression of 12,388 of those (53%) was supported by RNA-seq data (Table S4). Reciprocal best BLAST (Altschul et al., 1997) hit analysis of the *C. braunii* protein set revealed a high percentage presence of core gene sets:

96.43% of eukaryotic benchmarking universal single-copy orthologs (BUSCO, (Simao et al., 2015)), 98.65% CEGMA core eukaryotic genes (Parra et al., 2007), and 93.96% core gene families for green plants (Van Bel et al., 2012).

Gene prediction with Augustus (Keller et al., 2011) was performed following https://computationalbiologysite.wordpress.com/2013/07/25/incorporating-rnaseq-tophat-to-augustus/. Initial models were created based on the CEGMA output. RNA-seq data was mapped to the RepeatMasker masked *C. braunii* genome. Each accepted_hits.bam was sorted and processed with filterBam --uniq (--paired for paired data). Evidence of introns was extracted using bam2hints –intronsonly to obtain intron_hints.gff. The first round of Augustus was run with this as hints. An exon-exon junction database was constructed based on this output and bowtie was used to map the reads to the junctions. These mappings were further merged to the first intron hints and the second round of augustus was run. Gene prediction at this phase was manually investigated and confirmed genes on scaffold_0 and scaffold_2 were chosen and adjusted for the 5' and 3' ends of UTR based on RNA-seq mapping on Web Apollo (Lee et al., 2013). Thus, 120 manually inspected gene models were used to retrain Augustus. Construction of exon-part hints through wig file were performed according to http://augustus.gobics.de/binaries/readme.rnaseq.html. For the stranded RNA-seq data, forward and reverse mapped reads were separated with samtools and assigned the strand accordingly. Repeat hints were prepared by processing the gff file created by the RepeatMasker with "sed –e s/similarity/nonexonpart/ -e 's/Target.*/src=RM/'". Amino acid sequence of *A. thaliana* (TAIR10_pep_20110103_representative_gene_model_updated) and *P. patens* (P.patens.V6_filtered_cosmoss_proteins.fas) were mapped to the genome using exonerate and converted as hint data. The full-length EST sequences were mapped using blat (Kent, 2002) with -minIdentity=92 -extendThroughN parameters and converted to EST hints. All these hints were merged to a single hints file and the final run of Augustus was run with --gff3=on --UTR=on --alternatives-from-evidence=true --allow_hinted_splicesites=atac with a merged hints file. The output was collected and gene models predicted on the 11,808 scaffolds that were treated as *C. braunii* genome. Thus, we obtained 36,877 transcripts from 35,883 loci. For annotation see Table S4.

Assembly of organellar genomes

Organellar genomes were assembled using NOVOPlasty (Dierckxsens et al., 2017) v2.5.3. For chloroplast genome, two lanes of paired end data were processed using the *Chara vulgaris* chloroplast genome (NC_008097.1) as seed. This resulted in 4 possible reconstructions, two in 187 kbp and the remaining two in 200 kbp, i.e. contig arrangement 01+02+03+04+06, 01+04+05, 01+02+03+04+05, or 01+04+06. The differences are on whether 02 and 03 are inserted and whether the end is 05 or 06. 02 and 03 is contained in 01 and seems to represent an inverted repeat region and insertion of them would be excess. The 05 and 06 contain 27,447-bp common sequence, which is the small single copy region. Given there are about equal number of molecules that is flipped at the inverted repeat region, both reconstructions are equally valid and one is arbitrarily chosen. The mitochondrial genome was assembled using the *C. vulgaris* mitochondrial genome (NC_005255.1) as seed input and specifying the chloroplast genome obtained as above. This resulted in a single circularized assembly of 67,059 bp, which is very close to 67,737 of the *C. vulgaris* mitochondrial genome.

Repeat/TE annotation

Repetitive elements collectively contribute approximately 1.1 Gbp of the genome assembly. This estimate is probably low, given that highly similar repeats are challenging to assemble and that there is ~0.5 Gbp size difference between the ungapped (1.43 Gbp) assembly and C-value estimates (1.9 Gbp). Transposable elements (TEs) and unclassified repeats are abundant (61% and 37% of repeat annotation, respectively), with Gypsy-type LTR retrotransposons representing 24% (343 Mbp) of the ungapped assembly (Table S1G).

We have used the REPET package v2.4 to perform *de novo* identification, classification and annotation of repetitive elements in the *C. braunii* assembly as decribed in (Jouffroy et al., 2016). We first launched the TEdenovo pipeline on a sub-genome comprising contigs of size above 20 kb and representing a total of 362 Mb (12,655 contigs). We used default settings except that the minimum number of copies per group was set to 5 (minNbSeqPerGroup: 5), resulting in a library of 3,140 consensus sequences. This library was subsequently filtered by using the TEannot pipeline against the whole assembly and discarding consensus sequences without a single full length match, resulting in a library of 2,161 sequences. This filtered library was used to annotate the whole genome assembly using the TEannot pipeline. Threshold annotation scores were determined for each consensus as the 99th percentile of the scores obtained against a randomized sequence (whole genome reversed, not complemented and masked with TRF). Consensus sequences were then classified using the features detected with PASTEC followed by semi-manual curation. In addition to the HMM comparison against PFAM implemented in PASTEC, we have also used RPS-BLAST (-F T -e 1e-2) to search for more remote homologies against a library of CDD domains identified in the repbase library.

