



RESEARCH PAPER

Zygnema circumcarinatum UTEX 1559 chloroplast and mitochondrial genomes provide insight into land plant evolution

Lauren M. Orton^{1,*}, Elisabeth Fitzek^{2,3}, Xuehuan Feng⁴, W. Scott Grayburn¹, Jeffrey P. Mower^{5,6}, Kan Liu^{5,7}, Chi Zhang^{5,7}, Melvin R. Duvall¹ and Yanbin Yin^{4,*} 

¹ Biological Sciences, Northern Illinois University, 1425 W. Lincoln Hwy, DeKalb, IL 60115, USA

² Biology/Computational Biology, Bielefeld University, Universitätsstraße 25, D-33615, Bielefeld, Germany

³ Center for Biotechnology-CeBiTec, Universitätsstrasse 27, D-33615 Bielefeld, Germany

⁴ Department of Food Science and Technology, Nebraska Food for Health Center, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

⁵ Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

⁶ Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68583 USA

⁷ School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, NE 68583, USA

* Correspondence: lauren.orton@outlook.com or yyin@unl.edu

Received 30 October 2019; Editorial decision 17 March 2020; Accepted 19 March 2020

Editors: Robert Sharwood, Australian National University, Australia

Abstract

The complete chloroplast and mitochondrial genomes of Charophyta have shed new light on land plant terrestrialization. Here, we report the organellar genomes of the *Zygnema circumcarinatum* strain UTEX 1559, and a comparative genomics investigation of 33 plastomes and 18 mitogenomes of Chlorophyta, Charophyta (including UTEX 1559 and its conspecific relative SAG 698-1a), and Embryophyta. Gene presence/absence was determined across these plastomes and mitogenomes. A comparison between the plastomes of UTEX 1559 (157 548 bp) and SAG 698-1a (165 372 bp) revealed very similar gene contents, but substantial genome rearrangements. Surprisingly, the two plastomes share only 85.69% nucleotide sequence identity. The UTEX 1559 mitogenome size is 215 954 bp, the largest among all sequenced Charophyta. Interestingly, this large mitogenome contains a 50 kb region without homology to any other organellar genomes, which is flanked by two 86 bp direct repeats and contains 15 ORFs. These ORFs have significant homology to proteins from bacteria and plants with functions such as primase, RNA polymerase, and DNA polymerase. We conclude that (i) the previously published SAG 698-1a plastome is probably from a different *Zygnema* species, and (ii) the 50 kb region in the UTEX 1559 mitogenome might be recently acquired as a mobile element.

Keywords: Charophyte green algae, chloroplast, embryophyta, evolution, genomes, land plants, mitochondria, mitogenome, plastome, terrestrialization, *Zygnema*.

Introduction

The interest in Charophycean green algae (CGA) has recently grown due to their close phylogenetic relationship with land plants. In particular, CGA are of vital importance to investigating

and understanding the evolution of cell walls, cell wall biosynthesis, and the terrestrialization of plant life. The CGA include >10 000 species and >65 unique genera (McCourt

et al., 2004; Delwiche and Cooper, 2015). Charophyta are divided into six classes that are phylogenetically closer to Embryophyta (land plants) than to Chlorophyta (green algae). Within Charophyta, there are both the basal branching clades (which include Klebsormidiophyceae, Chlorokybophyceae, and Mesostigmatophyceae; KCM), and the higher branching clades (which include Zygnematophyceae, Charaphyceae, and Coleochaetophyceae; ZCC).

Initially, morphological assessments and molecular data of the CGA suggested that Charales were the closest relatives of the land plant lineage due to their complex body plan that appears analogous to that of typical land plants (Karol *et al.*, 2001). However, the ancestor of the land plant lineage would have probably been an organism with a simpler, less specialized body plan able to adapt, thrive on, and colonize the land (Delwiche and Cooper, 2015). Likewise, this ancestor would have also been able to quickly repurpose and utilize nutrients to fortify its cell walls and provide structure as well as produce compounds to prevent desiccation (Delwiche and Cooper, 2015).

Of particular interest, the class Zygnematophyceae has recently been classified as the closest sister group to the embryophytes (Wodniok *et al.*, 2011; Timme *et al.*, 2012; Wickett *et al.*, 2014). Within Zygnematophyceae, the genus *Zygnema* consists of nearly 100 species of freshwater algae located in temperate to tropical climates (Guiry and Guiry, 2013). *Zygnema circumcarinatum*, a filamentous charophycean alga with elongated cells and multiple star-shaped chloroplasts per cell, is found in stagnant waters as 'large bright green colored, tangled, floating masses' (Guiry and Guiry, 2013). This alga also produces a polysaccharide mucosa which is thought to have been favored in the transition from aquatic to terrestrial by preventing water loss in dry terrestrial environments (Becker and Marin, 2009; Timme and Delwiche, 2010; Herburger *et al.*, 2019).

Like many species of Zygnematophyceae, *Z. circumcarinatum* reproduces through conjugation of (+) and (−) mating types. This differentiation determines which organelle will be passed along to the offspring from the parent mating type (Sekimoto *et al.*, 2006). The shift from strict conjugation to oogamous reproduction better suits terrestrial adaptations; however, the specific origins and selection pressures for the shift to mating types and later oogamous reproduction are still unclear (Geng *et al.*, 2014). Additionally, the strain of focus in this study, UTEX 1559 (mt+), is identified as a spontaneous mutant of the strain Indiana University Culture Collection (IUCC) 42 (now UTEX 42), and consequently UTEX 1559 contains higher numbers and overall size of chloroplasts per cell (Gauch, 1966). As the IUCC has since moved to the UTEX Culture Collection, the UTEX 42 strain is the IUCC 42 equivalent.

These features of CGA, and in particular, *Z. circumcarinatum*, have provided researchers with new avenues to explore the evolution of land plant origins, and the adaptations to terrestrialization through phylogenomic methods. With this recent evidence placing Zygnematales as sister to Embryophyta, the interest in CGA has increased in research seeking to understand how the ancestors of land plants evolved the adaptations necessary for terrestrialization (Wodniok *et al.*, 2011; Timme *et al.*, 2012). We hypothesized that one such avenue

for exploring this evolution is to investigate the patterns of gene retention and/or loss in a pan-genome analysis of chlorophytes, charophytes, and embryophytes.

As of October 2019, when this paper was submitted, only two CGA nuclear draft genomes were available in GenBank, that of *Klebsormidium nitens* and *Chara braunii*. In the last 2 months of 2019, four additional draft nuclear genomes were published (Cheng *et al.*, 2019; Wang *et al.*, 2019): *Spirogloea muscicola*, *Mesotaenium endlicherianum*, *Mesostigma viride*, and *Chlorokybus atmophyticus*, which have significantly improved our understanding of the evolution of land plants by analyzing gene gains/losses of these nuclear genomes. However, the only available genome for *Z. circumcarinatum* is the plastome of strain SAG 698-1a (Turmel *et al.*, 2005). Thus, the goals of this research include to: (i) provide the complete chloroplast (plastome) and mitochondrial genomes (mitogenome) of *Z. circumcarinatum* strain UTEX 1559, of which the mitogenome is the first representative from *Zygnema*; and (ii) explore the gene content across Chlorophyta, Charophyta, and land plant representatives to determine those genes likely to be necessary for terrestrialization and colonization of land by the progenitor to land plants.

Materials and methods

Sequencing and assembly

Cultures of the filamentous alga *Z. circumcarinatum* UTEX 1559 (mt+) were obtained through the University of Texas at Austin's Culture Collection of algae (<https://utex.org/products/utex-1559>), and were incubated in flasks, and grown for 4 weeks on a shaker platform in a growth chamber with a 16 h/8 h light/dark cycle at 28 °C. Algae used in this study were cultured in MBBMA medium, which was prepared by adding arginine to MBBM (Van Etten *et al.*, 1983) at a final concentration of 0.02%. Other growth and harvest conditions were described previously (Fitze *et al.*, 2019). Cultures were harvested using vacuum filtration, stored overnight at −80 °C, and lyophilized prior to being pulverized for DNA extraction. This alga produces extracellular polysaccharide mucilage and that binds to the high molecular weight DNA and prohibits a sufficient extraction with regular DNA extraction kits such as the DNeasy plant kit (Qiagen) or NucleoSpin II (Machery-Nagel). Therefore, a modified cetyltrimethylammonium bromide (CTAB) protocol was used to decrease the risk of contamination by the extracellular polysaccharide mucilage (Porebski *et al.*, 1997). Here, an incubation with proteinase K (20 mg ml^{−1}) at 50 °C for 2 h followed by an incubation with CTAB extraction medium at 65 °C for 10 min was used to break open the cells. An overnight incubation with isopropanol at −20 °C ensured a sufficient DNA extraction. After three washing steps with 70% ethanol the pellet was dried using a SpeedVac system (Savant) and resuspended in sterile water. Extracted DNA samples were pooled and further purified using the DNA clean & concentrator kit (Zymo Research).

DNA extracts were quantified using gel electrophoresis (0.6% agarose gel, 125 mV, 25 min) and PicoGreen dye. A standard curve was generated with dilutions of salmon DNA. Purified DNA was sent to the Roy J. Carver Biotechnology Center, University of Illinois, Urbana-Champaign for sequencing on the Illumina HiSeq 2500 (Illumina, Inc., San Diego, CA, USA). Libraries of 250 nucleotide paired-end reads, mate-pair libraries of varying lengths, 3–5, 5–7, and 8–10 kb, and TruSeq Synthetic Long-Read (TSR) libraries were sequenced. Additional sequencing was generated using the Oxford Nanopore MinION (Oxford Nanopore Technologies, Ltd, Oxford, UK).

Illumina reads received post-sequencing were pooled and quality processed using Prinseq (Schmieder and Edwards, 2011) and FastQC (Andrews, 2010). Initial assembly of reads was done with SPAdes v3.8.1

(Bankevich *et al.*, 2012) and Soapdenovo2 (Luo *et al.*, 2012). Assemblies were compared for quality, and the SPAdes assembly was selected as having a higher quality assembly based on the QUAST web interface assessment tool (Gurevich *et al.*, 2013). Raw reads (TSR reads in particular) and assemblies were then identified using BLASTn (Sayers *et al.*, 2019) to determine organellar origin, and imported into Geneious Pro v 11.0.5 (Kearse *et al.*, 2012) for filtering of chloroplast, mitochondrial, and nuclear read pools based on BLASTn results. In addition, NOVOPlasty (Dierckxsens *et al.*, 2017) was run to assemble the plastome and mitogenome, which was used to verify the Geneious assemblies. MinION reads and contigs were further aligned to the assembled plastome as an additional means of assembly verification.

