Large Phylogenomic Data sets Reveal Deep Relationships and Trait Evolution in Chlorophyte Green Algae

Xi Li[1], Zheng Hou[1], Chenjie Xu[1], Xuan Shi[1], Lingxiao Yang[1], Louise A. Lewis[2], and Bojian Zhong [1],*

[1]College of Life Sciences, Nanjing Normal University, China
[2]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut, USA

*Corresponding author: E-mail: bjzhong@gmail.com

Abstract

The chlorophyte green algae (Chlorophyta) are species-rich ancient groups ubiquitous in various habitats with high cytological diversity, ranging from microscopic to macroscopic organisms. However, the deep phylogeny within core Chlorophyta remains unresolved, in part due to the relatively sparse taxon and gene sampling in previous studies. Here we contribute new transcriptomic data and reconstruct phylogenetic relationships of core Chlorophyta based on four large data sets up to 2,698 genes of 70 species, representing 80% of extant orders. The impacts of outgroup choice, missing data, bootstrap-support cutoffs, and model misspecification in phylogenetic inference of core Chlorophyta are examined. The species tree topologies of core Chlorophyta from different analyses are highly congruent, with strong supports at many relationships (e.g., the Bryopsidales and the Scotinosphaerales-Dasycladales clade). The monophyly of Chlorophyceae and of Trebouxiophyceae as well as the uncertain placement of Chlorodendrophyceae and Pedinophyceae corroborate results from previous studies. The reconstruction of ancestral scenarios illustrates the evolution of the freshwater-sea and microscopic–macroscopic transition in the Ulvophyceae, and the transformation of unicellular→colonial→multicellular in the chlorophyte green algae. In addition, we provided new evidence that serine is encoded by both canonical codons and noncanonical TAG code in Scotinosphaerales, and stop-to-sense codon reassignment in the Ulvophyceae has originated independently at least three times. Our robust phylogenetic framework of core Chlorophyta unveils the evolutionary history of phycoplast, cyto-morphology, and noncanonical genetic codes in chlorophyte green algae.

Key words: core Chlorophyta, phylotranscriptomics, systematic error, cyto-morphology, noncanonical genetic code.

Significance

Resolving phylogenetic relationships within and among the classes of chlorophyte green algae (Chlorophyta) remain elusive because of the antiquity of the clades and the evolutionary history of rapid radiations. We obtained higher support values at the nodes of Bryopsidales and the Scotinosphaerales that remain controversial in previous studies. We investigated large-scale genes and presented for the first time that serine is encoded by both canonical codons and noncanonical TAG code in Scotinosphaerales. Our phylogenetic inferences point toward three independent gains of stop-to-sense codon reassignment in *Scotinosphaera* sp. (Stop→Ser), Dasycladales (Stop→Gln), and Trentepohliales + Cladophorales + *Blastophysa* (Stop→Gln). The largely increased taxon and gene sampling compared with previous studies has provided a more robust phylogenetic framework of core Chlorophyta.

Introduction

The core Chlorophyta is an ancient, species-rich group of photosynthetic eukaryotes that includes the classes Pedinophyceae, Chlorodendrophyceae, and the UTC clade (Ulvophyceae, Trebouxiophyceae, and Chlorophyceae). Algae in the core Chlorophyta are of particular evolutionary

interest because they exhibit high morphological and cytological diversity (De Clerck et al. 2012; Leliaert et al. 2012), ranging from unicellular, colonial, and multicellular to a unique mechanism of wounding response as giant-celled siphonous/siphonocladous thalli forms (Mine et al. 2008; Cocquyt, Verbruggen, et al. 2010). Cytokinesis in the core Chlorophyta occurs by three different ways (as illustrated in fig. 1a): 1) microtubules-mediated furrowing with cell elongation in the Prasinophytes, Pedinophyceae, and most of species in the Ulvophyceae, as a relatively primitive mode (Marin 2012); 2) phycoplast-mediated furrowing without cell elongation in the Chlorodendrophyceae, Chlorophyceae, and Trebouxiophyceae, as a group of microtubules parallel to the division plane, which is essential in forming cell wall (Leliaert 2019); 3) phragmoplast-mediated cytokinesis as a well-developed mode evolved in land plants, some streptophytes, and Trentepohliales (an entirely terrestrial order in Ulvophyeae). Phragmoplast is an array of microtubules parallel to the spindle axis, which is essential in forming cell plates and cell wall (Leliaert et al. 2012; Leliaert 2019). The core Chlorophyta also occupies diverse habitats, being marine, freshwater, and terrestrial habitats (Dittami et al. 2017; Leliaert 2019) and has played a prominent role in the global ecosystem (O'Kelly 2007; Leliaert et al. 2012). The flourishing of marine algae may have acted as the principal source of cloud condensation nuclei and contributed to the "Snowball Earth" in the Neoproterozoic era (Gage et al. 1997; Pierrehumbert et al. 2011; Becker 2013; Rooney et al. 2014; Prave et al. 2016). Among the divergent chlorophyte lineages, the Ulvophyceae (seaweeds) contains groups with a noncanonical genetic code in nuclear genes (Gile et al. 2009; Cocquyt et al. 2010), but the origin and evolution of the noncanonical genetic code are ambiguous due to the lack of an accurate phylogenetic tree among ulvophytes and the core Chlorophyta (Del Cortona and Leliaert 2018; Fang et al. 2018). To better understand the evolutionary history of marine-freshwater transitions, cytological diversity, and codon reassignments of green algae, it is essential to determine a solid phylogenetic framework for the core Chlorophyta.

Although chloroplast phylogenetic studies uncovered five major classes of the core Chlorophyta, the deep-level (e.g., order or class) relationships within the core Chlorophyta have long been debated and the monophyly of the classes Ulvophyceae and Trebouxiophyceae remained ambiguous (Cocquyt, Verbruggen, et al. 2010; Novis et al. 2013; Fučíková et al. 2014; Melton et al. 2015; Turmel et al. 2017; Fang et al. 2018). Moreover, the phylogenetic placements of Chlorodendrophyceae and Pedinophyceae are unresolved using chloroplast and transcriptomic data (Turmel et al. 2015, 2016; One Thousand Plant Transcriptomes Initiative 2019; Del Cortona et al. 2020), as well as in the morphology-based studies (Stewart and Mattox 1975; Baudelet et al. 2017; Leliaert 2019). The algae in the class Chlorodendrophyceae and Prasinophyceae share the same cell wall composition (Baudelet et al. 2017). On the contrary, Pedinophyceae and the prasinophytes share the same type of microtubules-mediated cytokinesis, compared with the phycoplast-mediated cytokinesis in Chlorodendrophyceae, Chlorophyceae, and Trebouxiophyceae (fig. 1a) (Stewart and Mattox 1975; Leliaert 2019). The outgroup choice has been shown to be critical in deep phylogenomic analyses (De la Torre-Bárcena et al. 2009; Wilberg 2015; Borowiec et al. 2019), which may result in the discordant relationships in the early-diverging lineages of the core Chlorophyta (Chlorodendrophyceae and Pedinophyceae).

These inconsistent relationships among the major five classes within the core Chlorophyta may be in part due to poor taxon/gene sampling, resulting in both stochastic and systematic errors in phylogenetic analyses (Cocquyt, Verbruggen, et al. 2010; Smith et al. 2011; Fučíková et al. 2014; Sun et al. 2016; Fang et al. 2017, 2018; Turmel and Lemieux 2018). It has been suggested that estimating accurate species trees is difficult with a limited number of genes (Sayyari and Mirarab 2016) and inadequate models (Philippe et al. 2017). Further, the short and deep branches of the core Chlorophyta indicate that these algae may be subject to lineage-specific differences of rate variation, deviating amino acid/nucleotide composition (codon-usage biases), or incomplete lineage sorting (ILS) (Herron et al. 2009; Cocquyt, Verbruggen, et al. 2010; Cox et al. 2014; Liu et al. 2014; Irisarri et al. 2017). These issues can be ameliorated by using heterogeneous models, summary-coalescent methods, and large-scale molecular data with dense taxon sampling. Prespecified empirical mixture models (e.g., site-heterogeneous model C20) have been reported to reduce systematic error and suppress long-branch attraction artifact (Quang et al. 2008). The GHOST model (i.e., heterotachous model H4) can address lineage-specific asymmetries by recovering historical signal from heterotachously evolved lineages (Crotty et al. 2020). The multispecies summary-coalescent (MSC) model outperforms the concatenation method in the anomaly zone and ILS in species tree estimation (Xi et al. 2014; Edwards 2016; Edwards et al. 2016; Jiang et al. 2020). The anomaly zone is a region of species tree with very short branches, in which the most common gene tree contradicts the species tree (Degnan and Rosenberg 2009; Liu et al. 2010, 2015). The simulation and empirical studies have highlighted the importance of sufficient number of genes (e.g., >500) and observed the trend that species tree estimation error is reduced as the number of genes is increased (Mirarab et al. 2014; Sayyari and Mirarab 2016). Blom et al. (2017) also emphasized the need for large-scale data sets to better resolve challenging relationships and increase the supports in some nodes. Although large-scale data sets dramatically reduce stochastic errors in phylogenetic inference, they simultaneously amplify systematic errors when using incorrect models (Lartillot et al. 2007; Philippe et al. 2017). To reduce stochastic and systematic errors on phylogenetic inference, Philippe et al. (2017) argued that there is an

optimal compromise between taxon and gene sampling. Moreover, Xi et al. (2016) determined that the accuracy of species tree estimation is increased as more genes (500–2,000 genes) were analyzed, regardless of missing data, high gene rate heterogeneity, and ILS. Although species tree estimation can greatly benefit from large-scale data sets, MSC methods are sensitive to gene tree error (Mirarab and Warnow 2015; Zhang et al. 2018). The updated ASTRAL II and III can consider polytomies in input gene trees, and collapse low support branches in order to increase the accuracy of the gene tree estimation by reducing weak phylogenetic signal (Zhang et al. 2017, 2018).

