Chloroplast phylogenomic data from the green algal order Sphaeropleales (Chlorophyceae, Chlorophyta) reveal complex patterns of sequence evolution.

*Karolina Fučíková¹, Paul O. Lewis¹, and Louise A. Lewis¹

¹ Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs CT, USA

*Author for Correspondence: Karolina Fučíková, Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs CT, USA, 860-486-1811, 860-486-6364, karolina.fucikova@uconn.edu

Abstract

Chloroplast sequence data are widely used to infer phylogenies of plants and algae. With the increasing availability of complete chloroplast genome sequences, the opportunity arises to resolve ancient divergences that were heretofore problematic. On the flip side, properly analyzing large multi-gene data sets can be a major challenge, as these data may be riddled with systematic biases and conflicting signals. Our study contributes new data from nine complete and four fragmentary chloroplast genome sequences across the green algal order Sphaeropleales. Our phylogenetic analyses of a 56-gene data set show that analyzing these data on a nucleotide level yields a well-supported phylogeny – yet one that is quite different from a corresponding amino acid analysis. We offer some possible explanations for this conflict through a range of analyses of modified data sets. In addition, we characterize the newly sequenced genomes in terms of their structure and content, thereby further contributing to the knowledge of chloroplast genome evolution.

Keywords: algae, plastid, phylogenetic conflict, codon bias

Abbreviations: aa – amino acid, cp – chloroplast, mt – mitochondrion/mitochondrial, nt – nucleotide

1. Introduction

Organellar genomes in the Viridiplantae exhibit great variation in structure, size, and content (e.g., Lemieux et al., 2014a, 2014b, Turmel et al., 2015). Until recently, such variation has been documented using one or a few representatives per major lineage – embryophytes being the obvious, intensively studied exception to this rule. In the last year however, three studies have characterized chloroplast (Lemieux et al., 2014a) and mitochondrial (Fučíková et al., 2014a, Farwagi et al., 2015) genome diversity using denser sampling within a green algal class, order, and family, respectively.

The chlorophycean order Sphaeropleales contains freshwater algae that morphologically range from simple unicells to colonies of various degrees of complexity. Compared to their sister order Volvocales, which has been used for investigations of the evolution of multicellularity and also contains the versatile model species *Chlamydomonas*, Sphaeropleales are still fairly poorly understood in terms of their diversity and evolution. Recent studies (Fučíková et al., 2014a, 2014b, Tippery et al., 2012) have struggled with resolving the phylogenetic relationships in Sphaeropleales with certainty, particularly the relationships among deeply diverging and

morphologically simple genera (e.g., *Bracteacoccus*, *Chromochloris*, *Mychonastes*, and *Pseudomuriella*, which were also examined in the present study). There appears to be conflicting signal not only among data sets from different cellular compartments (nuclear, mitochondrial, and chloroplast), but also among individual genes from the same compartment.

Despite these potential issues, using organellar, particularly chloroplast genes for phylogenetic inference generally shows promise for resolving deep relationships among green plant lineages. In terms of phylogenetic utility, chloroplast data have a distinct advantage over mitochondrial mainly due to a larger and more constant number of genes, but also due to an overall better taxon sampling among available data (e.g., Fučíková et al., 2014a, Lemieux et al., 2014a, 2014b, Turmel et al., 2013).

Progressively denser taxon sampling of both chloroplast and mitochondrial genomes combined with more sophisticated analysis methods have greatly improved our understanding of green plant evolution in the past few years (Fučíková et al., 2014c, Lemieux et al., 2014a, 2014b, Ruhfel et al., 2014, Turmel et al., 2013, 2015). Usually, some way of accounting for biases and problems in the nucleotide (nt) data is employed. The commonly adopted approach has been to analyze a concatenated nt data set of protein-coding genes, often curating the data to remove highly variable gene regions, 3rd codon positions, or other fast-evolving sites using a site-stripping approach (e.g., Fučíková et al., 2014c, Ruhfel et al., 2014). Another alternative is translating and analyzing the data on the amino acid (aa) level and thus bypassing the often-assumed saturation of nt data, particularly 3rd positions, but even the translation approach may not overcome all problems (Rota-Stabelli et al., 2013).