Several unclassified consensus sequences have been classified in putative retrotransposons because they contain at least one of the following domains: cd00024 Chromatin organization modifier, cd00303 Retropepsins, cd01650 RT nLTR, cd01651 RT G2 intron, cd05482 Retropepsins, cd06095 Retropepsin, cd06222 RNase H, pfam00385 Chromo, pfam00552 Integrase, pfam00665 Integrase, pfam02093 Gag P30, pfam03708 Avian retrovirus envelope protein, pfam03732 Retrotransposon gag protein, pfam07727 Reverse transcriptase, pfam10536 Plant mobile domain, pfam13966 zinc-binding in reverse transcriptase, pfam13975 gag-polyprotein putative aspartyl protease, pfam13976 GAG-pre-integrase domain, and smart00298 Chromatin organization modifier domain.

Based on the REPET results, percentage overlap of protein coding gene models with TEs was assessed and added to Table S4. Gene models overlapping to 100% with TE evidence are considered true TE genes, while those overlapping to at least 50% (but less than 100%) might be protein-coding genes present in TE regions, or might encode TE-based proteins. All genes were kept in the gene catalog so that individual evaluation (e.g. based on the homology-based annotation) is possible.

Screening for whole genome duplication events

To identify whole genome duplication (WGD) events we employed the KeyS software (Rensing et al., 2007) to obtain Ks (synonymous substitution) distributions of paralogous genes for *C. braunii*. In brief, paralogous genes were defined by a self-BLAST retaining only BLAST hits that showed at least 50% query and subject coverage and an alignment length according to the

twilight zone *sensu* (Rost, 1999). Gene pairs with a BLAST identity of 98% or higher were further tested at the nucleic acid level to remove nearly identical sequences using optimal global alignments and a threshold of 98% identity. For nearly identical gene pairs only the longer sequence was kept and all gene pairs containing the shorter sequence were discarded (Rensing et al., 2007). The paralogous genes were further clustered using a minimal connectivity threshold of 50% (half linkage) and Ks values were calculated at the cluster nodes (representing duplication events rather than gene pairs) using the maximum likelihood method of CODEML implemented in PAML v4.7 (Yang, 2007).

The following procedure has been described recently (Lang et al., 2018), please see there for related citations. Briefly, we employed mixture modeling to find WGD signatures using the *mclust* v5.1 R package (Scrucca et al., 2016) to fit a mixture model of Gaussian distributions to the raw Ks and log-transformed Ks distributions. All Ks values ≤ 0.1 were excluded for analysis to avoid the incorporation of allelic and/or splice variants and to prevent the fitting of a component to infinity, while Ks values > 5.0 were removed because of Ks saturation. Further, only WGD signatures were evaluated between the Ks range of 0.235 (12.5 Ma ago) to account for recently duplicated gene pairs to Ks of 2.0 to account for misleading mixture modeling above this upper limit. Because model selection criteria used to identify the optimal number of components in the mixture model are prone to over fitting we also used SiZer and SiCon as implemented in the *feature* v1.2.13 R package (Duong et al., 2008) to distinguish components corresponding to WGD features at a bandwidth of 0.0188, 0.047, 0.094 and 0.188 (corresponding to 1 Ma, 2.5 Ma, 5 Ma and 10 Ma ago) and a significance level of 0.05.

Deconvolution of the overlapping distributions that can be derived from paranome-based Ks values without structural information shows that using mixture model estimation based on log-transformed Ks values mimics structure-based WGD predictions better than using raw Ks values, and can predict young WGD signatures and can pin point older WGD signatures (Lang et al., 2018). Since WGD signature prediction based on paranome-based Ks values can be misleading and is prone to over prediction we only considered Ks distribution peaks in a range of 0.235 to 2.0 as possible WGD signatures, thus excluding young paralogs potentially derived from tandem or segmental duplication and those for which accurate dating cannot be achieved due to high age (Fig. S3).

Genome comparison

C. braunii was compared with eight further Viridiplantae genomes. In addition to the genome length, GC content and the number of annotated genes, the mean intergenic and the mean intron length were calculated. The intergenic length was performed by extracting the genome regions not covered by the gff3 annotation file with bedtools complement (Quinlan, 2014) version 2.25.0. The intron length was calculated by extracting the distance between the annotated CDS regions. Both mean length and the corresponding standard deviation were calculated using awk (Table S1L). The gene density of the *C. braunii* genome is relatively sparse as compared to e.g. *A. thaliana*, *O. sativa* (rice) or two algae (*K. nitens* and *Chlamydomonas reinhardtii*), but similar to other Gbp-sized genomes like *Z. mays* or *H. vulgare* (Fig. 3 and Table S1L); the distance between genes is comparable to the approximately equal-sized *Z. mays* genome.

Comparative analysis of gene and transposons in selected plant and algae species

The genome sequences and annotations of *K. nitens*, *C. reinhardtii*, *A. thaliana*, *M. polymorpha*, *Oryza sativa*, *P. patens*, *C. braunii*, *Z. mays*, *H. vulgare* were downloaded and processed with GAG and the genome tools gff3 validator, to obtain consistent annotation files. For each annotated gene, intronic regions were inferred using the GenomeTools gff3 program. The *K. nitens* annotation file was manually curated for consistency with the other annotations and the GFF3 data standard.