The MinION nanopore reads were trimmed by using Nanofilt (De Coster *et al.*, 2018) and then assembled by using SMARTdenovo (Ruan, 2015) with default parameters, which has been shown to have exceptional performance according to recent literature (Fournier *et al.*, 2017; Istace *et al.*, 2017; Jayakumar and Sakakibara, 2019). The mitogenome assembly was mainly based on a single long nanopore contig assembled by SMARTdenovo, and corrected by Illumina reads in Geneious.

Annotations of the plastome were completed using NCBI's tBLASTn, BLASTx, and BLASTn, followed by verification in Geneious Pro v 11.0.5 once manual adjustment of the stop codon sites was completed where necessary. Verification was done by using the coding sequence (CDS) annotations from SAG 698-1a (NC_008117) (Turmel *et al.*, 2005) and mapping annotated genes on to the assembled UTEX 1559 plastome to account for gene rearrangements that may have occurred causing such variability between strains. Following that, CDS annotations from *Klebsormidium nitens* were also mapped to the assembled UTEX 1559 plastome to reduce bias from the SAG 698-1a CDS annotations and confirm proper verification of UTEX 1559 annotations.

The mitogenome annotations were completed using a combination of the Dogma annotation webserver (Wyman *et al.*, 2004), Mitofy webserver (Alverson *et al.*, 2010), NCBI's tBLASTn, BLASTx, and BLASTn, and the Glimmer3 program (Delcher *et al.*, 2007). Verification was done by both manually mapping both reads (Illumina and nanopore) and corresponding genes to the final assembly to determine accuracy of sequence assembly and annotations; and by using NCBI's ORFfinder, tBLASTn, BLASTx, and BLASTn. Manual adjustment of the stop codon sites was done where necessary.

The Mauve Aligner plugin (Darling *et al.*, 2010) in Geneious Pro v. 11.05 was used to compare the plastomes of *Z. circumcarinatum* UTEX 1559 and the previously published *Z. circumcarinatum* SAG 698-1a and determine the scale of gene rearrangements that had occurred between these two strains. Default settings were used.

Gene losses/gains

The CDSs in the plastome were identified across 33 species spanning the Chlorophyta, Charophyta, and early diverging Embryophyta. The CDSs were marked as present or absent in each species. The same methods were used for the mitogenomes of 18 species.

Genes were then coded into functional groups based on their Gene Ontology (GO) biological process and molecular function, and a survey was done to determine which of these functional groups were retained/lost through the chlorophyte, charophyte, and embryophyte lineages (Supplementary Tables S1, S2 at JXB online). A cursory search of the draft nuclear genome (in preparation) was performed for any genes determined to have been lost (not retained). This was done to determine if these genes had instead been transferred to the nucleus rather than completely lost from the genome. To do this, a homologous retained gene from a related species was used as a reference, and nuclear contigs of UTEX 1559 were mapped to the reference sequence using Geneious 11.0.5. Positive results were then analyzed using NCBI's tBLASTn.

A heatmap was constructed to show functional group saturation using the R package ggplot2 (Wickham, 2016).

Phylogeny reconstruction

A concatenated translation alignment of seven plastid genes (*rbcL*, *psbA*, *rps19*, *rpoB*, *atpA*, *clpP*, and *petA*) across the 33 species with representative

complete plastomes was created for phylogenetic reconstruction using Geneious v 11.0.5 software. Likewise, a concatenated translation alignment of four mitogenome genes (*cob*, *cox1*, *nad4*, and *nad5*) across the 18 species with representative mitogenomes was also created. Both alignments were analyzed separately using the jMODEL test program to determine the appropriate model for analysis (Darriba *et al.*, 2012). The GTR+I+G model fell within the 88 models at 100% confidence for both data sets and was selected as the best model. RAXML-HPC2 on XSEDE (Stamatakis, 2014), made available through the CIPRES Science Gateway (Miller *et al.*, 2010), was used to produce the best tree in each analysis. The Interactive Tree of Life (iTOL; Letunic and Bork, 2019) was used to visualize the tree, and the R package ggplot2 (Wickham, 2016) was used to create a heatmap showing representation across functional gene groups in the plastome and mitogenome separately.

Mitogenome structure

The mitogenome structure was investigated using Geneious v11.0.5 to visualize possible structural variants through mapping nanopore and Illumina reads to the assembled UTEX 1559 mitogenome using the 'Map to Reference' tool. Identical repeat regions were identified using both the 'Motif Search' tool and general 'Find in Sequence' tool. Repeats were annotated and investigated for circularity and other structural variations as described in Kozik *et al.* (2019).

Results

For UTEX 1559, through Illumina sequencing, we obtained a total of 324 895 194 paired-end reads (2×250 bp) and 48 429 TruSeq synthetic long reads (average length 3522 bp, range 1501–19 443bp). An additional 266 512 reads were generated using the Oxford Nanopore MinION sequencing method (average length 3121 bp, range 200–62 807 bp).

After the organellar genome assembly with these reads, the UTEX 1559 complete plastome (assembled mainly by Geneious and then verified by NOVOPlasty as well as mapping of nanopore contigs/reads; see the Materials and methods) was determined to be 157 548 bp in length with 123 genes (86 CDSs, three rRNAs, 34 tRNAs) (Table 1; Supplementary Fig. S1; Supplementary Table S1). The plastome has 66.3% AT richness. As with the previously published strain SAG 698-1a (165 372 bp long), there are no identifiable inverted repeat regions as are present in a number of clades in Embryophyta (Kolodner and Tewari, 1979). The mitogenome (a single contig was assembled from nanopore reads and then corrected by mapping Illumina reads using Geneious; see the Materials and methods) was determined to be 215 954 bp in length with 67 genes (40 CDSs, two rRNAs, 25 tRNAs) (Table 2; Supplementary Fig. S2; Supplementary Table S2). The mitogenome has an AT richness of 51.4%.

The plastome of *Z. circumcarinatum* UTEX 1559 was compared with that of SAG 698-1a using the Mauve Alignment plugin available in Geneious Pro v11.05, yielding evidence of wide-scale rearrangement of genes (Fig. 1). The average nucleotide identity between UTEX 1559 and SAG 698-1a plastomes is 85.69% (determined by ANI calculator; Rodriguez-R and Konstantinidis, 2016), which is surprisingly low considering that UTEX 1559 and SAG 698-1a are considered to be two strains of the same species.

Table 1. The 33 selected species that have plastomes available in GenBank

Taxonomic groups	Species (cpDNA)	Citation	Accession no.	Length (bp)	Total CDSs
Chlorophyta	<i>Chlamydomonas reinhardtii</i>	Maul et al. (2002)	NC_005353	203 828	69
Chlorophyta	<i>Chlorella heliozoae</i>	Fan et al. (2017)	NC_036805	124 353	78
Chlorophyta	<i>Ostreococcus tauri</i>	Robbens et al. (2007)	NC_008289	71 666	61
CGA-Klebsormidiophyceae	<i>Klebsormidium nitens</i>	Hori et al. (2014)	DF238762	181 482	119
CGA-Klebsormidiophyceae	<i>Klebsormidium flaccidum</i>	Civan et al. (2014)	NC_024167	176 832	107
CGA-Klebsormidiophyceae	<i>Entransia fimbriata</i>	Lemieux et al. (2016)	NC_030313	206 025	119
CGA-Klebsormidiophyceae	<i>Interfilum terricola</i>	Lemieux et al. (2014)	NC_025542	187 843	81
CGA-Chlorokybophyceae	<i>Chlorokybus atmophyticus</i>	Lemieux et al. (2007)	NC_008822	152 254	114
CGA-Mesostigmatophyceae	<i>Mesostigma viride</i>	Lemieux et al. (2000)	NC_002186	118 360	105
CGA-Charophyceae	<i>Nitella hyalina</i>	Pedrola-Monfort et al. (2016) ^a	KX306884	256 512	90
CGA-Charophyceae	<i>Chara braunii</i>	Nishiyama et al. (2018)	AP018555	187 771	96
CGA-Charophyceae	<i>Chara vulgaris</i>	Turmel et al. (2006)	NC_008097	184 933	105
CGA-Coleochaetophyceae	<i>Coleochaete scutata</i>	Lemieux et al. (2016)	NC_030358	107 236	82
CGA-Coleochaetophyceae	<i>Chaetosphaeridium globosum</i>	Turmel et al. (2002a)	NC_004115	131 183	98
CGA-Zygnematophyceae	<i>Roya anglica</i>	Civan et al. (2014)	NC_024168	138 275	93
CGA-Zygnematophyceae	<i>Roya obtusa</i>	Lemieux et al. (2016)	NC_030315	138 272	90
CGA-Zygnematophyceae	<i>Staurostrum punctulatum</i>	Turmel et al. (2005)	NC_008116	157 089	103
CGA-Zygnematophyceae	<i>Cosmarium botrytis</i>	Lemieux et al. (2016)	NC_030357	207 850	89
CGA-Zygnematophyceae	<i>Mesotaenium endlicherianum</i>	Civan et al. (2014)	NC_024169	142 017	87
CGA-Zygnematophyceae	<i>Closterium baillyanum</i>	Lemieux et al. (2016)	NC_030314	201 341	95
CGA-Zygnematophyceae	<i>Cylindrocystis brebissonii</i>	Lemieux et al. (2016)	NC_030359	136 938	90
CGA-Zygnematophyceae	<i>Netrium digitus</i>	Lemieux et al. (2016)	NC_030356	131 804	94
CGA-Zygnematophyceae	<i>Spirogyra maxima</i>	Lemieux et al. (2016)	NC_030355	129 954	91
CGA-Zygnematophyceae	<i>Zygnema circumcarinatum</i> UTEX 1559	This study	MT040697	157 548	86
CGA-Zygnematophyceae	<i>Zygnema circumcarinatum</i> SAG 698-1a	Turmel et al. (2005)	NC_008117	165 372	103
Embryophyta-liverwort	<i>Marchantia paleacea</i>	Shimada and Sugiura (1991)	NC_001319	121 024	89
Embryophyta-liverwort	<i>Marchantia polymorpha</i>	Bowman et al. (2017)	NC_037507	120 304	90
Embryophyta-hornwort	<i>Anthoceros formosae</i>	Kugita et al. (2003)	NC_004543	161 162	90
Embryophyta-hornwort	<i>Leiosporoceros dussii</i>	Villarreal et al. (2018)	NC_039750	155 956	86
Embryophyta-moss	<i>Physcomitrella patens</i>	Sugiura et al. (2003)	NC_005087	122 890	85
Embryophyta-moss	<i>Sphagnum palustre</i>	Shaw et al. (2016)	NC_030198	140 040	84
Embryophyta-moss	<i>Takakia lepidozoides</i>	Sadamitsu et al. (2015) ^a	NC_028738	149 016	61
Embryophyta-angiosperm	<i>Arabidopsis thaliana</i>	Sato et al. (1999)	NC_000932	154 478	85

^a Indicates a direct submission to GenBank, no accompanying manuscript; or unpublished.