In this study, we sequenced the transcriptomes of seven species and compiled public data sets from 10 nuclear genomes and 69 transcriptomes in the core Chlorophyta. To minimize potential impacts of gene-tree estimation error, we applied site-heterogeneous model in gene tree estimation and collapsed low bootstrap supports on ASTRAL-based species tree estimation. We demonstrated that species tree topologies generated from different methods and models were largely congruent. We found the Ulvophyceae to be nonmonophyletic and suggested a hard polytomy for the relationships among Chlorodendrophyceae, Pedinophyceae, and the UTC clade. We investigated large-scale genes and presented for the first time that serine is encoded by both canonical codons and noncanonical TAG code in Scotinosphaerales. Our phylogenetic inferences point toward three independent gains of stop-to-sense codon reassignment in *Scotinosphaera* sp. (Stop→Ser), Dasycladales (Stop→Gln), and Trentepohliales + Cladophorales + *Blastophysa* (Stop→Gln).

Results

Species Tree Concordance and Conflict from Phylogenomic Analyses

The number of ortholog groups (OGs) and the total lengths of amino acid alignments in each orthologous group are presented in table 1. Based on different column and taxon occupancy, we obtained various numbers of orthologous genes: 2,698 (50% column occupancy, 50% taxon occupancy), 1,984 (80%, 50%), 1,404 (50%, 80%), and 1,447 (80%, 80%) in the 70 species data set, whereas 2,629 (50%, 50%), 1,880 (80%, 50%), 1,432 (50%, 80%), and 1,415 (80%, 80%) OGs were retrieved in the 86 species data set.

The species tree topologies inferred from ASTRAL- and concatenation-based analyses were highly congruent across different data sets at the class level, except for the position of the Pedinophyceae, the Chlorodendrophyceae, and a few orders in the major classes (fig. 1b and supplementary figs. S1 and S2, Supplementary Material online). All the phylogenetic trees across different data sets strongly supported the monophyly of the Chlorophyceae, Trebouxiophyceae,

Pedinophyceae, and Chlorodendrophyceae (fig. 2 and supplementary figs. S3–S29, Supplementary Material online). Either Chlorodendrophyceae or Pedinophyceae was found to be sister to the remaining core Chlorophyta (fig. 1b). The Ulvophyceae was polyphyletic and separated into two clades across all the analyses, Ulvophyceae I (included most ulvophycean species) and Ulvophyceae II (Bryopsidales). Ulvophyceae II and the Chlorophyceae formed a clade, which was sister to Ulvophyceae I (fig. 2 and supplementary figs. S3–S29, Supplementary Material online). The incongruence between MSC and concatenation analyses mainly involved orders that contain only a few species (e.g., *Ignatius tetrasporus*, *Blastophysa* cf. *rhizopus*, and *Scotinosphaera* sp. in Ulvophyceae; *Leptosira obovata* and *Microthamnion kuetzingianum* in Trebouxiophyceae) or orders with limited taxon sampling in our study (e.g., *Oedogonium cardiacum*, *Chaetopeltis orbicularis*, and *Aphanochaete repens* in Chlorophyceae) (fig. 2 and supplementary figs. S1 and S2, Supplementary Material online).

The species trees based on concatenation analyses of 70 taxa were highly similar across the four data sets, except for the position of the Pedinophyceae, the Chlorodendrophyceae, and some ulvophyceans (fig. 1b and supplementary figs. S1 and S2, Supplementary Material online). The concatenation analyses using the site-heterogeneous model supported the Pedinophyceae as sister to the UTC clade (with SH-aLRT $\geq$ 88% and UFboot $\geq$ 93%), together being sister-group to the Chlorodendrophyceae with maximal bootstrap support (BS) (fig. 1b, T1). In contrast, the concatenation analyses using site-homogeneous and heterotachy models were totally congruent at the ordinal level or above (supplementary figs. S18–S25, Supplementary Material online), and fully supported the sister relationship of the Chlorodendrophyceae and the UTC clade, together being sister-group to the Pedinophyceae with maximal BS (fig. 1b, T2). T2 topology was also supported in all the 86 taxa concatenation analyses (supplementary figs. S30–S41, Supplementary Material online). A hard polytomy among Chlorodendrophyceae, Pedinophyceae, and the UTC clade was detected in most of the analyses (supplementary table S1, Supplementary Material online). The relationships among the Scotinosphaerales, Oltmannsiellopsidales, *Ignatius*, and the Ulvales-Ulotrichales clades across different data sets and analyses were instable in the analyses with 70 taxa (supplementary fig. S1, Supplementary Material online), but Scotinosphaerales+Dasycladales and Oltmannsiellopsidales consistently grouped together in all the analyses of 86 taxa (supplementary figs. S30–S41, Supplementary Material online), which was similar to T2 in supplementary figure S1, Supplementary Material online. All our concatenation analyses supported that *Blastophysa* cf. *Rhizopus* was sister to the Trentepohliales with maximum BS when using site-homogeneous and heterotachy models (supplementary fig. S2, Supplementary Material online, T1), but with low BS

Table 1

Four different column and taxon occupancy treatments

Data Partition/Missing Data	>50% Taxon Occupancy (Amino Acid)	>80% Taxon Occupancy (Amino Acid)
0.5 column occupancy (70 taxa)	2,698 (648,448)	1,404 (347,478)
0.8 column occupancy (70 taxa)	1,984 (415,483)	1,447 (320,498)
0.5 column occupancy (86 taxa)	2,629 (619,049)	1,432 (349,720)
0.8 column occupancy (86 taxa)	1,880 (390,307)	1,415 (307,954)
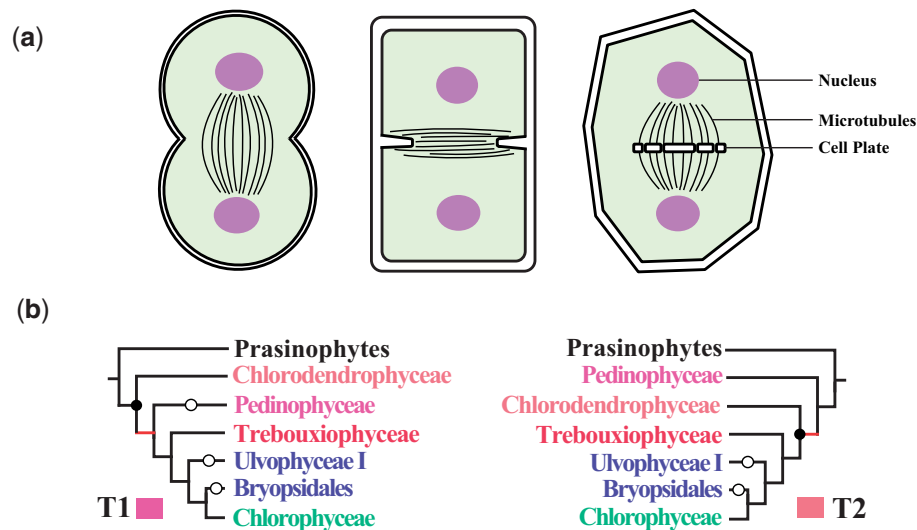
taxon/column occupancy (gene number)	concatenation approach			multispecies coalscent model		
	LG+F+G SH-aLRT/UFboot	LG+F+H4 SH-aLRT/UFboot	LG+F+G+C20 SH-aLRT/UFboot	Bootstrap cutoff	LG+G PP/BS	LG+F+G+C20 PP/BS
0.5/50% (2698)	100 / 100	100 / 100	88 / 97	BS10	0.67 / 54	0.82 / 84
				BS20	0.50 / 50	0.84 / 80
0.5/80% (1404)	100 / 100	100 / 100	96 / 97	BS10	0.75 / 55	0.75 / 80
				BS20	0.46 / 49	0.64 / 69
0.8/50% (1984)	100 / 100	100 / 100	99 / 99	BS10	0.48 / 56	0.76 / 88
				BS20	0.49 / 59	0.70 / 62
0.8/80% (1447)	100 / 100	100 / 100	88 / 93	BS10	0.37 / 55	0.81 / 78
				BS20	0.44 / 53	0.73 / 65

FIG. 1.—Different types of cytokinesis (adapted from Leliaert [2019]) and two recovered topologies and support values of the major classes of the core Chlorophyta. (*a*) The left: microtubules independent furrowing. The middle: phycoplast-dependent furrowing, a group of microtubules parallel to the nuclear division plane, which is necessary for maintaining a solid cell wall. The right: phragmoplast-mediated cytokinesis (e.g., Trentepohliales), which is an array of microtubules parallel to the spindle axis, guiding the formation of cell plates and new cell wall. (*b*) The purple boxes represent the topology on the left supporting Pedinophyceae as sister to the UTC clade; the pink boxes represent the topology on the right supporting Chlorodendrophyceae as sister to the UTC clade. The numbers in the boxes indicate support values obtained from the key node (red branch) in two topologies. Two hypotheses were proposed for the gain of the phycoplast (black dot), which has been lost (white dot) in the Pedinophyceae and the Ulvophyeae (T1) or subsequently lost only in the Ulvophyeae (T2).

and less consistent when using the site-heterogeneous model from 2,698 genes (supplementary fig. S2, Supplementary Material online, T2).