In this study we present nine new completely sequenced and four partially sequenced chloroplast genomes spanning the diversity of the green algal order Sphaeropleales. We describe the size, structure, content and gene order of the completely sequenced

genomes and phylogenetically analyze a data set of 56 protein-coding chloroplast genes to resolve deep-level evolutionary relationships in Sphaeropleales. The results of our nt and aa analyses differ in topology, which however is highly supported in both cases. We present some possible explanations and propose future directions for analyzing this and similar data sets.

2. Materials and Methods

2.1. Algal Culturing Conditions

Algal strains were obtained from the Culture Collection of Algae at the University of Texas at Austin (UTEX, <http://www.utex.org>), the Culture Collection of Algae at the University of Göttingen, Germany (SAG, <http://sagdb.uni-goettingen.de>), and the Culture Collection of Algae of Charles University in Prague (CAUP, <http://botany.natur.cuni.cz/algo/caup-list.html>). In addition to the strains listed in Table 1, partial genomic sequences were obtained for *Atractomorpha echinata* (UTEX 2309, GenBank KT369370 - KT369440), *Ourococcus multisporus* (UTEX 1240, GenBank KT369441 - KT369509), *Pediastrum duplex* var. *asperum* (UTEX LB1364; gene sequences available in Dryad; full cp genome will be published in a forthcoming study), and *Rotundella rotunda* (UTEX 2979, GenBank KT369317 - KT369369). The cultures were maintained under 16:8 light:dark cycle at 18°C and 70 $\mu\text{mol m}^{-2} \text{s}^{-1}$ in liquid media or on agar slants as specified by the culture collection of origin.

2.2. Genome Sequencing, Assembly, and Annotation

Genomic DNA was extracted using a PowerPlant DNA Isolation Kit (MO-BIO Laboratories, Carlsbad, CA), with a modified purification part of the protocol using chloroform separation and ethanol precipitation steps instead of column cleaning. DNA was subsequently shipped to Cold Spring Harbor Laboratories for TruSeq library preparation and sequencing on Illumina HiSeq2500, producing 2 x 100 bp paired reads. The reads were paired, trimmed, assembled, and annotated in Geneious v.R6 (Biomatters, www.geneious.com). For the large data sets of *Bracteacoccus minor* (46 million reads) and *Neochloris aquatica* (33 million reads), the program ABySS (Simpson et al., 2009) was used for initial assembly using a range of kmer sizes (25-50) to obtain the longest contigs possible. In most cases, a single cp contig was obtained from de novo assembly. In cases where multiple contigs were obtained (e.g., in *Chromochloris zofingiensis*), the contigs were subjected to a series of reference assemblies in Geneious (by mapping reads to the cp fragments and subsequently to the longer resulting fragments, often for 25 iterations or more) until we reached a point where the fragments could be confidently joined. The final contigs were verified by mapping paired reads to the consensus sequences and inspecting the resulting assemblies by eye for mismatches and unexpected drops in coverage. The coverage of the cp genomes varied from 40x to 2300x.

Annotation was aided by the combination of the following tools: DOGMA (Wyman et al., 2004, dogma.ccbb.utexas.edu/), RNAweasel (Lang et al., 2007), and BLAST (Altschul et al., 1990, 1997). Intron boundaries were determined based on alignments of the exon aa and nt sequences including sequences from other chlorophyceans. Intron conserved secondary structure elements were identified by RNAweasel, which also determined the intron subgroup affiliation. Sidedness index (C_S) was calculated according to Cui et al. (2006).