Subsequently, intact full-length long terminal repeat transposon elements (LTREs) were predicted using the GenomeTools LTRharvest and LTRdigest software (Steinbiss et al., 2009) utilizing a set of TE-associated PFAM domains and a compilation of eukaryotic tRNAs. The pipeline was implemented as a BASH/PBS shell script (run_LTR_harvest_digest.sh). The resulting set of candidate LTREs was filtered to contain 2 LTRs, >=1 protein domain match and 2 target site duplications. These filtered elements were considered to represent intact full-length LTREs whose nucleotide sequences were extracted and searched against the genome using Vmatch requiring >=80% sequence identity and 100 bp alignment length. Depending on the repeat content and genome size, genomes where either split at gap boundaries into preferably 100 Mbp stretches using the UCSC toolkit faSplit (A: Snakemake workflow: split_approach), or directly processed as a whole FASTA file (B: Snakemake workflow: vmatch_mask) (Köster and Rahmann, 2012). Resulting putative LTRE fragments were merged into non-redundant, non-overlapping regions using the reduce function implemented in the R/Bioconductor package GenomicRanges (A) (Lawrence et al., 2013) or the bedtools merge program (B).

Helitrons were predicted using the HelitronScanner software using the parameters reported for element inference and copy number prediction in plant genomes reported in the initial manuscript (Xiong et al., 2014). Additional fragments were inferred by matching 50 bp from the 3' terminus of each full-length helitrons against the respective genome utilizing Vmatch (Abouelhoda et al., 2004) following the same approach as described for LTREs. Resulting matches and full-length helitrons were merged into non-redundant, non-overlapping regions using the bedtools merge program. The pipeline was implemented in the Snakemake workflow in folder helitrons/.

Gene-to-gene, gene-to-LTRE, LTRE-to-gene and LTRE-to-LTRE distances were inferred using an R script utilizing the distanceToNearest function from the R/Bioconductor GenomicRanges package (get_distances.R/get_distances.sh). Subsequent data analysis and plotting was carried out and documented in the R Jupyter Notebooks: folder analysis/: analyseWindows.ipynb, Distances.ipynb, Introns.ipynb, Lengths.ipynb. All described, generated materials and software needed to reproduce this analysis are available from the accompanying Mendeley Data repository (doi:10.17632/9hzzf9m4kh.1), arranged as an archive ("ComparativeTE_and_genes.Lang.tar.gz") that contains input, output and scripts.

In-depth analyses of specific gene families

Cell wall biosynthesis

Glycosyltransferases in the *C. braunii* genome assembly were initially identified via BLAST, using the Carbohydrate Acting enZYme database (CAZY) as of 2016-06-01 as query and a cut-off value of 10^{-25}. The sequences were manually verified by alignment with known cell wall biosynthetic glucosyltransferases and deposited in Table S1H. Phylogenetic trees were

constructed using Phylogeny.fr with standard settings, starting with muscle alignment, curation of alignment by deletion of positions with gaps, and finally PhyML maximum likehood tree construction (Guindon et al., 2010). The phylogenetic trees (Data S1A, B) were statistically supported by approximate likelihood-ratio tests using default settings and values between 0 and 1 were obtained, as with bootstrap values. Approximate likelihood-ratio-test (aLRT) values were included when values were under 0.7 where *C. braunii* sequences are present.

Cell division

In order to compare the mode of cell division of algae and land plants we compiled a list of 221 *Arabidopsis* genes involved in cytokinesis (Table S1C), focusing on genes required for phragmoplast and PPB function. With these 221 *A. thaliana* proteins, a BLASTp (version 2.6.0+) search was performed against published plant and algal genomic/transcriptomic datasets (key resource table), including *C. braunii* and *K. nitens*. The e-value cutoff was set to 1E-4 and the number of database sequences to show alignment for was set to 3,000. The BLAST result was filtered according to (Rost, 1999) to keep homologous sequences only. Mutiple sequence alignments for phylogenetic trees of protein families were conducted using MAFFT (Katoh and Standley, 2013) in the automatic mode, and manually curated. The best fitting evolutionary model based was determined using ProtTest (Darriba et al., 2011) and applied in Bayesian phylogenetic inference using MrBayes (Ronquist et al., 2012) with two hot and two cold chains (Data S1Q-U) until the standard deviation of split frequencies dropped below 0.01 or for 6 mio generations (actin and cyclin).

Using the amplification score that shows potential gene expansion between *K. nitens* and *C. braunii* (Table S1C) we performed phylogenetic analyses as outlined above and found cyclin genes to be amplified in *C. braunii*, suggesting a more intricate regulation of the cell cycle as compared to *K. nitens*. While there is a single A1-type cyclin in both algae, the *C. braunii* genome encodes three B1-type cyclins (like *A. thaliana*), whereas *K. nitens* encodes only one (Table S1C, Data S1Q). We also found evidence that membrane trafficking is more elaborate; there are three genes coding for EXOCYST 70A in *A. thaliana*, two in *C. braunii* (and in the transcriptomes of several Zygnematophyceae), and a single gene in *K. nitens* (as in *Mesostigma viride* and Chlorophyta; Data S1R). With regard to the SNARE complex, we find that the *A. thaliana* NOVEL PLANT SNARE (NPSN) 11/12/13 clade contains two *C. braunii* (and two *Nitella mirabilis*) and a single *K. nitens* (and *M. viride*) protein (Data S1S).