Table 2. The 18 selected species that have mitogenomes available in GenBank

Taxonomic groups	Species (mtDNA)	Citation	Accession No.	Length (bp)	Total CDSs
Chlorophyta	<i>Chlamydomonas reinhardtii</i>	Vahrenholz et al. (1993)	NC_001638	15 758	8
Chlorophyta	<i>Chlorella heliozoae</i>	Fan et al. (2017)	KY629615	62 477	32
CGA-Klebsormidiophyceae	<i>Klebsormidium nitens</i>	Hori et al. (2014)	DF238763	106 468	35
CGA-Klebsormidiophyceae	<i>Entransia fimbriata</i>	Turmel et al. (2013)	NC_022861	61 645	35
CGA-Chlorokybophyceae	<i>Chlorokybus atmophyticus</i>	Turmel et al. (2007)	NC_009630	201 763	58
CGA-Mesostigmatophyceae	<i>Mesostigma viride</i>	Turmel et al. (2002b)	NC_008240	42 424	41
CGA-Coleochaetophyceae	<i>Chaetosphaeridium globosum</i>	Turmel et al. (2002a)	NC_004118	56 574	46
CGA-Charophyceae	<i>Nitella hyalina</i>	Lazaro-Gimeno et al. (2011) ^a	NC_017598	80 193	48
CGA-Charophyceae	<i>Chara vulgaris</i>	Turmel et al. (2003)	NC_005255	67 737	46
CGA-Zygnematophyceae	<i>Closterium baillyanum</i>	Turmel et al. (2013)	NC_022860	152 089	49
CGA-Zygnematophyceae	<i>Roya obtusa</i>	Turmel et al. (2013)	NC_022863	69 465	41
CGA-Zygnematophyceae	<i>Zygnema circumcarinatum</i> UTEX 1559	This study	MT040698	215 954	40
Embryophyta-liverwort	<i>Marchantia paleacea</i>	Oda et al. (1992)	NC_001660	186 609	76
Embryophyta-liverwort	<i>Marchantia polymorpha</i>	Bowman et al. (2017)	NC_037508	186 196	65
Embryophyta-hornwort	<i>Leiosporoceros dussii</i>	Villarreal et al. (2018)	NC_039751	212 153	23
Embryophyta-moss	<i>Physcomitrella patens</i>	Terasawa et al. (2007)	NC_007945	105 340	42
Embryophyta-moss	<i>Sphagnum palustre</i>	Liu et al. (2014)	NC_024521	141 276	40
Embryophyta-angiosperm	<i>Arabidopsis thaliana</i>	Unseld et al. (1997)	NC_037304	367 808	33

^a Indicates a direct submission to GenBank, no accompanying manuscript; or unpublished

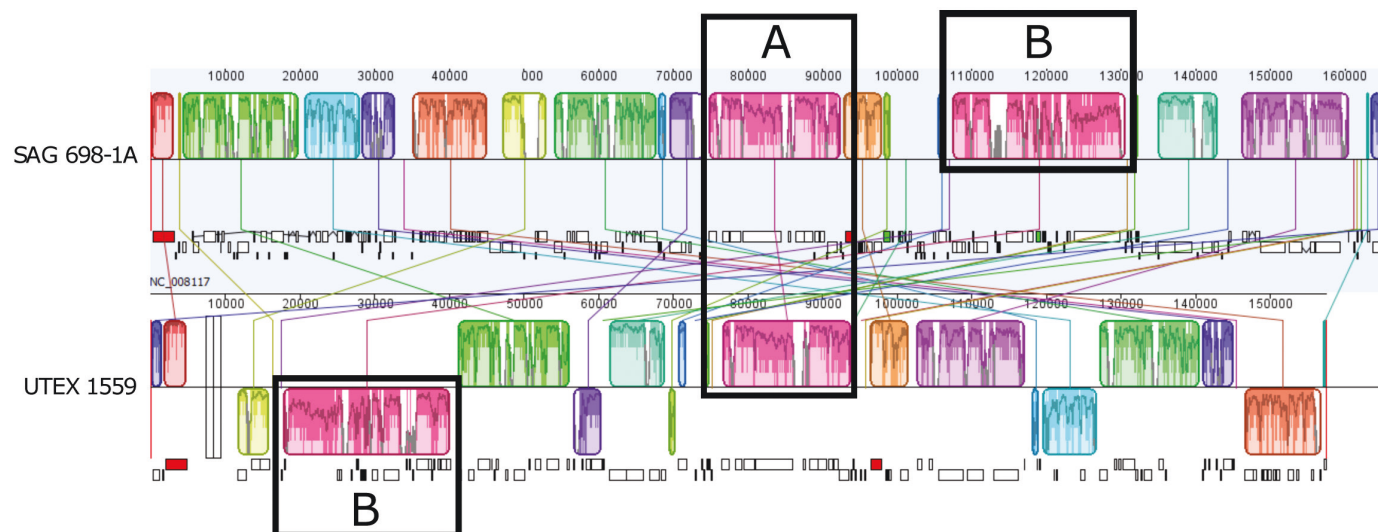


Fig. 1. The Mauve alignment of two *Z. circumcarinatum* strains, SAG 698-1a (top) and UTEX 1559 (bottom). Substantial rearrangements and differing plastome lengths are shown, indicating that these strains may not be conspecific. A box shows the region containing the *ccsA* (cytochrome *c* biogenesis) gene and the *trnI* gene, which are the very few well-preserved regions between the two plastomes. B box shows a large region falling between the *ndhJ* [NAD(P)H-quinone oxidoreductase subunit J] gene and the tRNA gene, *trnL* (*uag*), which is the largest region (100 kbp) that shares conserved gene orders within the region but rearranges in locations very dramatically relative to other regions.

Gene losses/gains

The CDSs in the plastomes were identified across 33 species spanning the Chlorophyta (three species), Charophyta (22 species), and selected Embryophyta (eight species). The sizes of the plastomes across the represented species ranged from 71 666 bp to 256 512 bp (Table 1) and the numbers of CDSs ranged between 61 and 119. Between strains UTEX 1559 and SAG 698-1a there is considerable variation in the presence/absence of certain CDSs. Specifically, SAG 698-1a (NC_008117) has 103 annotated CDSs, and among them 90 are functionally annotated (i.e. with a gene name). All these 90 CDSs have significant BLAST hits (E-value <1e-12) in UTEX 1559. In 63 of the 90 annotated CDSs, the sequence identities between the two strains are >90%. Inspecting the TFASTY (a more sensitive tool than tBLASTn, which also considers frameshifts and stop codons) sequence alignments, Pearson *et al.* (1997) found that the *rpl36* protein of SAG 698-1a actually matched a pseudogene in UTEX 1559 (Supplementary Table S3). However, the *rpl36* gene might have been misannotated in SAG 698-1a, as it encodes a much longer protein (75 amino acids) than *rpl36* proteins in other algae (37 amino acids). Inspecting the alignment in Supplementary Table S3, it was found that there is a methionine at position 40 in UTEX 1559, which corresponds to the start codon in other algal *rps36* genes. Therefore, *rps36* is likely to be present and functional in UTEX 1559. There are also 13 additional CDSs categorized as hypothetical ORFs in SAG 698-1a (ZyciCp033, ZyciCp036, ZyciCp066, ZyciCp068, ZyciCp069, ZyciCp070, ZyciCp071, ZyciCp072, ZyciCp073, ZyciCp074, ZyciCp086, ZyciCp090, and ZyciCp095). Only ZyciCp066 and ZyciCp07 have some moderate homology (1e-5 >E-value >1e-10) in the UTEX 1559 plastome.

The CDS presence/absence was further analyzed across all the studied plastomes. The chlorophyte CDSs do not contain

any dehydrogenase (*ndh* family) genes, while the majority of charophyte and embryophyte CDSs retain all or most dehydrogenase genes (Supplementary Table S1). Most chlorophytes and embryophytes have *petN*, encoding the cytochrome *b₆f* complex subunit 8, which functions as an electron transporter in PSII (UniProtKB-P0C1D4). However, in Charophyta, this gene seems to be missing in *Interfilum* (Supplementary Table S1).

PSI genes, *psaI* and *psaM*, are absent in *Chlamydomonas* (Supplementary Table S1). These two CDSs are present in the majority of the represented Charophyta and Embryophyta.

Between the chlorophyte, charophyte, and embryophyte lineages represented in this study, 49 chloroplast CDSs were retained. However, in examining the genes retained across only the charophytes and embryophytes, 52 CDSs were retained. These CDSs correspond to nine functional groups related to: transport, ribosomal proteins, synthesis, protein processes, photosynthesis, rRNAs, tRNAs, unknown genes (unknown function; excluding hypothetical or predicted ORFs), and 'other' (miscellaneous genes not belonging to any of the previously mentioned functional groups, yet not representative enough to be placed in their own category). These functional group classifications were based on GO terms for molecular function and biological processes (Fig. 2; Supplementary Table S1).

Additionally, a seven gene plastid phylogeny was constructed to detail the relationships between clades. Genes used to produce the phylogeny include: *rbcL*, *psbA*, *rps19*, *rpoB*, *atpA*, *clpP*, and *petA*. These seven genes were selected based on the following criteria: they were present in all species included in the analysis, varied in function and levels of positive selection, and were substantial in size and conservation across the data set. Zygnematophyceae is recovered as sister to Embryophyta (Fig. 2). Note that *Arabidopsis*, the only flowering plant included in this analysis, is clustered with some hornworts and Bryophyta but has a long branch.