The topologies of the species tree spanning different missing data and BS cutoffs in ASTRAL analyses were more similar than in concatenation analyses, as only the branching order of the Pedinophyceae and the Chlorodendrophyceae remained in conflict (fig. 1b and supplementary figs. S1 and S2, Supplementary Material online). Two topologies were recovered in the core Chlorophyta among our ASTRAL analyses using homogeneous model. The species trees with 10% BS value cutoff supported the first split segregating the Pedinophyceae and then the Chlorodendrophyceae, whereas the species trees with 20% BS value cutoff were less congruent among different data sets using the homogeneous model (fig. 1b). In

contrast, our ASTRAL analyses under a site-heterogeneous model supported that the Chlorodendrophyceae as the first split of core Chlorophyta, and the Pedinophyceae was sister to the UTC clade (fig. 1b, T1).

Relationships among the Major Orders in Chlorophyceae, Ulvophyceae, and Trebouxiophyceae

The relationships among major orders in Chlorophyceae, Ulvophyceae, and Trebouxiophyceae are presented in figure 2. The species trees were highly similar across different analyses, especially when applying the site-heterogeneous model. Therefore, we used the ASTRAL-based phylogenetic tree based on the site-heterogeneous model, with support branches lower than 20% collapsed from 2,698 genes as the reference species tree (fig. 2) to display relationships among the major orders within the UTC clade.

In agreement with previous studies, Chlorophyceae was monophyletic and consisted of five major orders (Lemieux et al. 2015; Fučíková et al. 2016, 2019; One Thousand Plant Transcriptomes Initiative 2019). The Chlamydomonadales was fully supported as sister to Sphaeropleales forming the CS clade. The CS clade was close to the OCC clade, including three orders (i.e., Oedogoniales, Chaetopeltidales, and Chaetophorales) (fig. 2). The monophyly and relationships within the Chlamydomonadales were fully supported. Spermatozopsis, Treubarinia, and Golenkinia had been previously reported within Chlamydomonadales (Nakada et al. 2008), whereas our phylogenomic results proposed that they were topologically closer to Sphaeropleales (fig. 2), as was supported in Lemieux et al. (2015) and Fučíková et al. (2019).

We sampled major orders in Ulvophyceae to expand taxon sampling. Our results showed that Ulvophyceae is not monophyletic and is divided into two groups: Ulvophyceae I and II
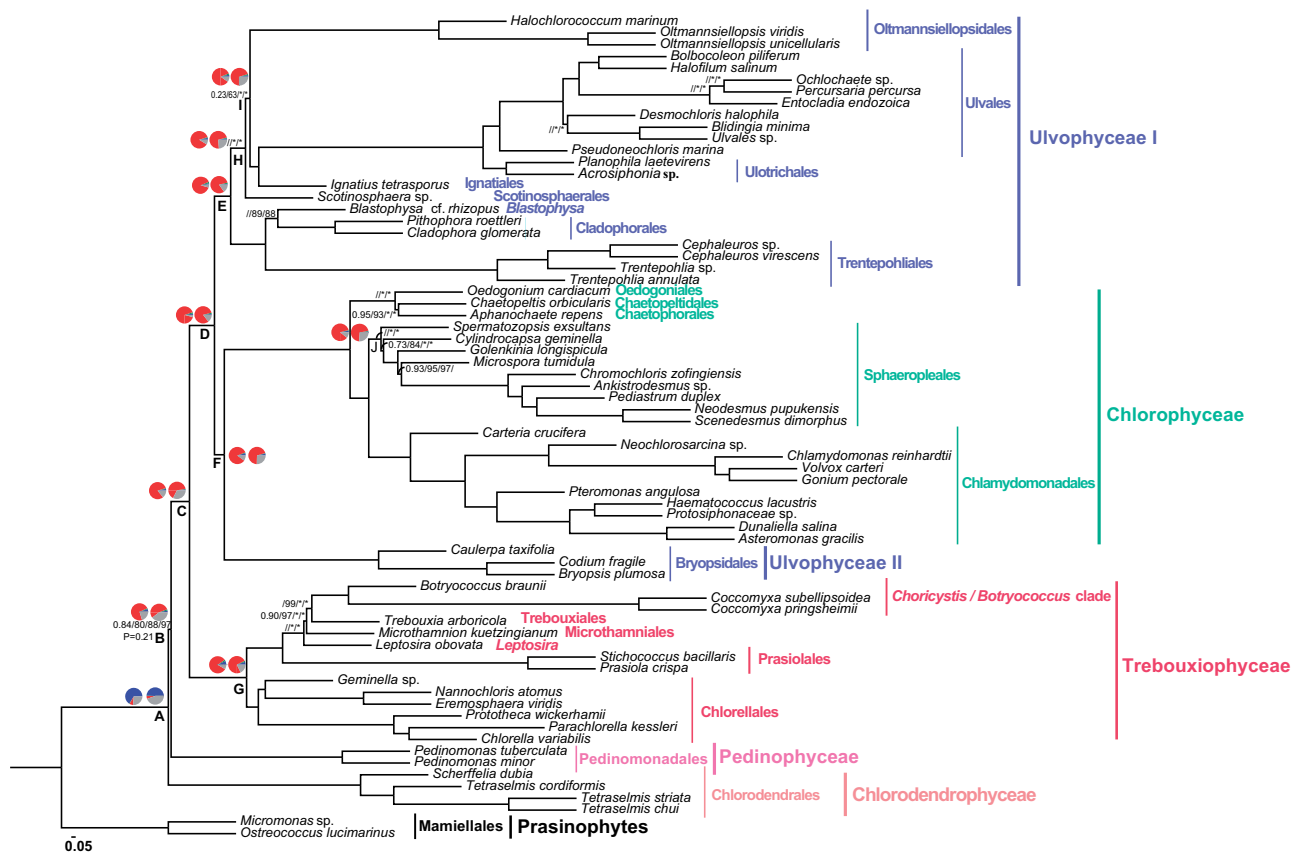

Fig. 2.—Resulting phylogeny of the core Chlorophyta obtained from a multilocus species tree analysis of 2,698 orthologous nuclear genes with site-heterogeneous model and 20% BS value cutoff based. Support values are shown only for branches receiving less than full support from PP/BS/SH-aLRT/UFboot analyses, respectively. The P value for the polytomy test is shown on node B. Conflicting topologies between ASTRAL and concatenation analyses of 2,698 genes with site-heterogeneous model and 20% BS value cutoff are indicated with asterisks. Pie charts at each node indicate the proportion of gene trees concordance and conflicts against the reference species tree, with 10% BS value cutoff on the left and 20% on the right: blue—the proportion of gene trees supporting the shown topology; green—the proportion of gene trees supporting the most common conflicting topology; red—the proportion of gene trees supporting all other supported conflicting topologies; gray—the proportion of gene trees below the bootstrap cutoff at that node. Nodes labeled A–J are with high gene tree discordances that have been discussed.

(fig. 2). Ulvophyceae I comprises two distinct clades: one clade consisting of Scotinosphaerales, Dasycladales, Ignatiales, and the UUO lineage (Ulvales, Ulotrichales, and Oltmannsiellopsidales) (fig. 3 and supplementary fig. S1, Supplementary Material online), and another clade consisting of Trentepohliales, Cladophorales, and *Blastophysa* (supplementary fig. S2, Supplementary Material online). Ulvophyceae II (Bryopsidales) was strongly supported as sister to Chlorophyceae in all our analyses.

Our results revealed the monophyly of Trebouxiophyceae (fig. 2). The placement of Chlorellales as sister to core Trebouxiophyceae was fully supported, and *Geminella* was nested within Chlorellales. Prasiolales was fully supported as sister to the remaining core Trebouxiophyceae (fig. 2).