Phytohormones: ETH

For the identification of putative homologs for ETH biosynthesis and signaling genes, BLASTp/tBLASTn searches were carried out against the *C. braunii* gene models and genome assembly using representative *A. thaliana* protein sequences as queries [ACS1 (AT3G61510), ACO1 (AT2G19590), ETR1 (AT1G66340), CTR1 (AT5G03730), EIN2 (AT5G03280), EIN3 (AT3G20770); Table S1J]. Translated sequences of putative ETH biosynthesis/signaling genes from *C. braunii* were then used as queries in reciprocal BLASTp searches to the *A. thaliana* protein database. Multiple ACO homologs were found in the *C. braunii* genome, however, the reciprocal BLASTp search suggests that these homologs are likely to be other oxidases. The other candidate *C. braunii* ETH biosynthesis/signaling protein sequences were manually verified and screened for essential protein domains [ACS (PR00753), ETR/ERS (ETH Binding

Domain), CTR1 (PF14381 and CD13999), EIN3 (PF04873 and C-terminal Signaling Domain), EBF (IPR001810)]. An additional search with BLASTP 2.8.0+ using the representative *A. thaliana* proteins as queries and the putative homologs as the subjects was performed.

Phytohormones: ABA

For the identification of putative homologs for ABA biosynthesis and signaling genes, BLASTn/BLASTp searches were carried out against the *C. braunii* gene models and genome assembly using representative *A. thaliana* genomic/protein sequences as queries (Table S1J). An additional search with BLASTP 2.8.0+ using the representative *A. thaliana* proteins as queries and the putative homologs as the subjects was performed. The obtained *C. braunii* protein sequences were manually verified and screened for essential protein domains [PSY (PF00494), PDS (PF01593), GTG1 (PF12537), SnRK/CPK (PF00069)].

Phytohormones: SL

For the identification of putative homologs for SL biosynthesis and signaling genes, BLASTn/BLASTp searches were carried out against the *C. braunii* gene models and genome assembly using representative *A. thaliana* genomic/protein sequences as queries (Table S1J). An additional search with BLASTP 2.8.0+ using the representative *A. thaliana* proteins as queries and the putative homologs as the subjects was performed. The obtained *C. braunii* protein sequences were manually verified and screened for essential protein domains [CCD (PF03055)].

Phytohormones: Jasmonates (JA), Salicylates (SA), Gibberellins (GA), Brassinosteroids (BR)

For the identification of putative homologs for JA, SA, GA and BR biosynthesis and signaling genes, BLASTn/BLASTp searches were carried out against the *C. braunii* gene models and genome assembly using representative *A. thaliana* genomic/protein sequences as queries (Table S1J). Canonical (land-plant like) signaling pathways for JA, SA, GA and BR have been shown to have arisen in land plants [JA - (Han, 2017); SA - (Wang et al., 2015)], vascular plants [GA - (Gao et al., 2008; Wang et al., 2015)] and seed plants [BR - (Vriet et al., 2015)] respectively. Consistent with these findings, none of the genes encoding steps in the biosynthesis or signaling pathways for GA, JA, SA or BR appear to be present in the *C. braunii* genome (Table S1J). However, JA was found in *C. australis* (Beilby et al., 2015), JA and SA were detected in *K. nitens* (Hori et al., 2014), and GA was detected in *Chara tomentosa*, suggesting a different synthesis than known in land plants as in the case of AUX and ABA (Table 1, Fig. 4).

Phytohormones: AUX transport

For the identification of putative homologs for AUX transporter genes, tBLASTn/BLASTp searches were carried out against the *C. braunii* gene models and genome assembly using representative *A. thaliana* genomic/protein sequences as queries (Table S1J and S11).

Predicted coding sequences of PIN proteins were manually aligned with representative PIN sequences from previously published alignments, PIN sequences from charophyte algae were obtained from the NCBI database. The PIN sequence of *K. nitens* (GAQ81096.1) originated from the complete genome assembly, other algal sequences were obtained from the SRA database (Leinonen et al., 2011) of individual sequencing project by using the BLASTn

algorithm, using the sequence from *K. nitens* as a query. The resulting hits were assembled with CAP3 (Huang and Madan, 1999) and repeatedly BLASTed against respective SRA databases to increase sequence length. Maximum-likelihood phylogenetic analysis was performed in MEGA 7.0 software using amino acid representation of highly conserved N- and C-terminal part of PIN sequence, LG+G+I substitution model and 500 bootstrap replicates (Data S1C, D).

Phytohormones: AUX signaling

For charophyte algae, mRNA sequences were downloaded and protein sequences were predicted with ESTScan v3.0.3 (Iseli et al., 1999) using the *A. thaliana* matrix [-M Arabidopsis_thaliana.smat]. Subsequently all proteins were screened with *hmmsearch* of the HMMer software suite (v3.1b2) for the abundance of the PFAM v30.0 domains: Auxin_resp (PF06507), AUX_IAA (PF02309), B3 (PF02362), F-box (PF00646) and F-box-like (PF12937) using either the gathering threshold [--cut_ga] option or an E-value of 0.1 for the complete sequence [-E 0.1] and an E-value of 0.1 for the domain [--domE 0.1] to account for possible sampling bias and cutoff bias of the curated PFAM model.