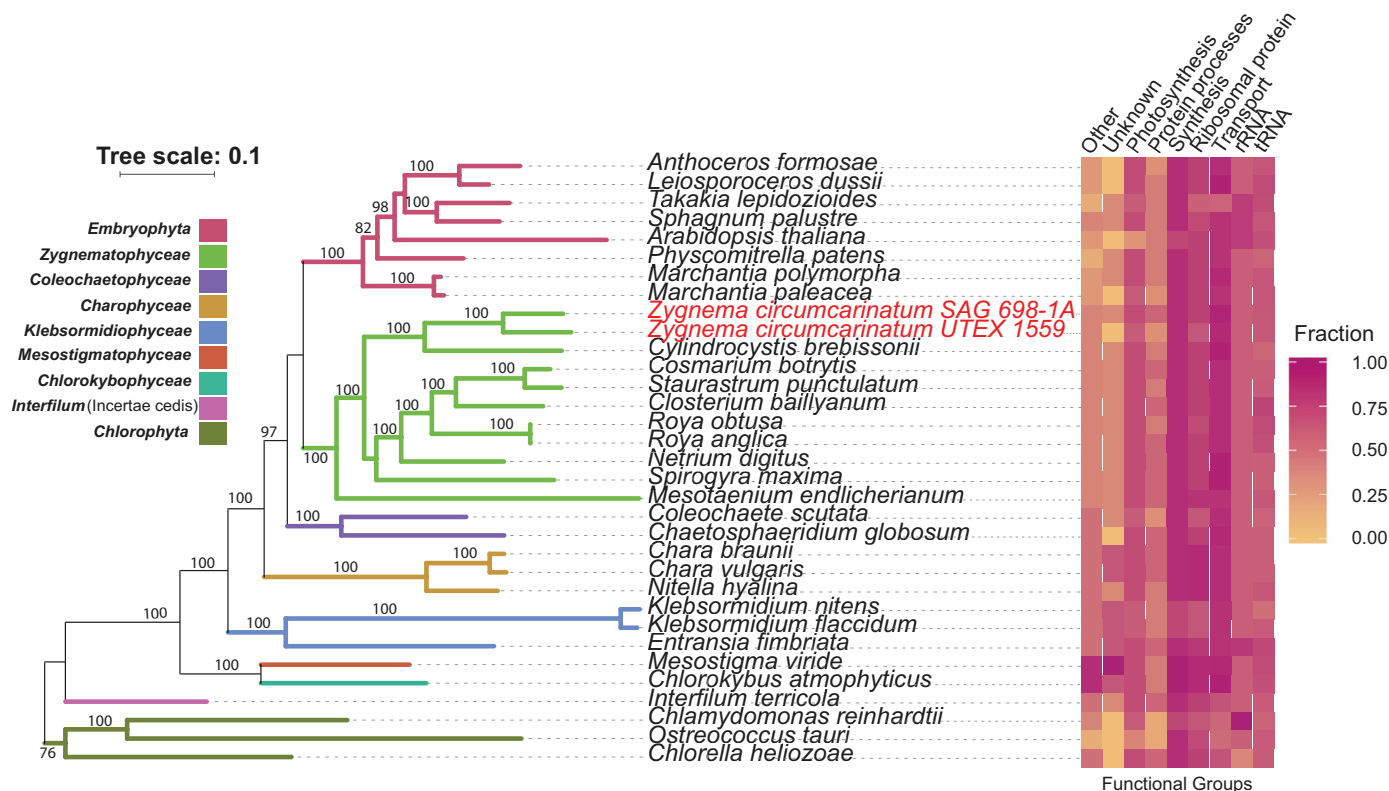


Fig. 2. A seven plastid gene phylogeny (see the Materials and methods for additional details) with a heatmap representing functional gene groups based on Gene Ontology terms. Fraction=fraction representation of total genes in each functional group. For example, UTEX 1559 has a fraction of 0.75 for the Transport function group, which means that 75% of transport genes present in all the studied plastomes exist in UTEX 1559.

Within the mitochondria, a great deal of variability is seen in Charophyta, with genome sizes ranging from 42 424 bp to 215 954 bp. UTEX 1559 now has the largest mitogenome (215 954 bp) among all published Charophyta. Across the mitogenomes of 18 representative species, 87 genes (46 CDSs, five rRNAs, 36 tRNAs) were catalogued and analyzed (Supplementary Table S2). As there is not a mitogenome from *Z. circumcarinatum* SAG 698-1a, there were no comparisons to be made between UTEX 1559 and SAG 698-1a. Instead, it was noted that across the chlorophyte, charophyte, and embryophyte species represented in this study, four CDSs were shared (*nad4*, *nad5*, *cob*, and *cox1*) (Supplementary Table S2). When the comparison is narrowed to include only charophyte and embryophyte representatives, it was determined that 28 CDSs were shared, a substantial increase. As compared across the entire data set, the two genes not retained in UTEX 1559 were: *rpl14* and *rps8* (Supplementary Table S2).

A four gene mitogenome phylogeny was constructed using the genes *cob*, *cox1*, *nad4*, and *nad5* with identical methods to the plastome phylogeny. These four genes were selected based on the following criteria: they were present in all species included in the analysis, and were substantial in size and conservation across the data set. Widely accepted monophyletic groups are not recovered in the mitogenome phylogeny (Fig. 3); for example, the two Chlorophyta species were not clustered, nor were the three Zygnematophyceae including UTEX 1559. Interestingly, the two Charophyceae species are now closer to land plants than Zygnematophyceae. All these are probably as a result of the broad representation of species included in this

study, compounded by the few available genes present across the entire data set.

All the 87 genes of the 18 species correspond to five functional groups (Fig. 3). These functional groups, in turn, correspond to genes in categories related to GO terms molecular functions and biological processes: transport, ribosomal proteins, rRNAs, tRNAs, and other (miscellaneous genes unrelated to either transport or ribosomal proteins, yet not representative enough to be placed in their own category).

Discussion

Zygnematophyceae have been recently classified as the closest sister group to Embryophyta (Wodniok et al., 2011; Timme et al., 2012; Wickett et al., 2014), and the Zygnematophyceae representative *Z. circumcarinatum* shares a number of similarities with its land plant relatives in regard to gene content (Fitzek et al., 2019). In this study, our focus was the organellar genomes of *Z. circumcarinatum* UTEX 1559.

Plastomes of UTEX 1559 and SAG 698-1a are too different to belong to the same species

For *Z. circumcarinatum*, there is one plastome from the strain SAG 698-1a in GenBank (NC_008117) (Turmel et al., 2005). Our UTEX 1559 strain and the previously sequenced SAG 698-1a strain have plastomes of different sizes: 157 548 bp and 165 372 bp, respectively (Table 1). The global sequence identity between them is surprisingly low at 85.69%. In bacteria, it is

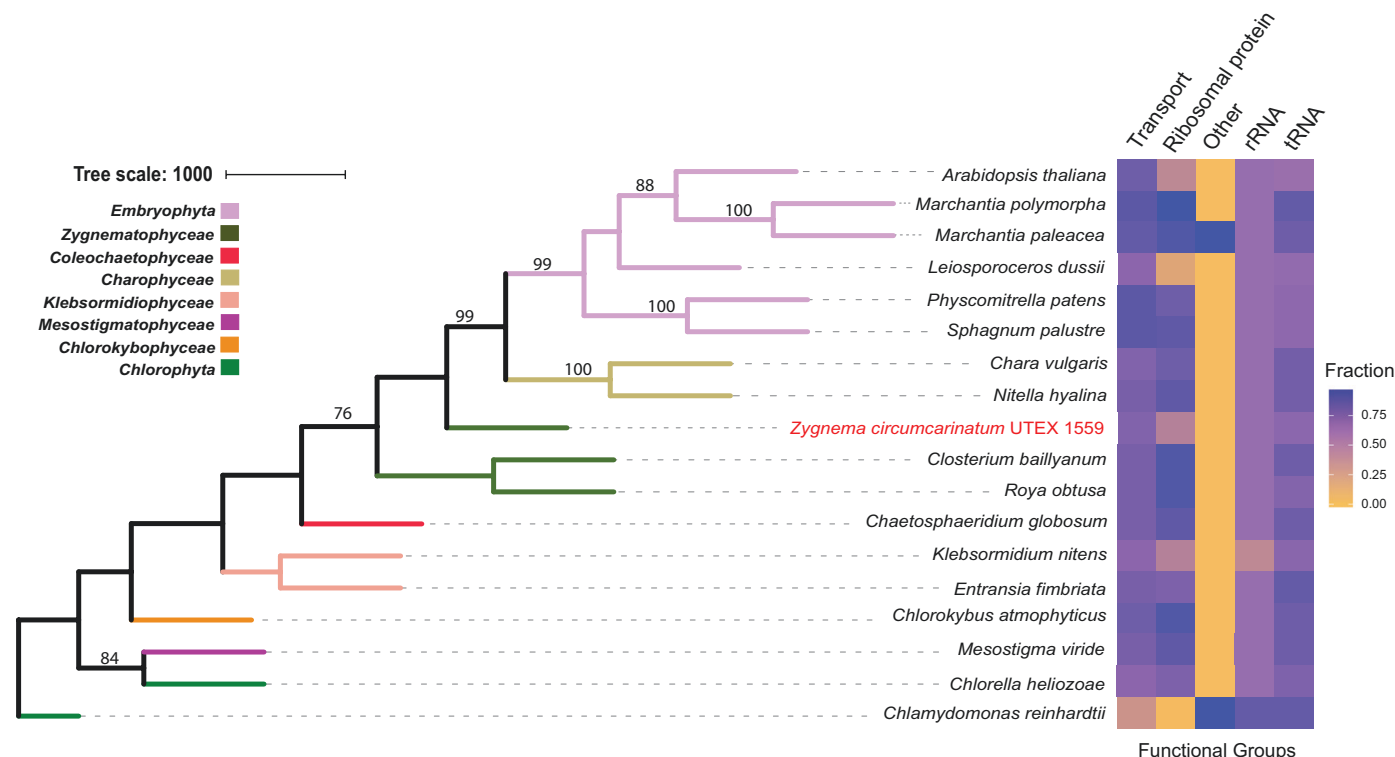


Fig. 3. A four mitogene phylogeny (see the Materials and methods for additional details) with a heatmap representing functional gene groups based on Gene Ontology terms. Fraction=fraction representation of total genes in each functional group. For example, UTEX 1559 has a fraction of 0.75 for the Transport function group, which means that 75% of transport genes present in all the studied mitogenomes exist in UTEX 1559.

common to use an average nucleotide identity <95% between two genomes as an indicator of two species (Konstantinidis and Tiedje, 2005). Although we are not aware of reports using plastome sequence identity to define plant and algal species, the low nucleotide identity between UTEX 1559 and SAG 698-1a suggests that they do not belong to the same species.