2.3. Phylogenetic analyses

2.3.1. Alignment and Pre-trimming

Chloroplast protein-coding genes were aligned using the default translation-aided alignment tool in Geneious. The large, hypervariable genes *ftsH*, *ycf1*, *rpoC1*, and *rps2* were not included in analyses, and only genes present in all ingroup and outgroup taxa were used for phylogenetic inference, rendering 56 protein-coding genes in the data set: *atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI*, *ccsA*, *cemA*, *clpP*, *petB*, *petD*, *petG*, *petL*, *psaA*, *psaB*, *psaC*, *psaJ*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbF*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*, *rbcL*, *rpl2*, *rpl5*, *rpl14*, *rpl16*, *rpl20*, *rpl23*, *rpl36*, *rpoA*, *rpoBa*, *rpoBb*, *rpoC2*, *rps3*, *rps4*, *rps7*, *rps8*, *rps9*, *rps11*, *rps12*, *rps14*, *rps18*, *rps19*, *tufA*, *ycf3*. Additionally, regions (codon positions) of uncertain homology were identified by eye and removed using an in-house python script (Dryad accession #XXX), which simultaneously produced a concatenated alignment. These pre-trimmed nt alignments were also translated by the same script to set up a corresponding set of single-gene and concatenated aa analyses.

2.3.2. Nucleotide Analyses

Individual genes were analyzed separately in addition to analyses of the concatenated data set. All nt analyses were partitioned by codon position and were conducted in the Bayesian (using MrBayes; Huelsenbeck and Ronquist, 2001, Ronquist and Huelsenbeck, 2003) and Maximum Likelihood (ML) framework (using RAxML; Stamatakis, 2014). The GTRGAMMAI (GTR+I+d Γ_4) model was

implemented in all nt analyses. Internode certainty (IC, Salichos et al., 2014) was calculated from the individual gene trees and mapped onto a Majority Rule-Extended (MRE) tree using RAxML. Gene support values (GSF), indicating what percentage of single-gene ML analyses supports a particular MRE branch, were calculated using PAUP* (Swofford, 2003) from ML best trees resulting from individual gene analyses.

MrBayes was run for 5,000,000 generations, sampling and printing every 500. Two independent MCMC runs using 4 chains (with the default heating schedule) were conducted per Bayesian analysis. Branch support was calculated from the posterior distribution of Bayesian trees after discarding the first 25% of the trees as burnin, and from 1000 ML bootstrap pseudoreplicates. A separate set of nucleotide analyses for individual genes and the concatenated data set was conducted in MrBayes using a codon model (nucmodel=codon). In addition, to account for potential heterotachy effects, the pre-trimmed concatenated nt data set was analyzed using a covariotide model in MrBayes with the remaining conditions set as specified above.

2.3.3. Amino Acid Analyses

Amino acid analyses were conducted using RAxML and MrBayes, implementing the LGF model of protein sequence evolution with gamma-distributed rates in RAxML (PROTGAMMALGF), as selected using RAxML in conjunction with the Perl script from <http://sco.h-its.org/exelixis/resource/download/software/ProteinModelSelection.pl>. In MrBayes, the aamodelpr=mixed and aamodelpr=GTR with rates=invgamma, and a covarion model were implemented in separate sets of analyses. Internode certainty (IC,

Salichos et al., 2014) was calculated from the individual gene trees and mapped onto the MRE tree using RAxML. Additionally, we investigated bias in codon usage for different amino acids using Kullback-Leibler (KL, Kullback and Leibler, 1951).

2.3.4. Site-Stripping and Ambiguous Recoding

Three site-removal approaches were used on the pre-trimmed concatenated data set to investigate the influence of potential saturation of fast-evolving sites and regions. Firstly, the fast-evolving genes *rpoA*, *rpoBa*, *rpoBb*, *rpoC2*, and *rps3* were removed from the data set (*rpoC1* was not included in any analyses), following the cautionary suggestions of Novis et al. (2013). To test whether this modification removes the conflict between nt and aa topologies, this data set was analyzed on nt and aa level using MrBayes. Secondly, 3rd codon positions were removed entirely from the pre-trimmed nt data set, which was subsequently analyzed using MrBayes. Thirdly, a series of stripped data sets was generated from the pre-trimmed concatenated alignment by progressively removing 10-90% of the fastest evolving sites using the CPO metric (Lewis et al., 2014). Fourthly, a data set was created with arginine, leucine, and serine codons made ambiguous to eliminate potential biases associated with synonymous codon usage, as documented by Cox et al (2014): Arg codons were recoded as MGN, Leu as YTN, and Ser as WSN. All modified data sets were then analyzed using MrBayes as described above.