The obtained results were used to classify the proteins into possible AUX gene families: ARFs [mandatory domains: Auxin_resp + B3; optional: AUX_IAA], Aux/IAA [mandatory: AUX_IAA - Auxin_resp] and TIR1/AFB [mandatory: F-box or F-box-like]. For the AUX gene familiy TIR1/AFB an additional BLAST search with BLAST+ (v2.5.0) [-matrix BLOSUM45 -evalue 1e-5] using representative *A. thaliana* genes as queries [AT3G62980.1 (TIR1), AT4G03190.1 (AFB1), AT3G26810.1 (AFB2), AT1G12820.1 (AFB3), AT4G24390.2 (AFB4), AT5G49980.1 (AFB5)] and the domain containing proteins as the subjects was performed. Only BLAST hits with a query coverage (alignment length / query length) of at least 50% and a minimal protein identity according to formula (2) of (Rost, 1999) were retained as possible AUX gene family candidates. Maximum-likelihood phylogenetic analysis for each AUX gene family was performed on manual curated multiple sequence alignments obtained via MAFFT (v7.305b) and the E-INS-i algorithm. *IQ-TREE* (Nguyen et al., 2015) v1.5.3 was applied using the standard non-parametric bootstrap option with 1,000 replicates and the best model selected by *IQ-TREE* (Table S1K, Data S1E-G).

Phytohormones: AUX,* in silico *modeling of* C. braunii *LRR FBPs.

Leucine-RichRepeat (LRR)-containing F-Box Proteins (FBPs) from *C. braunii* with sequence similarity to land plant LRR FBPs were *in silico* modeled using “intensive” modeling mode in Protein Homology/analogY Recognition Engine V 2.0 (Phyre2) (Kelley et al., 2015). Various PDB molecule templates (coronatine-insensitive protein 1: Chain B (c3ogmB) and Chain D (c3oglD); transport inhibitor response 1: Chain E (c2p1nE); f-box/lrr-repeat max2 homolog: Chain A (c5hywA), skp2: Chain C (c1fs2C) and Chain K (c1fqvk); and protein toll: Chain A (c4lxrA)) were sele-cted to model *C. braunii* LRR FBPs based on heuristics to maximize confidence, percentage identity and alignment coverage. Structural prediction from regions modeled *ab initio* are highly unreliable. The final models (color-coded by the confidence of the match to the templates overall) were submitted to 3DLigandSite server (Wass et al., 2010) to predict potential binding sites (gray structures cartoon depiction); see Data S1P.

Phytohormones: CK

In order to identify putative CK receptors, BLAST searches were carried out against the *C. braunii* gene models and genome assembly, using PpCHK4 and AHK4 as queries. The detected sequences were run against the Interpro and PFAM databases to detect the domains (histidine kinase and response regulators) which are found in CK receptors. Two sequences were identified containing the domain architecture of CK receptors (CHBRA123g00790 and CHBRA19g00270). In order to identify putative histidine phosphor transfer protein (HPT), a search with the HPT domain (Interpro IPR008207) was conducted and retrieved one sequence (CbHPT1, CHBRA650g00040) (Table S1J). For identification of the response regulators (type-A and type-B) we used the PFAM domains Response_reg (PF00072) and Myb_DNA-binding (PF00249) in an hmmsearch and did not find any gene models. In order to make sure that this result is not due to a missing or fragmentary gene model we also screened the available transcriptome data (transcripts were translated in all possible frames). While two A-type response regulators (RRA) could be detected in the transcriptome (comp31700c0seq1num3, comp64895c0seq1/2 rc num2, Table S1J/S1K, Data S1H), no combination of the two domains and thus no B-type (RRB) could be detected. All sequences harboring Response_reg domains were aligned with the response regulator domains of the *Arabidopsis* response regulators ARR1 and ARR14 (RRB) as well as ARR4 and ARR9 (RRA) and ARR 22 (RRC – not known to be involved in CK signaling) using the muscle implementation of the MEGA 7.0 suite. Using the alignment, a maximum likelihood tree was calculated with the pairwise distances estimated by a JTT model and 100 bootstrap samples. Again, two sequences were determined as RRAs. Of the *Chara* sequences in the RRB clade, again none contained a MYB domain (Data S1H).