These differences raised a question: are UTEX 1559 and SAG 698-1a really of the same species? In looking at the origins of each strain, UTEX 1559 was obtained by (Gauch, 1966) through the 'spontaneous mutation of IUCC 42 and contains increased size and numbers of chloroplasts per cell' (<https://utex.org/products/utex-1559>). In contrast, SAG 698-1a was collected from a freshwater ditch in Poselteich, Bohemia, Czech Republic in 1929 (https://sagdb.uni-goettingen.de/detailedList.php?str_number=698-1a). The SAG 698-1a culture was isolated by V. Czurda and is considered a relative of UTEX 42 (formerly IUCC 42, also isolated by V. Czurda in 1929).

It is unknown from 90 years ago how many *Z. circumcarinatum* strains were isolated by V. Czurda, and how these strains/species are closely related. However, the online resources available seem to indicate that UTEX 42 and SAG 698-1a are relatives. Gauch (1966) observed that UTEX 1559 had doubled cell size and chloroplast number compared with UTEX 42, and inferred that UTEX 1559 is a diploid mutant of the haploid UTEX 42. This polyploidy mutation that occurred in the UTEX 1559 strain might be one reason for the scale and magnitude of differences seen between the two assumed relative strains of *Z. circumcarinatum*. However, our plastome data shown here and the comparison of the genomic DNA reads/contigs

(in preparation) of UTEX 1559 and SAG 698-1a suggest that these two strains are very different. Additionally, we have also sequenced the mating type minus (–) strain SAG 698-1b, and found that its plastome is 100% identical to that of UTEX 1559. Furthermore, we did not detect conjugation between SAG 698-1b and SAG 698-1a using the protocol that was successful in Miller and Hoshaw (1974). Interestingly, the study by Stancheva *et al.* (2012) found that the *rbcL* gene sequence of UTEX 42 (wild type of UTEX 1559) was very different from the same gene encoded in the SAG 698-1a plastome (NC_008117) and postulated that 'the published chloroplast genome of strain SAG 698-1a is not that of *Z. circumcarinatum*, but rather some other species of *Zygnema*'. We found that the *rbcL* gene sequence is identical in UTEX 42 and UTEX 1559. Therefore, our finding using the whole-plastome comparison agrees with the speculation by Stancheva *et al.* (2012), and supports that the published 698-1a plastome (NC_008117) is probably from a species other than *Z. circumcarinatum*.

The results of the Mauve alignment reinforce the differences seen between the two *Z. circumcarinatum* strains, and suggest that a number of arrangements occurred between them. There is a solid anchor point of sequences between the CDS *ccsA* (cytochrome *c* biogenesis) and the tRNA *trnI* (*cau*). This anchor region is well preserved in both the UTEX 1559 and SAG 698-1a strains, as shown in the Mauve alignment (Fig. 1, box A). However, there are numerous regions in which the sequence has been shifted up- or downstream by large margins. One such shift of sequence is an ~100 000 bp shift upstream in UTEX 1559 (~80 000 base shift downstream in SAG 698-1a) from its position in SAG 698-1a, and this region

falls between the tRNA gene *trnL* (*uag*) and the CDS *ndhJ* [NAD(P)H-quinone oxidoreductase subunit J] (Fig. 1, box B).

Additional research into plastomes of charophyte algae has also shown large structural changes, including the loss of inverted repeat regions and uptake of foreign genes from phages and viruses (Lemieux *et al.*, 2016). Therefore, the differences observed between SAG 698-1a and UTEX 1559 are quite substantial and indicate that these may not be conspecific to *Z. circumcarinatum*. Further taxonomic review may be necessary to determine if SAG 698-1a was correctly identified in its initial collection by Czurda in 1929. Additionally, nuclear genome comparisons between both strains may also determine conspecificity.

Gene losses/gains in the plastome

The search of SAG 698-1a CDSs in UTEX 1559 found that all annotated CDSs are present in the UTEX 1559 plastome. However, many of these annotated CDSs have <90% amino acid sequence identities between the two plastomes, and most of the unannotated SAG 698-1a CDSs are missing in UTEX 1559, consistent with the fact that the SAG 698-1a plastome is 7824 bp larger.

Overall, the majority of functions of the genes analyzed in the plastome relate to transport and biosynthesis during photosynthesis. These functions are largely retained through the lineages ranging from Chlorophyta, to Charophyta, and into Embryophyta. This speaks to the obvious importance of the photosynthetic pathway and its functions that spanned aquatic to terrestrial life. However, the absence of the *ndh* (dehydrogenase) functional family of genes in chlorophyte representatives of this study is well documented (Peredo *et al.*, 2013; Lin *et al.*, 2017). The *ndh* genes encode the NAD(P)H dehydrogenase complex which allows transfer of electrons between the electron acceptors NADH and plastoquinone, and prevents overaccumulation of electrons in the thylakoid space.

The literature also suggests a precedent for the loss of genes in the *ndh* functional family in parasitic plants or plant species that have formed symbiotic relationships with bacteria or fungi (Cameron *et al.*, 2006; Delaux *et al.*, 2015; Kim *et al.*, 2015). Furthermore, (de Vries *et al.*, 2018b) explored the possibility that mutualistic relationships exist between many microorganisms and charophyte algae, and noted that beneficial bacteria (nitrogen fixers, etc.) were co-habiting environments with algae present in their analyses. Likewise, Knack *et al.* (2015) investigated the role of mutualistic relationships between streptophyte algae and microbes as it relates to terrestrialization in the land plant lineage. It is possible that mutualism may also have played a role in the gains or losses of certain genes in Charophyta.

Likewise, *matK*, a maturase gene found in the Embryophyta, is absent in the Chlorophytes, as well as in seven of the 19 represented Charophytes (*Interfilum*, *Mesostigma*, *Chlorokybus*, *Entransia*, *Coleochaete*, *Nitella*, and *Cylindrocapsa*). Interestingly, although we did not find *matK* hits in the *Nitella hyalina* plastome, *Nitella opaca* has one (AAO39160) (Hausner *et al.*, 2006). A tBLASTn search using *matK* of *Chaetosphaeridium* as the query (NP_683782.1) found at least two *matK* hits in

the mitogenome of *N. hyalina* but not in its plastome. Such mitogenome hits of *matK* were also observed in *Chara vulgaris*, which also has *matK* in its plastome. This suggests that gene transfers might have happened between the mitogenome and plastomes at least in some algal organisms. The product of the *matK* gene functions in the processing of precursor RNA through splicing group II introns in the chloroplast (UniProtKB-P56784). Phylogenetically, *matK* has been identified as a rapidly evolving gene due to high levels of nucleotide substitutions (Barthet *et al.*, 2015). This level of substitution has the potential to create pseudogenes or cause complete loss of genes (Wickett *et al.*, 2011; Villarreal *et al.*, 2013). It is interesting to note that the transitional nature of Charophyta is evident in the representation of *matK* in the species included in this study, with eight of the 19 Charophyte species lacking the *matK* gene, which is prevalent in the land plant lineage.

When comparing only the charophyte and embryophyte lineages, 52 CDSs were retained, and these retained genes in the chloroplast correspond to functional groups involving ATP synthesis, transport of molecules, ribosomal proteins, and photosystem-related proteins. The ancestor of the land plant lineage would have probably needed aspects of photoprotection to avoid overexposure in a terrestrial environment (Pierangelini *et al.*, 2017). This ancestor would more than likely have been able to efficiently protect itself against overexposure, or, conversely, it would have been able to effectively utilize the additional exposure in a terrestrial environment to increase photosynthetic efficiency. Therefore, the retention of many genes related to transport of molecules, and in particular electrons (*ndh* and *nad* complex genes, for example), seems fitting given the need for early land plants to adapt to the increase in photoexposure found in terrestrial environments.

In all, it comes as no surprise that the charophyte lineage would share many similar functional groups with the embryophytes which have since evolved and adapted to better suit terrestrial survival. Additionally, RNA sequencing studies conducted on osmotic stress found that PYL in the abscisic acid (ABA) signaling pathway was present in both UTEX 1559 and representative embryophytes (de Vries *et al.*, 2018a; Fitzek *et al.*, 2019). Early pathway intermediates for ABA are derived within the chloroplast, and it is likely that the functional groups analyzed in this study play a role in the transport, synthesis, and transcription/translation of inputs for this pathway, further expanding upon the importance of examining organellar genomes and their role in the adaptations to terrestrial environments. Interestingly, plastid-nucleus communication also plays an important role in land plant stress identification and response, and recent research has identified this trait as having been an important step in the colonization of land by the algal progenitor of the land plant lineage (de Vries *et al.*, 2018a; Zhao *et al.*, 2019).

Gene losses/gains in the mitogenome

Literature focusing on mitogenomes has extensively documented (Newton, 1988; Petersen *et al.*, 2006, 2017; Cotton *et al.*, 2015; Shearman *et al.*, 2016) that mitogenomes in Embryophyta are extremely variable in terms of the genome

composition, size, and structure, with some studies even suggesting that they can exist in multiple chromosomes. Likewise, RNA editing potentially leads to inconsistencies in mitogenomes which may result in higher substitution rates, and overall variability/flexibility in their genomes (Petersen *et al.*, 2006; Cuenca *et al.*, 2010). This variability is also echoed in Charophyta with genome sizes ranging from 42 424 bp to 215 954 bp (Table 2). In examining the composition of the mitogenomes represented in this study, it was discovered that very few genes were retained across all three lineages (chlorophyte, charophyte, and embryophyte), implying that there were few genes of Chlorophyta that were retained across the lineages. Those genes that were retained across all three lineages pertained to transport (*nad4* and *nad5*) as well as those related to the cytochrome complex (*cob* and *cox1*) and RNAs (*rrnS*, *rrnL*, *trnM*, *trnQ*, and *trnW*).

The genes not retained in the UTEX 1559 mitogenome were *rpl14* and *rps8*. However, they are present in the UTEX 1559 plastome (Supplementary Table S1). These genes pertained to ribosomal proteins. A number of studies into land plant mitogenomes have described not only their variability, but also their ability to ‘offload’ or ‘take up’ genes from other organelles as well as other organisms through horizontal gene transfer (Cuenca *et al.*, 2012). The loss of genes for ribosomal proteins is a common and recurrent theme throughout the evolution of plant and algal mitochondria (Adams *et al.*, 2002; Mower *et al.*, 2012; Turmel *et al.*, 2013; Petersen *et al.*, 2017). The mitogenome of *Zostera* (Alismatales) has lost nearly all genes encoding ribosomal proteins, and a few were found to have been transferred to the nucleus (Petersen *et al.*, 2017).