3. Results

3.1. Chloroplast genome size, structure and content

The plastid genomes of Sphaeropleales ranged from 102.7 kb to 220.4 kb and contained inverted repeats (IR) between 6.5 and 35.5 kb in size (Table 1, Fig. 1). The IR region consistently contained 3 rRNA genes (*rrf*, *rrl*, and *rrs*) and two tRNA genes (*trnA(ugc)* and *trnI(gau)*). The *psbA* gene was present in the IR in 6 of the fully sequenced genomes, as was *trnS(gcu)*. *Neochloris aquatica* harbored the most genes in the IR region: *rrf*, *rrl*, *rrs*, *trnA(ugc)*, *trnI(gau)*, *trnS(gcu)*, *psbA*, *atpH*, *rpl12*, and part of each *atpF* and *cemA*. The number of introns in the IR region varied from 0 to 12 and the largest IR (*Kirchneriella aperta*, 35.5 kb) contained the most introns, present in *rrs* (1), *rrl* (7), and *psbA* (4).

The same suite of protein-coding genes was present in all taxa, with the exception of *Neochloris*, which lacked *chlB*, *chlL*, and *chlN*. Only in the previously published *Acutodesmus* genome, the *rps2* gene was present in two separate fragments (open reading frames), and was intact in the remaining taxa.

3.2. Phylogenetic analyses

3.2.1. Nucleotide Analyses

The pre-trimmed data set contained 35,712 characters and 23 chlorophycean taxa (14 Sphaeropleales, 5 Volvocales, 2 Chaetophorales and one of each Chaetopeltidales and Oedogoniales) and had 0.43% missing data (indels or incompletely sequenced genes in some taxa). All alignments (untrimmed, pre-trimmed, and further modified for supporting analyses) as well as their resulting trees are available in Dryad.

In all concatenated analyses (but not always in the single-gene analyses), the main groups of Chlorophyceae were recovered as monophyletic: Sphaeropleales (ingroup), Volvocales (*Chlamydomonas*, *Dunaliella*, *Gonium*, *Pleodorina*, and *Volvox*), and the OCC clade (Oedogoniales - *Oedogonium*, Chaetopeltidales - *Floydiella*, and Chaetophorales – *Schizomeris* and *Stigeoclonium*). Relationships among outgroup taxa were consistent with the current understanding of Chlorophyceae (Fig. 2), with Chaetophorales and Chaetopeltidales as sister groups within OCC, and *Dunaliella* sister to a monophyletic Volvocaceae (as recovered e.g., in Lemieux et al., 2014a). The concatenated nt analyses (ML and Bayesian) partitioned by codon position yielded fully resolved and well supported trees. The ML best tree topology was identical to the Bayesian consensus tree (shown in Fig. 2). The MRE tree calculated from individual gene trees generated by RAxML nt analyses was similar in topology to the concatenated topology, with the exception of the position of *Mychonastes homosphaera*, which (in the MRE tree) was sister to *Kirchneriella* and *Ourococcus*, instead of being the second-deepest diverging lineage in Sphaeropleales. In this majority rule tree, IC and GSF were low for most nodes (Fig. 3).

3.2.2. Amino Acid Analyses

The concatenated aa analyses yielded consistent topologies (Bayesian consensus tree for MrBayes and best ML tree for RAxML) with all nodes receiving high statistical support. The GTR and mixed aa models in MrBayes yielded identical topologies and support (mixed model results are represented in Fig. 2), as did the analysis under the covarion model. With one exception (marked with an asterisk in Fig. 2), the Bayesian codon analysis yielded a consensus tree of an identical topology as the aa analyses, but with much lower support for several nodes within Sphaeropleales. These nodes generally were ones that were inconsistent with the nt topology.