Photorespiration

In land plants, the canonical photorespiratory pathway employs 8 enzymes, namely 2PG-phosphatase (PGPase), glycolate oxidase (GOX), glutamate:glyoxylate aminotransferase (GGT), glycine decarboxylase (GDC), serine hydroxymethyltransferase (SHMT), serine/alanine:glyoxylate aminotransferase (SGT), hydroxypyruvate reductase (HPR) and glycerate 3-kinase (GLYK) (Bauwe et al., 2010). Particularly, the glycolate oxidation step, which is performed by GOX in the plant peroxisomes, is catalysed by glycolate dehydrogenase in the mitochondrium of the green algae *C. reinhardtii* (Nakamura et al., 2005) and in the cytosol of cyanobacteria. To analyze the photorespiration in the Charophyte algae *C. braunii*, the protein sequences of enzymes from *A. thaliana* were used to identify homologue proteins in *C. braunii* by a BLASTp similarity search against the Chbra.pep.20151207.orcae database (Table S1M). To verify, if *C. braunii* also possess genes to oxidize glycolate via a glycolate dehydrogenase like Chlorophytes and cyanobacteria do, the polyphyletic proteins from *C. reinhardtii* (ABG36932.1) and *Synechocystis sp.* PCC 6803 (Sll0404 and Slr0806) were used as templates in similarity searches. To verify, if a putative glycolate oxidase prefers the substrate glycolate over lactate, three amino acids in the active site that were shown to be responsible for the substrate preference (Hackenberg et al., 2011) were analyzed. To this end, the putative glycolate oxidase from *C. braunii* and verified glycolate oxidase proteins of the land plants *A. thaliana* and *Spinacia oleracea*, the red alga *Cyanidioschyzon merolae* and characterized L-lactate oxidase proteins from the cyanobacterium *Nostoc* sp. PCC 7120 and the bacterium *Aerococcus viridans* were aligned and the corresponding amino acids in the active sites of the proteins compared.

Retrograde signaling and PAPs

Protein data from the genomes of *C. reinhardtii*, *K. nitens*, *C. braunii*, and *P. patens* was screened for orthologs of the flowering plant-type retrograde signaling pathway or PAPs *via* a reciprocal best BLASTp approach using *A. thaliana* sequenes as query. For GUN1, the BLASTp analyses were repeated using reciprocal pHMMER surveys. To further pinpoint the relation of *Cb*GUN1 to other PPRs, the high similarity *K. nitens* protein GAQ81958.1 was used as a query in BLASTP (2.2.26) search to a database comprising the NCBI nr dataset as of January 2015 supplemented with *K. nitens*, *Pinus taeda* 1.01, and *P. patens* v3.3 Ppav3.3 datasets and 912 hit sequences were retrieved through (http://moss.nibb.ac.jp/cgi-bin/blast-nr-Kfl). Two *C. braunii* proteins Cbr_g9159.t1 (GUN1) and Cbr_g31394.t1, and a *M. polymorpha* protein Mapoly0154s0039.1 were added to this set. From this set, top 500 hits with GAQ81958.1 were retrieved and aligned with mafft version 6.811b and converted to nexus format file through (http://moss.nibb.ac.jp/cgi-bin/selectNalign). The alignment was edited to retain 242 aa (others were excluded; further 47 proteins that showed low conservation in the retained regions were deleted). The nexus file was subjected to http://moss.nibb.ac.jp/cgi-bin/makenjtree to construct a NJ tree based on JTT distance with 1,000 bootstraps using PHYLIP 3.695. Sequences identical within the retained 242 aa sites were treated as a single OTUs and 381 OTUs remained in the final tree. The organism name the sequence originated was recovered using NCBI taxonomydb (ftp://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid/prot.accession2taxid.gz, ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz). The subcellular localization of PAPs was predicted using three online tools (Table S1N).

Transcription factors and transcriptional regulators

Transcription associated proteins (TAPs) comprise transcription factors (TFs, acting in sequence-specific manner, typically by binding to *cis*-regulatory elements) and transcriptional regulators (TRs, acting on chromatin or via protein-protein interaction. We classified all *C. braunii* proteins into 122 families and sub families of TAPs by first screening the proteins for domains and then applying a domain-based rule set to distinguish the TAPs (Lang et al., 2010; Wilhelmsson et al., 2017). We compared this genome-wide classification with genomic protein sets from *Cyanidioschyzon merolae*, *C. reinhardtii*, *Cyanophora paradoxa*, *K. nitens* and several land plants, as well as with transcriptomic data of Charophyta (Timme et al., 2012), *M. polymorpha* and ferns (Table S1Q, S1Z). The phylogenetic tree for the trihelix family (Data S1J) was inferred as mentioned above for the cell division related families.

For the Homeodomain (HD) and basic Helix-Loop-Helix (bHLH) phylogenetic analyses (Table S1O, S1P), the *C. braunii* genome was searched using a BLASTp query that was assembled from the previously characterized bHLH and HD protein sequences (Catarino et al., 2016) in *At*, *A. thaliana*; *Os*, *O. sativa*; *Sm*, *Selaginella moellendorffi*; *Pp*, *P. patens*; *Mp*, *M. polymorpha; Kf*, *K. nitens; Cr*, *C. reinhardtii*; *Ot*, *Ostreococcus tauri*; *Vc*, *Volvox carteri; Cm*, *C. Merolae* with the addition of bHLH proteins sequences from *Cv*, *Coccomyxa subellipsoidea* (previously *Chlorella vulgaris*. The results of the BLASTp search were analyzed manually to ensure the presence of the HD or the bHLH conserved domain using SMART and PFAM. All protein sequences were aligned using MAFFT (Katoh and Standley, 2013) and further manually aligned independently for HD and bHLH. The Maximum likelihood analysis was carried out

using PhyML (Guindon et al., 2010) 3.0, using the JTT amino acid substitution model and a predicted gamma distribution. Branch support was tested using a Shimodaira-Hasegawa-like approximate likelihood ratio test (SH-like aLRT). The generated unrooted trees were visualised using MEGA 6.0.