Indeed, genes missing in UTEX 1559 organellar genomes might have been transferred into its nuclear genome. Although a tBLASTn search has found homologs in the UTEX 1559 draft contigs for several genes missing in the plastome and mitogenome, the low quality of this draft nuclear genome assembly does not allow us to make confident conclusions at this moment. This search will need to be revisited once our draft assembly of the nuclear genome (in preparation) is finalized.

The gene *sdh4* was also among those genes retained in the UTEX 1559 mitogenome. This succinate dehydrogenase subunit 4 (UniProtKB-Q941A0) is involved in the tricarboxylic acid cycle (carbohydrate metabolism). Although both *sdh3* and *sdh4* are retained in the UTEX 1559 mitogenome, a common trend is also noted in many land plant phylogenies with *sdh3* present and *sdh4* lost (Petersen *et al.*, 2017) or pseudogenized (Mower *et al.*, 2012), as is seen in *Arabidopsis* (NC_037304).

It comes as no surprise that the mitochondrion has a narrowed scope of functional groups given that its main role is to produce ATP via the process of cellular respiration. The majority of genes analyzed in this study pertained to electron transport or ribosomal proteins (Fig. 3).

Mitogenome structure

Mitogenomes are typically displayed as a single ‘master circle’ structure. However, recent research has shed light on mitogenomes existing in numerous structural variants or ‘isoforms’ (Oldenburg and Bendich, 1998; Shearman *et al.*,

2016; Kozik *et al.*, 2019). Many mitogenomes maintain excessively repetitive sequences, making assembly difficult through conventional bioinformatic means, thus obscuring mitogenome structure further (Shearman *et al.*, 2016). These repeats are regions in which recombination can occur, leading to many different isoforms as well (Kozik *et al.*, 2019). In our assembly, we observed that the UTEX 1559 mitogenome assembles into a circular form (Supplementary Fig. S2), which was strongly supported by the assembly of long MinION nanopore reads. However, there is an atypical sequence imbedded within the ‘master circle’ (bases 72 683–122 613; ~50 kb) between 86 bp flanking repeats of the UTEX 1559 mitogenome. This region is unique in that, during our initial assessment, we found that it does not have any significant BLAST hits in published organellar genomes, except for one quality, identifiable, nucleotide sequence: *trnM*-CAT. This created the illusion of a gene desert bearing extremely low gene density between these 86 bp flanking repeats, contributing to an overall decrease in gene density as compared with the remaining Charophyte species in the mitogenome analysis (Fig. 4).

Nanopore long reads were continuously sequenced through this region (no ambiguity among the reads), indicating that it is a persistent feature in the mitogenome of UTEX 1559, and not an assembly error. Illumina read coverage of this region is also consistent and maintains a high depth of reads. It is generally accepted that a driving force for terrestrialization was, in fact, the horizontal transfer of necessary gene families from mutualistic microorganisms to a land plant progenitor (Yue *et al.*, 2012). The unusually large size of this region seems to indicate that it would be an unlikely candidate sequence to have been horizontally transferred from a single donor organism and remain so clearly intact. We also investigated recombination as a possible means of introduction for this sequence which is flanked by large repeat regions.

Interestingly, after an exhaustive search using amino acid sequences queried in tBLASTn, BLASTn, BLASTx, BLASTp, and FASTY (Pearson *et al.*, 1997), this gene desert of ~50 kb returned a number of hits for gene sequences. These hits included: primase, *rsmA*, DNA-dependent RNA polymerase, and DNA polymerase (Fig. 5; Supplementary Table S4). Subsequently, we explored the possibility that this region is a mobile element in the mitogenome, as there are a number of documented cases of nuclear transfer of DNA into the mitochondria by way of transposable elements (TEs) (Knoop *et al.*, 1996; Alverson *et al.*, 2010; Navarro, 2017). These TEs can then contribute to gene deserts (Lee *et al.*, 2015). Using the online Dfam database (Hubley *et al.*, 2016), our sequence was queried for matches with TEs, and returned no hits. Next, we used NCBI’s ORFfinder to explore ORFs in this ~50 kb region. The resulting ORFs were then queried against the NCBI protein database using BLASTp, and quality hits were further investigated (Fig. 5; Supplementary Table S4). The presence of the DNA and RNA polymerases, as well as a DNA primase provides strong evidence for classifying this region as a mobile element in the UTEX 1559 mitogenome. Additionally, it is likely that this mobile element may have arisen rather recently, as there is evidence to suggest that large mutations in organellar genomes are often cleared from the genome quickly

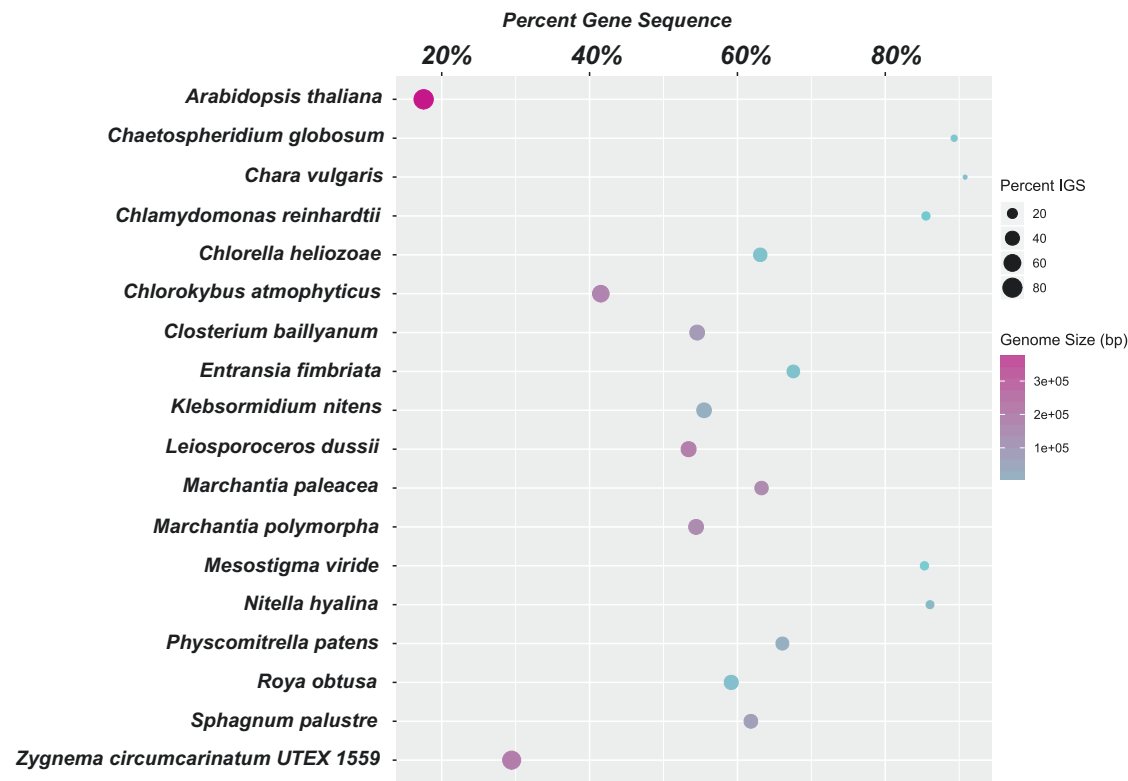


Fig. 4. Gene density graph of the mitogenomes included in this study. x-axis, percentage of gene sequence in each mitogenome; y-axis, species. The size of the circle represents the percentage of the intergenic spacer (IGS). The color of the circle is a representation of the total size of the mitogenome. All species except *A. thaliana* and *Z. circumcarinatum* UTEX 1559 maintain gene densities of $\geq 40\%$. UTEX 1559 has a gene density of 29.4% excluding the genes present in the mobile element. When the additional genes present in the mobile element are considered, UTEX 1559 still maintains $<40\%$ gene density at 39.4%.

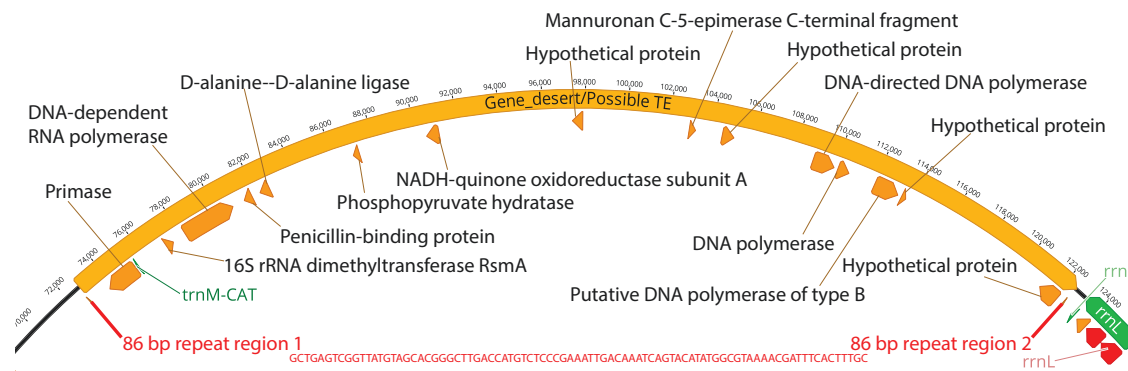


Fig. 5. Illustration of the 50 kb 'gene desert' region in the UTEX 1559 mitogenome. In total, 16 genes (15 CDSs and one tRNA) are found in this region by using ORFfinder and sequence similarity searches. The functional descriptions of the best BLAST hits of each CDSs are indicated (details can be found in [Supplementary Table S4](#)). The two 86 bp direct repeats are indicated (red color) and the sequence is provided. The image was exported from Geneious Pro v 11.0.5. The whole circular mitogenome illustration can be found in [Supplementary Fig. S2](#).