Gene Tree and Species Tree Concordance and Conflicts

The gene tree concordance and conflicts were most sensitive to different BS cutoff values and were not strongly influenced by the missing data and models (supplementary figs. S42–S49, Supplementary Material online). We found that gene tree conflicts were more common at deep-level relationships and short internodes within core Chlorophyta. Increasing the BS cutoffs from 10% to 20% dramatically reduced gene tree conflicts at the deep nodes (e.g., nodes A-G, fig. 2) and at some uncertain internodes (e.g., nodes H and I in the Ulvophyceae and node J in the Chlorophyceae, fig. 2). Gene tree discordance was not impacted at shallow nodes (e.g., within-genus). The Robinson-Foulds (RF) distances between species trees from all the ASTRAL analyses and the reference tree were extremely low (supplementary table S2, Supplementary Material online). No more than two bipartitions among 70 species (i.e., RF distance of 0, 2, or 4) were different in all the ASTRAL analyses, whereas most of the RF distances were zero when using the site-heterogeneous model. The RF distances between species trees from the concatenation analyses and the reference tree were higher, although no more than ten bipartitions were different. The RF distances were decreased when site-heterogeneous model was applied in both ASTRAL and concatenation analyses.

Noncanonical Genetic Code in the Ulvophyceae

The noncanonical genetic codes in the Ulvophyceae are presented in table 2, figure 3, and supplementary figure S50, Supplementary Material online. The presence of noncanonical TAG/TAA codons in *actin*, *EF-1a*, *GPI*, *GAPA*, and *histone H3* genes in *Cephaleuros virescens*, *Cephaleuros* sp., *Trentepohlia annulata*, *Pithophora roettleri*, and *Cladophora glomerata* were newly identified in our study (table 2, fig. 3, and supplementary fig. S50, Supplementary Material online), although all these noncanonical codon usages were expected given the presence of such alternative codes in other members of these orders (Cocquyt et al. 2010). Like the analyses of the ten genes, the investigation of 30 ulvophycean species in the 3,690 OGs showed a broad distribution of noncanonical code. Here, we illustrate a few of cases of noncanonical genetic code (fig. 3 and supplementary fig. S50, Supplementary Material online). The alignments of the 3,690 OGs are available at Figshare (https://figshare.com/s/17b87d35616f7ee2e7ff). Glutamine is encoded by both canonical CAG/CAA codons and noncanonical TAG/TAA codons in Dasycladales and Trentepohliales + Cladophorales + *Blastophysa* clade, which is in line with previous studies (Cocquyt et al. 2010; Del Cortona et al. 2020). We also found that when *actin*, *EF-1alpha*, *GAPA*, and *histone H3* genes have the noncanonical TAR→Gln code, TGA is used as a functional stop codon (e.g., *actin* gene of *P. roettleri*) (fig. 3 and supplementary fig. S50, Supplementary Material online), as reported by Cocquyt et al. (2010). On the contrary, when these genes have only the canonical code, TAR and TGA are used as stop codons (e.g., *histone H3* gene of *P. roettleri* and *Cephaleuros* sp.) (fig. 3 and supplementary fig. S50, Supplementary Material online). Although the presence of noncanonical nuclear genetic code has been briefly mentioned in Del Cortona et al. (2020), to the best of our knowledge, there are few studies investigating the usages of genetic code in the Scotinosphaerales. Unlike Trentepohliales and Cladophorales, we observed that in *Scotinosphaera* sp. (Scotinosphaerales), serine is encoded by both canonical codons and noncanonical TAG code in *GPI*, *actin* genes, and many genes from 3,690 OGs (fig. 3 and supplementary fig. S50, Supplementary Material online). In *Scotinosphaera* sp., TAA is used as a stop codon in *actin*, *EF-1alpha*, and *histone H3* genes, whereas *GAPA* gene has TGA as a stop codon. Its stop codon in *GPI* was uncertain due to the incomplete sequence in *Scotinosphaera* sp. As expected, genes from members of UUO clade and *Ignatius* appear to use standard nuclear genetic codes (Cocquyt et al. 2010; Del Cortona and Leliaert 2018). There are no genes that have been found to terminate with a TAG codon in *Scotinosphaera* sp. in neither the five nuclear genes nor the 3,690 OGs due to the incompleteness of transcriptomes (without good quality of annotation) and this requires more investigation. We investigated the chloroplast genome of *Scotinosphaera* sp. (Fang et al. 2018) and found that the chloroplast genes mostly use TAA as stop code, whereas in a few cases end with TAG. Thus, we conjecture that some nuclear genes of Scotinosphaerales are likely to use TAG as stop code when they have only the canonical code.

Discussion

Impacts of outgroups, missing data, BS cutoffs, and model misspecification on the phylogenetic inference of the core Chlorophyta.

FIG. 3.—Selected regions of the alignments illustrating genetic code alteration on TAR (Stop→Gln) in Dasycladales and Trentepohliales + Cladophorales + *Blastophysa* (in blue boxes) and TAG (Stop→Ser) in Scotinosphaerales (in red boxes) in a–f. Stop codons are shown as asterisk. (a) GPI; (b) actin; (c) Gene996; (d) Gene1041; (e) Gene2764; (f) Gene3083. (g) Phylogenetic distribution of the noncanonical codes is indicated with horizontal bars (TAR→Gln in blue and TAG→Ser in red).

The novelties of our study are as follows: 1) improve the robustness of phylogenetic inference within the core Chlorophyta, which remains challenging in the One Thousand Plant Transcriptomes Initiative (2019) and Del Cortona et al. (2020) (e.g., the placement of Bryopsidales and the Scotinosphaerales-Dasycladales clade); 2) investigate the impacts of outgroup choice on the placement of the early-diverging lineages of the core Chlorophyta (i.e., Chlorodendrophyceae and Pedinophyceae); 3) evaluate the influence of missing data, BS cutoffs, models, and concatenation versus MSC methods on phylogenetic reconstruction; 4) demonstrate the evolutionary history of noncanonical genetic codes in the Ulvophyceae and uncover for the first time the TAG→Ser codon reassignment in Scotinosphaerales. We

Table 2

The presence of noncanonical nuclear genetic code

	Actin	EF-1α	GPI	GAPA	Histone
Scotinosphaera sp.[a]	Ser	No	Ser	No	No
Cladophora glomerata[b]	No	Gln	Gln	Gln	Gln
Pithophora roettleri[b]	Gln	Gln	No	Gln	No
Cephaleuros virescens[c]	No	No	No	No	Gln
Cephaleuros sp.[c]	No	Gln	No	No	No
Trentepohlia annulata[c]	No	No	No	Gln	No
Trentepohlia sp.[c]	No	No	No	No	No
Oltmannsiellopsis viridis[d]	No	No	No	No	No
Oltmannsiellopsis unicellularis[d]	No	No	No	No	No
Halochlorococcum marinum[d]	No	No	No	No	No

Note—Noncanonical nuclear genetic codes are indicated by Gln and Ser, representing stop-to-sense codon reassignments of TAR (stop→Gln) and TAG (stop→Ser). Canonical nuclear genetic codes are indicated by "No."
[a]Scotinosphaerales;
[b]Cladophorales;
[c]Trentepohliales;
[d]Oltmannsiellopsidales.

found that large-scale data sets (1,404–2,698 genes, table 1) and broad taxon sampling were beneficial to phylogenetic inferences, regardless of the extent of missing data. In agreement with previous analyses of One Thousand Plant Transcriptomes Initiative (2019) and Del Cortona et al. (2020), the common branch uniting the UTC clade received maximal support by all methods applied. The Chlorophyceae was confirmed as monophyletic, and we found the Ulvophyceae not to be monophyletic, which was consistent with several phylogenetic analyses of chloroplast genomes (Fučíková et al. 2014; Melton et al. 2015; Turmel et al. 2017; Fang et al. 2018).

It is worth noting that missing data and models affected the placement of the Scotinosphaerales in the concatenation analyses using 70 taxa, in which Scotinosphaerales was sister to Oltmannsiellopsidales or *Ignatius* (supplementary fig. S1, Supplementary Material online). On the contrary, Scotinosphaerales + Dasycladales and Oltmannsiellopsidales consistently grouped together with full support in all the concatenation analyses with 86 taxa (supplementary fig. S30–S41, Supplementary Material online). However, Scotinosphaerales + Dasycladales was moderately supported (posterior probability [PP] = 0.61 and BS = 83) as sister to Trentepohliales + Cladophorales + *Blastophysa* in Del Cortona et al. (2020). Ulvophyceae II (Bryopsidales) was strongly supported as sister to Chlorophyceae in all our 70 and 86 taxon analyses (figs. 2 and 3), and the nonmonophyly of the Ulvophyceae (i.e., Bryopsidales + Chlorophyceae) was also observed in the ASTRAL-based analyses in One Thousand Plant Transcriptomes Initiative (2019) with weak support

(PP = 0.35) and concatenation-based analyses in Del Cortona et al. (2020) with relatively high support (BS = 95). On the contrary, monophyly of the Ulvophyceae was fully supported in concatenation-based analyses in One Thousand Plant Transcriptomes Initiative (2019) (i.e., Bryopsidales as sister to Trentepohliales + Cladophorales + *Blastophysa*) and lowly supported (PP = 0.63) in the ASTRAL-based analyses in Del Cortona et al. (2020) (i.e., Bryopsidales as sister to the Ulvophyceae I). We suggest that the higher support value and consistency (e.g., at the placement of Bryopsidales and Scotinosphaerales + Dasycladales clade) in our study may be due to larger-scale gene sampling (2,698/2,629 genes), comparing to 410 and 539 in One Thousand Plant Transcriptomes Initiative (2019) and Del Cortona et al. (2020), respectively.