MADS box sequences were identified using the aforementioned domain-based rule set to distinguish the TAPs (Lang et al., 2010). Phylogenies were calculated with MrBayes (Huelsenbeck and Ronquist, 2001) applying mixed AA model for 50,000,000 generations based on an amino acid alignment of Type I and Type II MADS-domain proteins from a broad set of land plants together with MADS-domain proteins from charophytes. Sequences were aligned with MAFFT (Katoh and Standley, 2013) applying E-INS-i mode. Intron structure was determined by using the transcript sequence as query for BLAST searches against the genome scaffolds. Subsequently, the genomic region that harbors the gene was extracted and aligned to the transcript sequence.

Motor proteins

PFAM domains related to the three classes of motor proteins were retrieved from the whole predicted proteomes of *C. braunii*, *C. reinhardtii*, *P. patens*, and *A. thaliana* using Interproscan (Table S1S). These selected domain signatures not only include the true motors but also domains associated with the tasks the motors have to fulfill in a cell. Since motor proteins are comparably long gene prediction on draft genomes can lead to a slight overestimation of domain numbers. Thus, retrieved predicted gene structures were examined, whether they reside adjacent to another predicted gene encoding for a motor protein part. If the domain structures from known complete proteins conformed with a fusion of two or more adjacent gene models in *C. braunii*, we used this fused gene model for further analysis.

Action potential related ion channels and transport proteins

Ion channels, transporters and pumps predicted to be involved in electrical signaling in plants were identified in the *C. braunii* genome *via* a tBLASTn/BLASTp approach using *A. thaliana* sequences as bait as well as on the basis of PFAM domains. Subsequent BLASTp searches of retrieved sequences against TAIR10 (https://www.arabidopsis.org) and SWISSPROT were employed to identify closest homologs. Finally, sequences were were classified into respective transporter families according to TCDB (Saier et al., 2016) and ARAMEMNON (Schwacke et al., 2003) (Table S1R). When partially split models were found, they were manually annotated with reference to RNA-seq evidence through a genome browser at https://chara.asrc.kanazawa-u.ac.jp/Cbr1/jbrowse/.

LysM-RLKs

The *C. braunii* genome was screened for LysM-RLK genes via tBLASTn using Medicago NFP and Rice CERK1 as bait sequences (Table S1V). Hits with E-value < 10^{-30} were collected and deduplicated. These sequences were aligned using MAFFT (Katoh and Standley, 2013) with LysM-RLKs from embryophytes and *Nitella mirabilis*. Using MEGA 6.0 the best substitution model (JTT+G) was determined and a maximum likelihood tree was inferred using all sites and 100 bootstrap resamplings (Fig. 5C, Data S1L-N).

PPR proteins

Genomic protein sets were scanned for presence of the PFAM domain PPR (http://pfam.xfam.org/family/PF01535) using HMMscan. The number of proteins harboring two or more PPR domains were considered PPR proteins putatively involved in organellar RNA editing (Maier et al., 2008) and are shown in Table S1Y.

ROS-associated genes

21 families belonging to the well-known reactive oxygen species (ROS) gene network were searched using as a first screen the follwing PFAM. PF00141 for Class III Prx (CIII) and Ascorbate Prx (APx and APx-R), PF00199 and PF06628 for catalases (Kat), PF00255 for glutathione Prx (GPx), PF00578 and PF08534 for peroxiredoxin family, PF03098 for dioxygenase (DiOx), PF08022, PF01794, PF08030 and PF08414 for NADPH Oxidase (RBOH) and Ferric reduction oxidase (FRO), PF02777 and PF00080 for superoxide dismutase family (MnSOD, FeSOD, Cu/ZnSOD), PF00462 for Glutaredoxins superfamily, PF01786 for Alternative Oxidase (AOX and PTOX), PF02298 for Blue-copper-binding protein superfamily, PF00210 for ferritin (FER), PF13417 for dehydroascorbate reductase (DHAR), PF07992 and PF02852 for Monodehydroascorbate reductase (MDAR) and Glutathione reductase (GR), PF07992, PF02943 and PF00085 for thioredoxin superfamily and PF01070 Glycolate Oxidases (GOx). *Arabidopsis* sequences belonging to the "ROS gene network" have been used to confirm the *C. braunii* families affiliation.

Only alpha-DiOxygenase (DiOx) and APx-R were not detected in the *C. braunii* assembly. The 19 other families have been found in *C. braunii* with various conservation rates (Table S1X). Among these families, Class III peroxidases (Prx), described as secreted peroxidases, are usually members of a large family. The *C. braunii* genome contained 14 homologous sequences (Table S1X), which is much lower as compared with flowering plants (73 in *A. thaliana*) but higher than in *K. nitens* (3). All the 14 sequences are derived from a single gene in an ancestor of *C. braunii* as they form a presumably monophyletic clade (Data S1O). Before these duplication events only one or a few initial sequences may have existed, implied by the single sequence detected in *Chlorokybus atmophyticus* transcriptome data (Timme et al., 2012) and the low number of three sequences found in *K. nitens*. The CIII Prx protein sequences from *K. nitens* (3 sequences), *C. braunii* (14 sequences), *P. patens* (57 sequences) and *A. thaliana* (73 sequences) were aligned using MAFFT and the tree constructed using Maximum Likelihood implemented in MEGA (Data S1O).