(Burke et al., 2018). A BLAST search against the recently released draft genomes of *Spirogloea muscicola* and *Mesotaenium endlicherianum* (Cheng et al., 2019) also confirmed the absence of this 50 kb region in these Zygnematophyceae genomes. When this mobile element and the genes it contains are included in the gene density assessment, the gene density in UTEX 1559 increases 10% (Fig. 4). It is clear that this mobile element accounts for a sizeable portion of the UTEX 1559 mitochondrial gene density. Future research may seek to compare congeneric and conspecific lineages in *Zygnema* to possibly determine if this mobile element persists in any other

related mitogenomes. Furthermore, given these unique features present in the UTEX 1559 mitogenome, it is important to note that there are likely to be numerous other isoforms that exist, and these different mitogenome structures present an additional avenue of exploration in future research efforts geared toward investigating mitogenome structural isoforms. Future directions of this research will investigate the likely mobile element and alternative isoforms of the mitogenome, and explore the nuclear genomes of chlorophytes, charophyte, and embryophytes to expand on the functions of genes retained through these lineages, and provide a more complete

understanding of those functions and related genes necessary for terrestrialization within the nuclear genome of UTEX 1559 and its available charophyte counterparts.

Supplementary data

Supplementary data are available at *JXB* online.

Fig. S1. Circular representation of the UTEX 1559 plastome output from Geneious Pro v 11.0.5

Fig. S2. Circular representation of the UTEX 1559 mitogenome output from Geneious Pro v 11.0.5

Table S1. Gene presence/absence and GO function groups in 33 plastomes

Table S2. Gene presence/absence and GO function groups in 18 mitogenomes

Table S3. Sequence alignment of the SAG 698–1a ribosomal protein L36 (encoded by the *rpl36* gene) against the UTEX 1559 plastome

Table S4. Gene annotation in the 50 kb ‘gene desert’ of the UTEX 1559 mitogenome

Acknowledgements

This work was mainly supported by the National Science Foundation CAREER award (DBI-1933521) and the United States Department of Agriculture (USDA) (58-8042-9-089) to YY. This work was partially completed utilizing the Holland Computing Center of the University of Nebraska, which receives support from the Nebraska Research Initiative. We would also like to acknowledge the Plant Molecular and Bioinformatics Center at Northern Illinois University for additional financial support of this project. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or USDA.

References

Adams KL, Qiu YL, Stoutemyer M, Palmer JD. 2002. Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. *Proceedings of the National Academy of Sciences, USA* **99**, 9905–9912.

Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, Palmer JD. 2010. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Molecular Biology and Evolution* **27**, 1436–1448.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.

Bankevich A, Nurk S, Antipov D, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19**, 455–477.

Barthet MM, Moukarzel K, Smith KN, Patel J, Hilu KW. 2015. Alternative translation initiation codons for the plastid maturase MatK: unraveling the pseudogene misconception in the Orchidaceae. *BMC Evolutionary Biology* **15**, 210.

Becker B, Marin B. 2009. Streptophyte algae and the origin of embryophytes. *Annals of Botany* **103**, 999–1004.

Bowman JL, Kohchi T, Yamato KT, et al. 2017. Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell* **171**, 287–304.e15.

Burke SV, Ungerer MC, Duvall MR. 2018. Investigation of mitochondrial-derived plastome sequences in the Paspalum lineage (Panicoidae; Poaceae). *BMC Plant Biology* **18**, 152.

Cameron DD, Leake JR, Read DJ. 2006. Mutualistic mycorrhiza in orchids: evidence from plant–fungus carbon and nitrogen transfers in the green-leaved terrestrial orchid *Goodyera repens*. *New Phytologist* **171**, 405–416.

Cheng S, Xian W, Fu Y, et al. 2019. Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. *Cell* **179**, 1057–1067.e14.

Civañ P, Foster PG, Embley MT, S  neca A, Cox CJ. 2014. Analyses of charophyte chloroplast genomes help characterize the ancestral chloroplast genome of land plants. *Genome Biology and Evolution* **6**, 897–911.

Cotton JL, Wysocki WP, Clark LG, Kelchner SA, Pires JC, Edger PP, Mayfield-Jones D, Duvall MR. 2015. Resolving deep relationships of PACMAD grasses: a phylogenomic approach. *BMC Plant Biology* **15**, 178.

Cuenca A, Petersen G, Seberg O, Davis JI, Stevenson DW. 2010. Are substitution rates and RNA editing correlated? *BMC Evolutionary Biology* **10**, 349.

Cuenca A, Petersen G, Seberg O, Jahren AH. 2012. Genes and processed paralogs co-exist in plant mitochondria. *Journal of Molecular Evolution* **74**, 158–169.

Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* **5**, e11147.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**, 772.

De Coster W, D’Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669.

Delaux PM, Radhakrishnan GV, Jayaraman D, et al. 2015. Algal ancestor of land plants was preadapted for symbiosis. *Proceedings of the National Academy of Sciences, USA* **112**, 13390–13395.

Delcher AL, Bratke KA, Powers EC, Salzberg SL. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679.

Delwiche CF, Cooper ED. 2015. The evolutionary origin of a terrestrial flora. *Current Biology* **25**, R899–R910.

de Vries J, Curtis BA, Gould SB, Archibald JM. 2018a. Embryophyte stress signaling evolved in the algal progenitors of land plants. *Proceedings of the National Academy of Sciences, USA* **115**, E3471–E3480.

de Vries S, de Vries J, von Dahlen JK, Gould SB, Archibald JM, Rose LE, Slamovits CH. 2018b. On plant defense signaling networks and early land plant evolution. *Communicative & Integrative Biology* **11**, 1–14.

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

Fan W, Guo W, Van Etten JL, Mower JP. 2017. Multiple origins of endosymbionts in Chlorellaceae with no reductive effects on the plastid or mitochondrial genomes. *Scientific Reports* **7**, 10101.

Fitzek E, Orton L, Entwistle S, Grayburn WS, Ausland C, Duvall MR, Yin Y. 2019. Cell wall enzymes in *Zygnema circumcarinatum* UTEX 1559 respond to osmotic stress in a plant-like fashion. *Frontiers in Plant Science* **10**, 732.

Fournier T, Gounot JS, Freel K, Cruaud C, Lemainque A, Aury JM, Wincker P, Schacherer J, Friedrich A. 2017. High-quality de novo genome assembly of the *Dekkera bruxellensis* yeast using nanopore MinION sequencing. *G3* **7**, 3243–3250.

Gauch HG. 1966. Studies on the life cycle and genetics of *Zygnema*. Ithaca, NY: Cornell University Press.

Geng S, De Hoff P, Umen JG. 2014. Evolution of sexes from an ancestral mating-type specification pathway. *PLoS Biology* **12**, e1001904.

Guiry MD, Guiry GM. 2013. AlgaeBase. National University of Ireland, Galway: <http://www.algaebase.org>.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075.

Hausner G, Olson R, Simon D, Johnson I, Sanders ER, Karol KG, McCourt RM, Zimmerly S. 2006. Origin and evolution of the chloroplast *trnK* (matK) intron: a model for evolution of group II intron RNA structures. *Molecular Biology and Evolution* **23**, 380–391.

Herburger K, Xin A, Holzinger A. 2019. Homogalacturonan accumulation in cell walls of the green alga *Zygnema* sp. (Charophyta) increases desiccation resistance. *Frontiers in Plant Science* **10**, 540.