The gene tree–species tree conflict did not decrease with increasing column and taxon occupancy (supplementary figs. S42–S49, Supplementary Material online), in contrast to the studies of Lemmon et al. (2009) and Roure et al. (2013). This may result from the relatively low number of genes and taxa sampled in their studies, because limited characters could indirectly decrease the accuracy of phylogenetic inferences by uneven or unrepresentative sampling of gene histories (Xi et al. 2016; Parks et al. 2018). Gene tree conflicts in species tree inference are commonly reported (Degnan and Rosenberg 2006, 2009; Parks et al. 2018). Gene tree discordances were reduced after increasing the BS cutoffs at these focal nodes (fig. 2). This mirrors trends observed in Parks et al. (2018) that suggested gene tree conflicts tend to have low BS. Some nodes with high/full support were accompanied with great gene tree discordances in our study, mostly in short internodes at deep level (e.g., nodes A-J, fig. 2 and supplementary figs. S42–S49, Supplementary Material online), as have been reported in Sayyari and Mirarab (2016). For example, the placement of Chlorodendrophyceae and Pedinophyceae remain controversial (fig. 1b, node B in fig. 2), as reported in previous studies (Turmel et al. 2015, 2016; Del Cortona et al. 2020). The outgroup choice affected the placement of the Chlorodendrophyceae and Pedinophyceae, when using the site-heterogeneous model in concatenation-based analyses. Pedinophyceae was observed as the earliest-diverging lineage of the core Chlorophyta in all the concatenation analyses using 86 taxa, regardless of the substitution model used (supplementary figs. S30–S41, Supplementary Material online). This is consistent with the concatenation analyses based on 70 taxa with site-homogeneous and heterotachy models (fig. 1b, T2). However, we note that the placement of Pedinophyceae changed when using the site-heterogeneous model in the concatenation analyses of 70 taxa (fig. 1b, T1). As shown in Del Cortona et al. (2020), our analyses also suggest a hard polytomy for the relationships among Chlorodendrophyceae, Pedinophyceae, and the UTC clade (supplementary table S1, Supplementary Material

online), which likely results from their antiquity and short time span of diversification (O'Kelly 2007; Cocquyt et al. 2010).

RF distances supported that the use of the MSC model was beneficial to the reliability of phylogenetic reconstruction. We assumed that the reliability of ASTRAL-based species tree estimation can be increased with a sufficiently large number of genes, regardless of the extent of missing data, gene rate heterogeneity, and ILS, as reported by Xi et al. (2016). The RF distances between species trees were relatively higher when inferred with the homogeneous model than with the site-heterogeneous model (supplementary table S2, Supplementary Material online). Whereas RF distances between trees were relatively lower in all the ASTRAL-based analyses than in concatenation-based analyses. This indicated greater congruence among species trees from ASTRAL-based analyses (supplementary table S2, Supplementary Material online). Our analyses using the heterotachy model were highly similar to those carried out using the homogeneous model (fig. 2 and supplementary figs. S1 and S2, Supplementary Material online), thus we infer that our data sets were not sensitive to the issues of asymmetries. Two main reasons are suggested: 1) the reference tree we used in the RF distances analyses was based on the MSC model; and 2) the support for concatenation methods may be spuriously high (Liu and Edwards 2009; Roch and Steel 2015; Edwards et al. 2016).

Habitat and Trait Evolution in the Core Chlorophyta

The evolution of the phycoplast is thought to be a key step and an early innovation during the evolution of the core Chlorophyta. Chlorodendrophyceae and UTC classes are characterized by the possession of a phycoplast-mediated cytokinesis, which was subsequently lost in the Ulvophyceae (fig. 1b), as reported in previous studies (Leliaert et al. 2012; Fučíková et al. 2014). Marin (2012) emphasized that the evolution of the phycoplast (fig. 1a) was significant, as it could mediate cell division without cell elongation. Based on our study, we propose two hypotheses for the gain and loss of the phycoplast (fig. 1b): 1) depending on the topology of T1, the phycoplast evolved in the common ancestor of the core Chlorophyta and was lost independently in the Pedinophyceae and the Ulvophyceae; 2) depending on the topology of T2, the phycoplast evolved in the common ancestor of Chlorodendrophyceae and the UTC clade and was lost only in the Ulvophyceae. Further investigation using denser taxon sampling from Chlorodendrophyceae and Pedinophyceae will contribute to a better understanding of evolution of the phycoplast.

The analyses of ancestral character reconstruction illustrated that the early green algae might originate from unicellular and marine environments, and later successfully colonize to freshwater and terrestrial in the core Chlorophyta (except the Ulvophyceae) (fig. 4). Freshwater to marine transitions have occurred several times in the core Chlorophyta,

especially in the Ulvophyceae, which is a variety of group living mainly in marine, ranging from unicellular to macroscopic multicellular, siphonous, and siphonocladous forms. Del Cortona et al. (2020) hypothesized that the transition of planktonic to benthic habitat played a crucial role in the diversification and macroscopic growth in ulvophycean species during the Cryogenian. The dominance of marine habitat for giant-celled algae (e.g., Acetabularia, Bryopsis, Caulerpa in the Ulvophyceae) (fig. 4) indicates that they have evolved distinct mechanisms for maintaining cell–cell osmotic conditions (Bisson et al. 2006; Mine et al. 2008). This freshwater to marine transformation may associate with the loss of phycoplast in the Ulvophyceae, considering that Ulvophyceae and Prasinophytes share similar type of microtubules-mediated cytokinesis and marine habitat.

Our phylogenetic analyses provided a solid framework and important insights into cytomorphological evolution in the Ulvophyceae (fig. 4). We confirmed that multicellularity evolved independently in different clades of the Ulvophyceae, such as the Ulvales-Ulotrichales clade (Cocquyt, Verbruggen, et al. 2010; Del Cortona et al. 2017, 2020). The transformation of unicellular→colonial→multicellular occurred at least once in the Ulvophyceae (fig. 4). We infer that the ancestor of the UUO clade was a unicellular organism, as the Scotinosphaerales, early divergent Oltmannsiellopsidales and Ignatius taxa were unicellular and colonial types. Multicellularity has evolved in the Ulvales-Ulotrichales clade, which contains complex morphological types, ranging from unicellular or colonial, multicellular to macroscopic siphonocladous (fig. 4). This transformation of microscopic to macroscopic algae (fig. 3g) mirrors the unicellular→colonial→multicellular theory (Niklas 2014; Del Cortona and Leliaert 2018), which hypothesized that cell–cell simple adhesive (colonial), complex cell–cell communication, and indeterminate growth (multicellular) developed from the co-option of genes in the unicellular ancestor.

Genetic Code Evolution in the Ulvophyceae

Our phylogenomic inference allowed us to shed light on genetic code evolution in the Ulvophyceae. By analyzing publicly available and our newly sequenced transcriptomes, we detected a noncanonical genetic code in the Trentepohliales and Cladophorales with TAR (Stop→Gln) codon reassignment in genes (table 2, fig. 3, and supplementary figs. S38–S41, Supplementary Material online) as reported by Gile et al. (2009) and Cocquyt et al. (2010). However, these authors did not sample the Scotinosphaerales. In our study, we detected TAG (Stop→Ser) codon reassignment in the actin and GPI genes of Scotinosphaera. TAG is used exclusively by Scotinosphaerales at positions where serine is predominant in other Ulvophyceae (fig. 3 and supplementary fig. S50, Supplementary Material online). Furthermore, we investigated all of the 30 ulvophycean species in the 3,690 OGs,
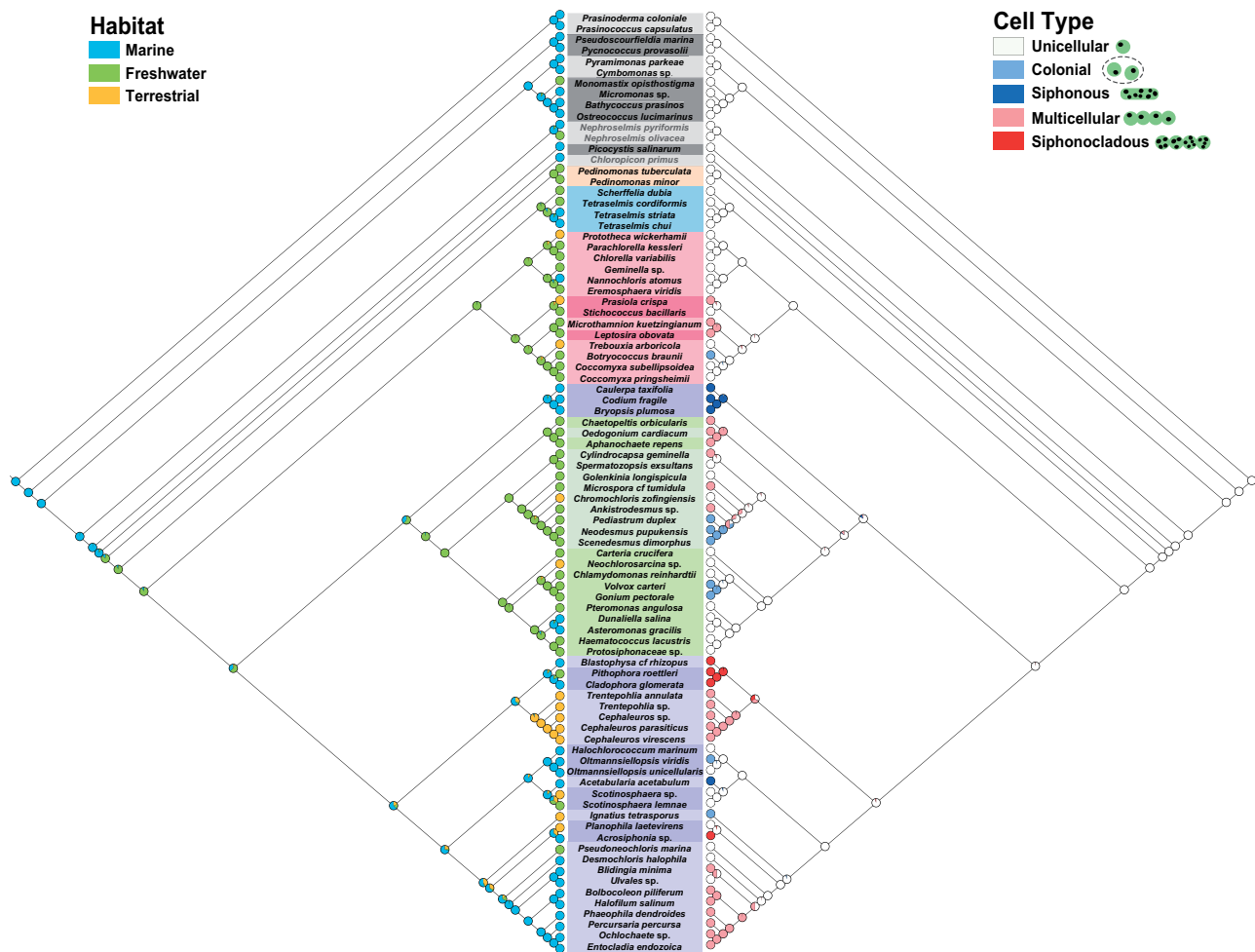
FIG. 4.—The inferred ancestral character reconstruction of habitats (left) and cell types (right) based on a likelihood method and plotted on the tree in supplementary figure S38, Supplementary Material online.