UBQ proteasome system (UPS)

Arabidopsis genes encoding components of the plant Ubiquitin proteasome system (UPS) were manually selected and used as query sequences in a tBLASTn analysis to identify respective orthologous genes in the *C. braunii* genome. Hits with E-values < 10^{-10} were collected and annotated following a reciprocal best BLASTp approach using TAIR10 (Table S1I).

QUANTIFICATION AND STATISTICAL ANALYSES

All details of the applied statistics (*e.g.* for RNAseq-based differential gene expression analysis) are provided alongside the respective analysis in the Methods Details section. For the differential gene expression analysis between antheridia, oogonia, and zygotes, three true

biological replicates were sequenced and used for the statistical analysis (computed using DESeq2). No sequencing points, i.e. samples, were removed during the analysis.

DATA AND SOFTWARE AVAILABILITY

Raw Illumina (DRA004353, DRA006568) and PacBio (DRA006569) genomic sequence data have been deposited in the DDBJ Sequence Read Archive (DRA) at the DNA Data Bank of Japan (DDBJ) under BioProject PRJDB3348. The main scaffolds are available as entries BFEA01000001-BFEA01011654, the accompanying organisms scaffolds as BFBZ01000001-BFBZ01016437. The chloroplast genome is available as AP018555, the mitochondrial as AP018556. Raw Illumina RNA-seq data used for annotation (DRA006080, DRA002641) have been deposited in the DRA at the DDBJ under BioProject PRJDB3228. Raw Illumina RNA-seq data of reproductive stages have been deposited to NCBI SRA (PRJNA445548). The genome and its annotation is available for human curation *via* the ORCAE interface at the URL: http://bioinformatics.psb.ugent.be/orcae/. The data is freely available for browsing as well as for bulk downloads and blast searches. Persons who would like to contribute and edit the data using the web interface will have to request an account by sending an email. Any change made to gene structures will be processed automatically by adding protein domains (running interpro) and best-blast hits. These changes will be shared with the community immediately. 69,969 ABI reads of a cDNA library (minimum length of 100 bp) have been deposited at the DDBJ under the accession numbers LU106825 to LU176793 (Table S1D). Alignments that are the basis for the phylogenetic trees as well as the genome comparison datasets resulting in Fig. 3 have been deposited as Mendeley Datasets (doi:10.17632/9hzzf9m4kh.1).

Supplemental Tables and Files

The supplemental/supporting information is arranged into:

- a PDF containing Figures S1-S7;
- a PDF containing phylogenetic trees and alignments Data S1A-U;
- five Excel spreadsheets containing Tables S1-5 (with indexing of sheets);
- alignments and supporting data for the genome comparisons (related to Fig. 3) in Mendeley (doi:10.17632/9hzzf9m4kh.1).

Supplemental Tables

Table S1, related to STAR methods: details of assembly, annotation and comparative analyses, with index in first sheet.

Table S1A: Genome libraries and accession numbers

Table S1B: Libraries used for assembly

Table S1C: Cell division

Table S1D: EST data deposited in DDBJ

Table S1E: RNA-seq used for annotation

Table S1F: Repeatmasker results

Table S1G: Repetitive elements

Table S1H: Cell wall biosynthesis

Table S1I: UBQ proteasome system

Table S1J: Phytohormones

Table S1K: Auxin signaling and transport

Table S1L: Genome comparison

Table S1M: Photorespiratory pathway

Table S1N: PAP localization prediction

Table S1O: bHLH and HD TFs comparison

Table S1P: bHLH and HD TFs *C. braunii*

Table S1Q: Transcription factors and transcriptional regulators

Table S1R: Ion channels

Table S1S: Motor proteins

Table S1T: Assembled bacterial genomes

Table S1U: Most abundant bacterial genera

Table S1V: LysM RLKs

Table S1W: Reproductive transcriptome

Table S1X: ROS network

Table S1Y: PPR proteins

Table S1Z: Transcription factors and transcriptional regulators gene Ids

Table S1AA: genome and transcriptome datasets used for comparative studies

Table S2, related to Fig. 5/6 and STAR methods: Differential gene expression analyses of rerproductive stages, with index in first sheet.

Table S2A: zygote versus oogonia

Table S2B: oogonia versus antheridia

Table S3, related to Fig. 5/6 and STAR methods: Gene Ontology analyses of the differential expression data in Table S2, with index in first sheet.

Table S3A: zygote versus oogonia up GO enrichment

Table S3B: zygote versus oogonia up genes

Table S3C: zygote versus oogonia down GO enrichment

Table S3D: zygote versus oogonia down genes

Table S3E: oogonia versus antheridia up GO enrichment

Table S3F: oogonia versus antheridia up genes

Table S3G: oogonia versus antheridia down GO enrichment

Table S3H: oogonia versus antheridia down genes

Table S4, related to Fig. 5/6 and STAR methods: *C. braunii* protein coding gene annotation (Gene Ontology, best blast hits, trihelix TFs, expression data, overlap with TE evidence, decontamination).

Table S5, related to STAR methods: Decontamination analyses.

S5A summary

S5B underlying data

Data S1, related to STAR methods: phylogenetic trees and alignments.