- Hori K, Maruyama F, Fujisawa T, *et al.* 2014. *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nature Communications* **5**, 3978.
- Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AF, Wheeler TJ. 2016. The Dfam database of repetitive DNA families. *Nucleic Acids Research* **44**, D81–D89.
- Istace B, Friedrich A, d'Agata L, *et al.* 2017. *De novo* assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience* **6**, 1–13.
- Jayakumar V, Sakakibara Y. 2019. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. *Briefings in Bioinformatics* **20**, 866–876.
- Karol KG, McCourt RM, Cimino MT, Delwiche CF. 2001. The closest living relatives of land plants. *Science* **294**, 2351–2353.
- Kearse M, Moir R, Wilson A, *et al.* 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649.
- Kim HT, Kim JS, Moore MJ, Neubig KM, Williams NH, Whitten WM, Kim JH. 2015. Seven new complete plastome sequences reveal rampant independent loss of the *ndh* gene family across orchids and associated instability of the inverted repeat/small single-copy region boundaries. *PLoS One* **10**, e0142215.
- Knack JJ, Wilcox LW, Delaux PM, Ané JM, Piotrowski MJ, Cook ME, Graham JM, Graham LE. 2015. Microbiomes of streptophyte algae and bryophytes suggest that a functional suite of microbiota fostered plant colonization of land. *International Journal of Plant Sciences* **176**, 405–420.
- Knoop V, Unseld M, Marienfeld J, Brandt P, Sünkel S, Ullrich H, Brennicke A. 1996. *copa*-, *gypsy*- and *LINE*-like retrotransposon fragments in the mitochondrial genome of *Arabidopsis thaliana*. *Genetics* **142**, 579–585.
- Kolodner R, Tewari KK. 1979. Inverted repeats in chloroplast DNA from higher plants. *Proceedings of the National Academy of Sciences, USA* **76**, 41–45.
- Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences, USA* **102**, 2567–2572.
- Kozik A, Rowan BA, Lavelle D, Berke L, Schranz ME, Michelmore RW, Christensen AC. 2019. The alternative reality of plant mitochondrial DNA: one ring does not rule them all. *PLoS Genetics* **15**, e1008373.
- Kugita M, Yamamoto Y, Fujikawa T, Matsumoto T, Yoshinaga K. 2003. RNA editing in hornwort chloroplasts makes more than half the genes functional. *Nucleic Acids Research* **31**, 2417–2423.
- Lee SI, Kim JH, Park KC, Kim NS. 2015. LTR-retrotransposons and inter-retrotransposon amplified polymorphism (IRAP) analysis in *Lilium* species. *Genetica* **143**, 343–352.
- Lemieux C, Otis C, Turmel M. 2000. Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. *Nature* **403**, 649–652.
- Lemieux C, Otis C, Turmel M. 2007. A clade uniting the green algae *Mesostigma viride* and *Chlorokybus atmophyticus* represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. *BMC Biology* **5**, 2.
- Lemieux C, Otis C, Turmel M. 2014. Chloroplast phylogenomic analysis resolves deep-level relationships within the green algal class Trebouxiophyceae. *BMC Evolutionary Biology* **14**, 211.
- Lemieux C, Otis C, Turmel M. 2016. Comparative chloroplast genome analyses of streptophyte green algae uncover major structural alterations in the klebsormidiophyceae, coleochaetophyceae and zygnematophyceae. *Frontiers in Plant Science* **7**, 697.
- Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* **47**, W256–W259.
- Lin CS, Chen JJW, Chiu CC, *et al.* 2017. Concomitant loss of NDH complex-related genes within chloroplast and nuclear genomes in some orchids. *The Plant Journal* **90**, 994–1006.
- Liu Y, Medina R, Goffinet B. 2014. 350 my of mitochondrial genome stasis in mosses, an early land plant lineage. *Molecular Biology and Evolution* **31**, 2586–2591.
- Luo R, Liu B, Xie Y, *et al.* 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**, 18.
- Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH, Stern DB. 2002. The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. *The Plant Cell* **14**, 2659–2679.
- McCourt RM, Delwiche CF, Karol KG. 2004. Charophyte algae and land plant origins. *Trends in Ecology & Evolution* **19**, 661–666.
- Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: 2010 Gateway Computing Environments Workshop (GCE), New Orleans, LA, 1–8.
- Miller RD, Hoshaw RW. 1974. Cell width as a taxonomic character with special reference to *Zygnema circumcarinatum* Czurda. *British Phycological Journal* **9**, 145–148.
- Mower JP, Sloan DB, Alverson AJ. 2012. Plant mitochondrial genome diversity: the genomics revolution. In: Wendel JF, Greilhuber J, Dolezel J, Leitch IJ, eds. *Plant genome diversity volume 1: plant genomes, their residents, and their evolutionary dynamics*. Vienna: Springer Vienna, 123–144.
- Navarro C. 2017. The mobile world of transposable elements. *Trends in Genetics* **33**, 771–772.
- Newton KJ. 1988. Plant mitochondrial genomes: organization, expression and variation. *Annual Review of Plant Physiology and Plant Molecular Biology* **39**, 503–532.
- Nishiyama T, Sakayama H, de Vries J, *et al.* 2018. The chara genome: secondary complexity and implications for plant terrestrialization. *Cell* **174**, 448–464.e24.
- Oda K, Yamato K, Ohta E, Nakamura Y, Takemura M, Nozato N, Akashi K, Ohshima K. 1992. Transfer RNA genes in the mitochondrial genome from a liverwort, *Marchantia polymorpha*: the absence of chloroplast-like tRNAs. *Nucleic Acids Research* **20**, 3773–3777.
- Oldenburg DJ, Bendich AJ. 1998. The structure of mitochondrial DNA from the liverwort, *Marchantia polymorpha*. *Journal of Molecular Biology* **276**, 745–758.
- Pearson WR, Wood T, Zhang Z, Miller W. 1997. Comparison of DNA sequences with protein sequences. *Genomics* **46**, 24–36.
- Peredo EL, King UM, Les DH. 2013. The plastid genome of *Najas flexilis*: adaptation to submersed environments is accompanied by the complete loss of the NDH complex in an aquatic angiosperm. *PLoS One* **8**, e68591.
- Petersen G, Cuenca A, Zervas A, Ross GT, Graham SW, Barrett CF, Davis JI, Seberg O. 2017. Mitochondrial genome evolution in Alismatales: size reduction and extensive loss of ribosomal protein genes. *PLoS One* **12**, e0177606.
- Petersen G, Seberg O, Davis JI, Stevenson DW. 2006. RNA editing and phylogenetic reconstruction in two monocot mitochondrial genes. *Taxon* **55**, 871–886.
- Pierangelini M, Ryšánek D, Lang I, Adlassnig W, Holzinger A. 2017. Terrestrial adaptation of green algae *Klebsormidium* and *Zygnema* (Charophyta) involves diversity in photosynthetic traits but not in CO₂ acquisition. *Planta* **246**, 971–986.
- Porebski S, Bailey LG, Baum BR. 1997. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Molecular Biology Reporter* **15**, 8–15.
- Robbens S, Derelle E, Ferraz C, Wuyts J, Moreau H, Van de Peer Y. 2007. The complete chloroplast and mitochondrial DNA sequence of *Ostreococcus tauri*: organelle genomes of the smallest eukaryote are examples of compaction. *Molecular Biology and Evolution* **24**, 956–968.
- Rodriguez-R LM, Konstantinidis KT. 2016. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Preprints* **4**, e1900v1901.
- Ruan J. 2015. SMARTdenovo: ultra-fast *de novo* assembler using long noisy reads. <https://github.com/ruanjue/smartdenovo>.
- Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S. 1999. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Research* **6**, 283–290.
- Sayers EW, Agarwala R, Bolton EE, *et al.* 2019. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **47**, D23–D28.
- Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864.
- Sekimoto H, Tanabe Y, Tsuchikane Y, Shirotsaki H, Fukuda H, Demura T, Ito M. 2006. Gene expression profiling using cDNA microarray analysis of the sexual reproduction stage of the unicellular charophycean

alga *Closterium peracerosum-strigosum-littorale* complex. Plant Physiology **141**, 271–279.

Shaw AJ, Devos N, Liu Y, Cox CJ, Goffinet B, Flatberg KI, Shaw B. 2016. Organellar phylogenomics of an emerging model system: *Sphagnum* (peatmoss). Annals of Botany **118**, 185–196.

Shearman JR, Sonthirod C, Naktang C, Pootakham W, Yoocha T, Sangsarakru D, Jomchai N, Tragoonrung S, Tangphatsornruang S. 2016. The two chromosomes of the mitochondrial genome of a sugarcane cultivar: assembly and recombination analysis using long PacBio reads. Scientific Reports **6**, 31533.

Shimada H, Sugiura M. 1991. Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. Nucleic Acids Research **19**, 983–995.

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

Stancheva R, Sheath RG, Hall JD. 2012. Systematics of the genus *Zygnuma* (Zygnematophyceae, Charophyta) from Californian watersheds. Journal of Phycology **48**, 409–422.

Sugiura C, Kobayashi Y, Aoki S, Sugita C, Sugita M. 2003. Complete chloroplast DNA sequence of the moss *Physcomitrella patens*: evidence for the loss and relocation of *rpoA* from the chloroplast to the nucleus. Nucleic Acids Research **31**, 5324–5331.

Terasawa K, Odahara M, Kabeya Y, Kikugawa T, Sekine Y, Fujiwara M, Sato N. 2007. The mitochondrial genome of the moss *Physcomitrella patens* sheds new light on mitochondrial evolution in land plants. Molecular Biology and Evolution **24**, 699–709.

Timme RE, Bachvaroff TR, Delwiche CF. 2012. Broad phylogenomic sampling and the sister lineage of land plants. PLoS One **7**, e29696.

Timme RE, Delwiche CF. 2010. Uncovering the evolutionary origin of plant molecular processes: comparison of *Coleochaete* (Coleochaetales) and *Spirogyra* (Zygnematales) transcriptomes. BMC Plant Biology **10**, 96.

Turmel M, Otis C, Lemieux C. 2002a. The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. Proceedings of the National Academy of Sciences, USA **99**, 11275–11280.

Turmel M, Otis C, Lemieux C. 2002b. The complete mitochondrial DNA sequence of *Mesostigma viride* identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants. Molecular Biology and Evolution **19**, 24–38.

Turmel M, Otis C, Lemieux C. 2003. The mitochondrial genome of *Chara vulgaris*: insights into the mitochondrial DNA architecture of the last common ancestor of green algae and land plants. The Plant Cell **15**, 1888–1903.

Turmel M, Otis C, Lemieux C. 2005. The complete chloroplast DNA sequences of the charophyte green algae *Staurastrum* and *Zygnuma* reveal that the chloroplast genome underwent extensive changes during the evolution of the Zygnematales. BMC Biology **3**, 22.

Turmel M, Otis C, Lemieux C. 2006. The chloroplast genome sequence of *Chara vulgaris* sheds new light into the closest green algal relatives of land plants. Molecular Biology and Evolution **23**, 1324–1338.

Turmel M, Otis C, Lemieux C. 2007. An unexpectedly large and loosely packed mitochondrial genome in the charophyte green alga *Chlorokybus atmophyticus*. BMC Genomics **8**, 137.

Turmel M, Otis C, Lemieux C. 2013. Tracing the evolution of streptophyte algae and their mitochondrial genome. Genome Biology and Evolution **5**, 1817–1835.

Unsel M, Marienfeld JR, Brandt P, Brennicke A. 1997. The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. Nature Genetics **15**, 57–61.

Vahrenholz C, Riemen G, Pratje E, Dujon B, Michaelis G. 1993. Mitochondrial DNA of *Chlamydomonas reinhardtii*: the structure of the ends of the linear 15.8-kb genome suggests mechanisms for DNA replication. Current Genetics **24**, 241–247.

Van Etten JL, Burbank DE, Xia Y, Meints RH. 1983. Growth cycle of a virus, PBCV-1, that infects Chlorella-like algae. Virology **126**, 117–125.

Villarreal JC, Turmel M, Bourgouin-Couture M, Laroche J, Salazar Allen N, Li FW, Cheng S, Renzaglia K, Lemieux C. 2018. Genome-wide organellar analyses from the hornwort *Leiosporoceros dussii* show low frequency of RNA editing. PLoS One **13**, e0200491.

Villarreal JC, Forrest LL, Wickett N, Goffinet B. 2013. The plastid genome of the hornwort *Nothoceros aenigmaticus* (Dendrocerotaceae): phylogenetic signal in inverted repeat expansion, pseudogenization, and intron gain. American Journal of Botany **100**, 467–477.

Wang S, Li L, Li H, et al. 2019. Genomes of early-diverging streptophyte algae shed light on plant terrestrialization. Nature Plants **6**, 95–106.

Wickett NJ, Forrest LL, Budke JM, Shaw B, Goffinet B. 2011. Frequent pseudogenization and loss of the plastid-encoded sulfate-transport gene *cysA* throughout the evolution of liverworts. American Journal of Botany **98**, 1263–1275.

Wickett NJ, Mirarab S, Nguyen N, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proceedings of the National Academy of Sciences, USA **111**, E4859–E4868.

Wickham H. 2016. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag. <https://ggplot2.tidyverse.org>.

Wodniok S, Brinkmann H, Glöckner G, Heidel AJ, Philippe H, Melkonian M, Becker B. 2011. Origin of land plants: do conjugating green algae hold the key? BMC Evolutionary Biology **11**, 104.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics **20**, 3252–3255.

Yue J, Hu X, Sun H, Yang Y, Huang J. 2012. Widespread impact of horizontal gene transfer on plant colonization of land. Nature Communications **3**, 1152.

Zhao C, Wang Y, Chan KX, et al. 2019. Evolution of chloroplast retrograde signaling facilitates green plant adaptation to land. Proceedings of the National Academy of Sciences, USA **116**, 5015–5020.