and reinforced our hypotheses of the noncanonical genetic codes comparing to the previous studies (Gile et al. 2009; Cocquyt et al. 2010; Del Cortona et al. 2020). Noncanonical nuclear genetic codons are universally present in the nuclear genes of Scotinosphaerales, Dasycladales, and Trentepohliales + Cladophorales + *Blastophysa* clade (fig. 3 and supplementary fig. S50, Supplementary Material online).

This appears to be an alternative scenario of stop-to-sense codon reassignment, compared with Cocquyt et al. (2010), in which Scotinosphaerales was missing and Bryopsidales + Dasycladales was close to Cladophorales + *Blastophysa* clade. Based on our phylogenetic tree in figure 3g, we infer that the stop-to-sense codon reassignments in the Ulvophyceae are likely to have originated independently at least three times: 1) in the ancestor of the monophyletic Trentepohliales + Cladophorales + *Blastophysa* clade with TAR (Stop→Gln) codon reassignment; 2) in Dasycladales with TAG (Stop→Gln) codon reassignment; and 3) in Scotinosphaerales with TAG (Stop→Ser) codon reassignment (fig. 3g).

Materials and Methods

Taxon Sampling and Culture Conditions

We newly sequenced the transcriptomes of seven species (six in the Ulvophyceae and *Protosiphonaceae* sp. in the Chlorophyceae). *Blidingia minima* (NIES-1837) (Hamana et al. 2013), *Ulvales* sp. (as "*Halochlorococcum*" sp. NIES-1838), *Protosiphonaceae* sp. (as "*Pseudendoclonium*" sp. NIES-2501), and *Scotinosphaera* sp. (NIES-154) were obtained from National Institute of Environmental Studies in Tsukuba (NIES), Tsukuba, Japan. *Cephaleuros* sp. HZ-2017, *P. roettleri*, and *Trentepohlia* sp. were sourced from the Freshwater Algae Culture Collection at the Institute of Hydrobiology (FACHB-collection), Chinese Academy of Sciences. All strains were cultured at 20°C under a 12 h:12 h L:D period at a photon flux density of 30 μmol photons m^{-2} s^{-1}.

Data Sources and Transcriptome Assembly

The new transcriptome data were obtained using Illumina nova-seq PE150 (2 × 150 bp) technologies. Library construction and

sequencing were performed at Novogene Bioinformatics Technology Co., Ltd (Beijing, China). Clean reads were obtained according to the following steps: firstly, we removed Illumina adapters; then we removed reads containing more than 10% ploy-N; lastly, we removed reads containing more than 50% low-quality bases (Q ≤ 20). The contigs were assembled de novo with Trinity (Grabherr et al. 2011) using a kmer value of 25, and clustered using Corset (Davidson and Oshlack 2014) with default parameters. The longest contig in each cluster was chosen as unigene for further analyses. In the 70 taxa analyses, we retrieved 8 additional whole genomes and 53 transcriptomes in Chlorophyta mainly from the Phytozome v 12 (http://phytozome.jgi.doe.gov/pz/portal.html, last accessed February 10, 2021) and the 1KP Project (Matasci et al. 2014; One Thousand Plant Transcriptomes Initiative 2019) (http://www.onekp.com, last accessed February 10, 2021). Transcriptomes of *Caulerpa taxifolia* and *Oltmansiellopsis unicellularis* were retrieved from Ranjan et al. (2015) and Cooper and Delwiche (2016), respectively. The prasinophytes *Ostreococcus lucimarinus* and *Micromonas* sp. were designated as outgroups and used to root the phylogeny in the 70 taxa analyses. To assess the sensitivity of the outgroup sampling and integrate recently published data of four ulvophycean species (*Cephaleuros parasiticus*, *Phaeophila dendroides*, *Acetabularia acetabulum*, and *Scotinosphaera lemnae*) (Del Cortona et al. 2020), we increased taxon sampling to 86 species and 14 prasinophytes species were designated as outgroups, including Chloropicophyceae and Picocystophyceae, Nephroselmidophyceae, Mamiellophyceae, Pyramimonadophyceae, Pseudoscourfieldiales, and Palmophyllophyceae. The details of the 86 species were described in supplementary table S3, Supplementary Material online.

Selection of Orthologs

Data from eight published whole genomes of Chlorophyceae (*Chlamydomonas reinhardtii*, *Chromochloris zofingiensis*, *Dunaliella salina*, and *Volvox carteri*), Trebouxiophyceae (*Chlorella variabilis* and *Coccomyxa subellipsoidea*), and Prasinophytes (*O. lucimarinus* and *Micromonas* sp.) were used to identify putative single-copy orthologs in Chlorophyta using OrthoMCL v2.0.9 (Li et al. 2003), with default parameters. This resulted in a data set of 13,790 clusters of putative orthologs. We sorted 5,926 clusters of single-copy genes present in at least four of eight species, and then selected 3,960 single-copy genes that existed in both *C. reinhardtii* (Chlorophyceae) and *C. subellipsoidea* (Trebouxiophyceae). The 3,960 OGs were further used as references in Orthograph v0.6.1 (Petersen et al. 2017) with default settings to obtain orthologs from genomic/transcriptomic data of 84 other taxa.

Alignment Processing and Trimming

Amino acid sequences of the 3,960 OGs were aligned using MAFFT v7.310 (Katoh and Standley 2013) with the L-INS-I algorithm. Alignments were trimmed using TrimAl v1.2 (Capella-Gutiérrez et al. 2009) by three steps: 1) the low-quality and ambiguous alignment regions were removed using a heuristic method (-automated1), which can automatically select between the gappyout method (consider gap scores distribution only) and strict method (consider both gap and similarity scores distribution), depending on the average identity score and the number of sequences in a given alignment; 2) to examine the impact of missing data on species tree inference, we further applied distinct trimming strategies using TrimAl prior to tree reconstruction, as column occupancy cutoffs of 0.5 (relaxed) or 0.8 (strict) with -g 0.5 or 0.8; 3) we removed short and empty sequences with -resoverlap 0.5 -seqoverlap 50. To explore the impacts of variation in taxon occupancy, we set cutoffs as 50% (≥ 36 or 43 species) or 80% (≥ 57 or 68 species) out of 70 or 86 species to each subset using Geneious v10.2.3 (http://www.geneious.com/, last accessed February 10, 2021).

Phylogenetic Inference

Amino acid sequences of 2,698, 1,984, 1,447, and 1,404 OGs were separately used in ASTRAL- and concatenation-based species tree estimation based on 70 taxa. Additional concatenation-based analyses were conducted based on 86 taxa. Substitution models were chosen by IQ-TREE v1.5.5 (Nguyen et al. 2015). The gene trees were reconstructed using the site-homogeneous PROTGAMMALG model with RAxML v8.2.11 (Stamatakis 2006) and the site-heterogeneous LG+C20+F + G model with IQ-TREE v1.5.5 for all alignments from each of the four data sets (table 1), and 100 random replicates were performed to calculate bootstrap value. To minimize potential impacts of gene-tree estimation error on multi-locus-based species tree estimation (Mirarab and Warnow 2015), we collapsed low support branches (i.e., BS <10%/20%) from the gene trees by Newick utilities (htpp://cegg.unige.ch/newick_utils). The multi-locus-based species trees with local PP and BS (100 replicates each), and polytomy test among Chlorodendrophyceae, Pedinophyceae, and the UTC clade were implemented in ASTRAL III v5.5.9 (Zhang et al. 2018). The concatenation-based species trees were reconstructed using site-homogeneous LG+F + G, site-heterogeneous LG+C20+F + G, and heterotachous GHOST LG+F+H4 models implemented in IQ-TREE 1.5.5 with ultrafast bootstrapping (UFboot) (Minh et al. 2013) and SH-aLRT (Guindon et al. 2010) testing supports (1,000 replicates each).

Gene tree concordance and conflicts against the reference tree (i.e., the ASTRAL-based species tree from 2,698 genes trees of 70 taxa based on a site-heterogeneous model, with BS ≥ 20%) were quantified using PhyParts (bitbucket.com/

blackrim/phyparts; accessed July 12, 2019), with both 10% and 20% BS threshold. We rooted all the gene trees from different data sets with *O. lucimarinus*, *Micromonas* sp., or both, using root. R (https://github.com/lixi8507/core-Chlorophyta, last accessed February 10, 2021). Gene trees that lack either *O. lucimarinus* or *Micromonas* sp. were removed from the PhyParts analyses. We summarized and visualized bipartition support using PhyPartsPieCharts.py (https://github.com/mossmatters/phyloscripts/tree/master/phypart-spiecharts, last accessed February 10, 2021) and ETE3 Python package (http://etetoolkit.org./). The pairwise RF distances (Robinson and Foulds 1981) between species trees based on subsets and the reference species tree were calculated by the RFdistances.twoFiles.v2.py script from the RF Distances Filter (Simmons et al. 2016) (https://github.com/dbsloan/rf_distances_filter, last accessed February 10, 2021). The normalized RF distances were calculated by dividing the plain RF distance by $2(n - 3)$, where n represents the number of taxa.

Ancestral State Reconstruction

The habitats and cell types of 86 species of Chlorophyta were recoded as follows: Habitat: marine, freshwater, and terrestrial; Cell type: unicellular, colonial, siphonous, multicellular, and siphonocladous. The concatenation-based phylogeny from 2,629 OGs of 86 taxa with the site-heterogeneous model was used to guide the ancestral state reconstruction. The ancestral state of habitat and cell type was reconstructed with Mesquite v3.61 (http://www.mesquiteproject.org, last accessed February 10, 2021) (Maddison and Maddison 2019) using Tracing Character History option based on a likelihood method, which find the ancestral states at different nodes with maximum probability (Schluter et al., 1997; Pagel, 1999). One-parameter Mk1 model was used to estimate the evolutionary rate of categorical characters, in which the forward and backward rates between states are equal (Lewis 2001).

Genetic Codes

We downloaded reference sequences of 10 nuclear genes (the Elongation Factor-1 Alpha [1 alpha]/Elongation Factor-Like [EFL], actin, GPI, GAPA, histone H3, OEE1, 40 S ribosomal protein S9, and 60 S ribosomal proteins L3 and L17) from different ulvophycean species that contain both noncanonical and canonical genetic codes in Ulvophyceae (Gile et al. 2009; Cocquyt et al. 2010) from NCBI (https://www.ncbi.nlm.nih.gov/, last accessed February 10, 2021). To investigate the existence of noncanonical genetic codes in 10 species of Trentepohliales, Oltmannsiellopsidales, Cladophorales, and Scotinosphaerales, the additional transcriptomes were blasted against the 10 genes reference database by BLASTN v2.7.1+ (Altschul et al. 1997) with *E*-value cutoff of 10^{-10}. The sequences with best BLAST hit were aligned using translation align tool in Geneious v10.2.3. To reinforce our hypothesis of the noncanonical genetic codes, we investigated all of the 30

ulvophycean species in the nucleotide sequence of 3,690 OGs generated in Orthograph v0.6.1. Orthograph executes pHMM-based and BLAST reciprocal search to target ortholog genes and it infers frameshift-corrected open reading frame by using Exonerate (Slater and Birney 2005). All of the nucleotide sequences of the 3,690 OGs were aligned by TranslatorX (Abascal et al. 2010) with settings "-p M -t T -w 1 -c 1." Noncanonical nuclear genetic codes were widely detected by screening the alignments in Geneious. These alignments were deposited in the Figshare: https://figshare.com/s/17b87d35616f7ee2e7ff. Noncanonical nuclear genetic TAG/TAA codons were found in some taxa at positions coding for glutamine by canonical nuclear genetic CAG/CAA codons in the remaining ulvophycean species. We also detected that many genes in Scotinosphaerales used TAG codon at a conserved position of serine (TCC/TCG).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Data Availability

The raw sequence reads are deposited in NCBI Short Read Archive. Individual accessions numbers are shown in supplementary table S3, Supplementary Material online. The sequence alignments are available from the Figshare: https://figshare.com/s/17b87d35616f7ee2e7ff

Literature Cited

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequence guided by amino acid translations. Nucleic Acids Res. 38(Web Server issue):W7–W13.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389–3402.

Baudelet P, Ricochon G, Linder M, Muniglia L. 2017. A new insight into cell walls of Chlorophyta. Algal Res Biomass Biofuels Bioprod. 25:333–371.

Becker B. 2013. Snow ball earth and the split of Streptophyta and Chlorophyta. Trends Plant Sci. 18(4):180–183.

Bisson MA, Beilby MJ, Shepherd VA. 2006. Electrophysiology of turgor regulation in marine siphonous green algae. J Membr Biol. 211(1):1–14.

Blom M, Bragg JG, Potter S, Moritz C. 2017. Accounting for uncertainty in gene tree estimation: summary-coalescent species tree inference in a challenging radiation of Australian lizards. Syst Biol. 66(3):352–366.

Borowiec ML, et al. 2019. Compositional heterogeneity and outgroup choice influence the internal phylogeny of the ants. Mol Phylogenet Evol. 134:111–121.

Capella-Gutiérrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973.

Cocquyt E, et al. 2010. Complex phylogenetic distribution of a non-canonical genetic code in green algae. BMC Evol Biol. 10:327.

Cocquyt E, Verbruggen H, Leliaert F, De Clerck O. 2010. Evolution and cytological diversification of the green seaweeds (Ulvophyceae). Mol Biol Evol. 27(9):2052–2061.

Cooper, E, Delwiche, C. 2016. Green algal transcriptomes for phylogenetics and comparative genomics. Figshare. doi:10.6084/m9.figshare.1604778. Accessed March 18, 2019.

Cox CJ, Li B, Foster PG, Embley TM, Civán P. 2014. Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. Syst Biol. 63(2):272–279.

Crotty SM, et al. 2020. GHOST: recovering historical signal from heterotachously evolved sequence alignments. Syst Biol. 69(2):249–264.

Davidson NM, Oshlack A. 2014. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. Genome Biol. 15(7):410.

De Clerck O, Bogaert KA, Leliaert F. 2012. Diversity and evolution of algae: primary endosymbiosis. Adv Bot Res. 64:55–86.

de la Torre-Bárcena JE, et al. 2009. The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data. PLoS One 4(6):e5764.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2(5):e68.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol. 24(6):332–340.

Del Cortona A, et al. 2017. The plastid genome in Cladophorales green algae is encoded by hairpin plasmids. Curr Biol. 27(24):3771–3782.

Del Cortona A, et al. 2020. Neoproterozoic origin and multiple transitions to macroscopic growth in green seaweeds. Proc Natl Acad Sci U S A. 117(5):2551–2559.

Del Cortona A, Leliaert F. 2018. Molecular evolution and morphological diversification of ulvophytes (Chlorophyta). Perspect Phycol. 5(1):27–43.

Dittami SM, Heesch S, Olsen JL, Collen J. 2017. Transitions between marine and freshwater environments provide new clues about the origins of multicellular plants and algae. J Phycol. 53(4):731–745.

Edwards SV. 2016. Phylogenomic subsampling: a brief review. Zool Scr. 45(S1):63–74.

Edwards SV, et al. 2016. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Mol Phylogenet Evol. 94(Pt A):447–462.

Fang L, et al. 2018. Improving phylogenetic inference of core Chlorophyta using chloroplast sequences with strong phylogenetic signals and heterogeneous models. Mol Phylogenet Evol. 127:248–255.

Fang L, Leliaert F, Zhang Z, Penny D, Zhong B. 2017. Evolution of the Chlorophyta: insights from chloroplast phylogenomic analyses. J Syst Evol. 55(4):322–332.

Fučíková K, et al. 2014. New phylogenetic hypotheses for the core Chlorophyta based on chloroplast sequence data. Front Ecol Evol. 2:63.

Fučíková K, Lewis LA, Lewis PO. 2016. Comparative analyses of chloroplast genome data representing nine green algae in Sphaeropleales (Chlorophyceae, Chlorophyta). Data Brief. 7:558–570.

Fučíková K, Lewis PO, Neupane S, Karol KG, Lewis LA. 2019. Order, please! Uncertainty in the ordinal-level classification of Chlorophyceae. PeerJ 7:e6899.

Gage DA, et al. 1997. A new route for synthesis of dimethylsulphoniopropionate in marine algae. Nature 387(6636):891–894.

Gile GH, Novis PM, Cragg DS, Zuccarello GC, Keeling PJ. 2009. The distribution of Elongation Factor-1 Alpha (EF-1alpha), Elongation Factor-Like (EFL), and a non-canonical genetic code in the ulvophyceae: discrete genetic characters support a consistent phylogenetic framework. J Eukaryot Microbiol. 56(4):367–372.

Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29(7):644–652.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59(3):307–321.

Hamana K, Niitsu M, Hayashi H. 2013. Occurrence of homospermidine and thermospermine as a cellular polyamine in unicellular chlorophyte and multicellular charophyte green algae. J Gen Appl Microbiol. 59(4):313–319.

Herron MD, Hackett JD, Aylward FO, Michod RE. 2009. Triassic origin and early radiation of multicellular volvocine algae. Proc Natl Acad Sci U S A. 106(9):3254–3258.

Irisarri I, et al. 2017. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat Ecol Evol. 1(9):1370–1378.

Jiang X, Edwards SV, Liu L. 2020. The multispecies coalescent model outperforms concatenation across diverse phylogenomic data sets. Syst Biol. 69(4):795–812.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 7(Suppl 1):S4.

Leliaert F. 2019. Green algae: Chlorophyta and Streptophyta. In: Schmidt TM, editor. Encyclopedia of microbiology. 4th ed. Burlington (MA): Academic Press. p. 457–468.

Leliaert F, et al. 2012. Phylogeny and molecular evolution of the green algae. Crit Rev Plant Sci. 31(1):1–46.

Lemieux C, Vincent AT, Labarre A, Otis C, Turmel M. 2015. Chloroplast phylogenomic analysis of chlorophyte green algae identifies a novel lineage sister to the Sphaeropleales (Chlorophyceae). BMC Genomics 15(1):264.

Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst Biol. 58(1):130–145.

Lewis PO. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst Biol. 50(6):913–925.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13(9):2178–2189.

Liu L, Edwards SV. 2009. Phylogenetic analysis in the anomaly zone. Syst Biol. 58(4):452–460.

Liu L, Xi Z, Davis CC. 2015. Coalescent methods are robust to the simultaneous effects of long branches and incomplete lineage sorting. Mol Biol Evol. 32(3):791–805.

Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol Biol. 10:302.

Liu Y, Cox CJ, Wang W, Goffinet B. 2014. Mitochondrial phylogenomics of early land plants: mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias. Syst Biol. 63(6):862–878.

Maddison WP, Maddison DRV. 2019. Mesquite:a modular system for evolutionary analysis. Version 3.61. 2019. 12. 26. Available from: http://www.mesquiteproject.org. Accessed March 2, 2021.

Marin B. 2012. Nested in the Chlorellales or independent class? Phylogeny and classification of the Pedinophyceae (Viridiplantae) revealed by molecular phylogenetic analyses of complete nuclear and plastid-encoded rRNA operons. Protist 163(5):778–805.

Matasci N, et al. 2014. Data access for the 1,000 Plants (1KP) project. Gigascience 3:17.

Melton JT, Leliaert F, Tronholm A, Lopez-Bautista JM. 2015. The complete chloroplast and mitochondrial genomes of the green macroalga Ulva sp. UNA00071828 (Ulvophyceae, Chlorophyta). PLoS One 10(4):e0121020.

Mine I, Menzel D, Okuda K. 2008. Morphogenesis in giant-celled algae. Int Rev Cell Mol Biol. 266:37–83.

Minh BQ, Nguyen MAT, Von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol. 30(5):1188–1195.

Mirarab S, et al. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30(17):i541–i548.

Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31(12):i44–i52.

Nakada T, Misawa K, Nozaki H. 2008. Molecular systematics of Volvocales (Chlorophyceae, Chlorophyta) based on exhaustive 18S rRNA phylogenetic analyses. Mol Phylogenet Evol. 48(1):281–291.

Nguyen L, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol Biol Evol. 32(1):268–274.

Niklas KJ. 2014. The evolutionary-developmental origins of multicellularity. Am J Bot. 101(1):6–25.

Novis PM, Smissen RD, Buckley TR, Gopalakrishnan K, Visnovsky G. 2013. Inclusion of chloroplast genes that have undergone expansion misleads phylogenetic reconstruction in the Chlorophyta. Am J Bot. 100(11):2194–2209.

O'Kelly CJ. 2007. The origin and early evolution of green plants. In: Falkowski PG, Knoll AH, editors. Evolution of primary producers in the sea. Vol. 13. Burlington (MA): Elsevier Academic Press. p. 287–309.

One Thousand Plant Transcriptomes Initiative 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574(7780):679–685.

Pagel M. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Syst Biol. 48(3):612–622.

Parks M, Wickett NJ, Alverson AJ. 2018. Signal, uncertainty, and conflict in phylogenomic data for a diverse lineage of microbial eukaryotes (Diatoms,Bacillariophyta). Mol Biol Evol. 35(1):80–93.

Petersen M, et al. 2017. Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. BMC Bioinformatics 18(1):111.

Philippe H, et al. 2017. Pitfalls in supermatrix phylogenomics. Eur J Taxon. 283:1–25.

Pierrehumbert RT, Abbott DS, Voigt A, Koll DDB. 2011. Climate of the Neoproterozoic. Annu Rev Earth Planet Sci. 39(1):417–460.

Prave AR, Condon DJ, Hoffmann KH, Tapster S, Fallick AE. 2016. Duration and nature of the end-Cryogenian (Marinoan) glaciation. Geology 44(8):631–634.

Quang LS, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24(20):2317–2323.

Ranjan A, Townsley BT, Ichihashi Y, Sinha NR, Chitwood DH. 2015. An intracellular transcriptomic atlas of the giant coenocyte Caulerpa taxifolia. PLoS Genet. 11(1):e1004900.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Math Biosci. 53(1–2):131–147.

Roch S, Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor Popul Biol. 100C:56–62.

Rooney AD, et al. 2014. Re-Os geochronology and coupled Os-Sr isotope constraints on the Sturtian snowball Earth. Proc Natl Acad Sci U S A. 111(1):51–56.

Roure B, Baurain D, Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol Biol Evol. 30(1):197–214.

Sayyari E, Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol Biol Evol. 33(7):1654–1668.

Schluter D, Price T, Mooers AO, Ludwig D. 1997. Likelihood of ancestor states in adaptive radiation. Evolution. 51(6):1699–1711.

Simmons MP, Sloan DB, Gatesy J. 2016. The effects of subsampling gene trees on coalescent methods applied to ancient divergences. Mol Phylogenet Evol. 97:76–89.

Slater G, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6(1):31.

Smith DR, et al. 2011. The GC-rich mitochondrial and plastid genomes of the green alga Coccomyxa give insight into the evolution of organelle DNA nucleotide landscape. PLoS One 6(8):e23624.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22(21):2688–2690.

Stewart KD, Mattox KR. 1975. Some aspects of mitosis in primitive green algae: phylogeny and function. Biosystems 7(3–4):310–315.

Sun L, et al. 2016. Chloroplast phylogenomic inference of green algae relationships. Sci Rep. 6:20528.

Turmel M, de Cambiaire JC, Otis C, Lemieux C. 2016. Distinctive architecture of the chloroplast genome in the Chlorodendrophycean green algae Scherffelia dubia and Tetraselmis sp. CCMP 881. PLoS One 11(2):e148934.

Turmel M, Lemieux C. 2018. Evolution of the plastid genome in green algae. Adv Bot Res. 85:157–193.

Turmel M, Otis C, Lemieux C. 2015. Dynamic evolution of the chloroplast genome in the green algal classes Pedinophyceae and Trebouxiophyceae. Genome Biol Evol. 7(7):2062–2082.

Turmel M, Otis C, Lemieux C. 2017. Divergent copies of the large inverted repeat in the chloroplast genomes of ulvophycean green algae. Sci Rep. 7(1):994.

Wilberg EW. 2015. What's in an outgroup? The impact of outgroup choice on the phylogenetic position of Thalattosuchia (Crocodylomorpha) and the origin of Crocodyliformes. Syst Biol. 64(4):621–637.

Xi Z, Liu L, Davis CC. 2016. The impact of missing data on species tree estimation. Mol Biol Evol. 33(3):838–860.

Xi Z, Liu L, Rest JS, Davis CC. 2014. Coalescent versus concatenation methods and the placement of Amborella as sister to water lilies. Syst Biol. 63(6):919–932.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19(S6):153.

Zhang C, Sayyari E, Mirarab S. 2017. ASTRAL-III: increased scalability and impacts of contracting low support branches. In: Meidanis J, Nakhleh L, editors. RECOMB international workshop on comparative genomics. London: Springer. p. 53–75.