The MRE calculated from individual gene trees generated by RAxML aa analyses was similar in topology to the nt MRE tree in that *Mychonastes homosphaera* was sister to *Kirchneriella* and *Ourococcus*. The MRE tree resembled the topology yielded by the concatenated aa analyses in placing *Bracteacoccus* as the second deepest-diverging sphaeroplealean lineage. Internode certainty and GSF were low for most nodes of the MRE tree (Fig. 3).

3.2.3. Site Stripping and Ambiguous Recoding

Removal of *rpoA*, *rpoBa*, *rpoBb*, *rpoC2*, and *rps3* from the pre-trimmed data set did not affect the inferred relationships among Sphaeropleales on either the nt or the aa level and the conflict between the nt and aa topologies persisted (data not shown). Stripping 3rd positions from the pre-trimmed data set resulted in a topology that was different from either of the topologies presented in Fig. 2. The sister relationship of *Chlorotetraedron* and *Neochloris* was consistent with the aa analyses, as was the deep divergence of *Bracteacoccus*, but the placements and relative relationships of *Chromochloris*, *Mychonastes*, and *Pseudomuriella* were unlike either nt or aa topologies with (*Chromochloris* + *Pseudomuriella*) being the third-deepest divergence in Sphaeropleales after (*Ankyra* + *Atractomorpha*) and *Bracteacoccus*, and *Mychonastes* being sister to the clade containing *Chlorotetraedron*, *Neochloris*, *Pediastrum*, *Acutodesmus*, *Rotundella*, *Kirchneriella*, and *Ourococcus*. All nodes in the tree were highly supported (Supplementary materials).

The purpose of our site-stripping experiment was to determine whether the fastest-evolving sites are responsible for the differences between nt and aa topologies (Fig. 2). Progressive site stripping yielded a variety of topologies that on average became less and less well supported with increasing site removal. The placement of *Mychonastes* varied markedly among the stripped data sets,

further indicating the problematic nature of this taxon.. In general we may conclude that stripping the fastest nucleotides does not result in a topology more similar to the aa topology, but only decreases overall branch support. The analysis set with ambiguated Arg, Leu, and Ser codons yielded a fully supported tree that was nearly identical to the amino acid tree, with the exception of the sister relationship between *Neochloris* and *Pediastrum*. All alignments and resulting trees are available in the Dryad repository.

4. Discussion

4.1. Chloroplast Genome Size, Structure and Content

Despite finding considerable variation in chloroplast genome size in the order Sphaeropleales, our study concludes that this variation is much smaller compared to other chlorophycean groups – the order Volvocales and the OCC clade (Fig. 1, Table 1, Bélanger et al., 2006, Brouard et al., 2008, 2010, 2011, de Cambiaire et al., 2006, Maul et al., 2002, Smith et al., 2010, 2013). However, one of the highlights of our study is the newly discovered variation in size and content of inverted repeats. Because inverted repeats do not occur in most OCC members, IR variation can only be compared to Volvocales, whose IRs range between 14 and 42 kb (the largest was demonstrated in *Chlamydomonas moewusii* via physical mapping, Turmel et al., 1987), whereas most Sphaeropleales have IRs shorter than 14 kb, with three exceptions: *Kirchneriella* (36 kb), *Neochloris* (18 kb), and *Pseudomuriella* (22 kb). In the first two species, the extreme sizes are results of numerous intron insertions in IR genes (*rrs* and *rrl* in both, and *psbA* in *Kirchneriella*; while *Neochloris* has 9 introns in *psbA*, this gene is not part of the IR). In *Pseudomuriella* the increase in IR size appears to result from both intron insertion (*rrl* and *psbA*) and spacer region expansion. We can compare the enlarged IR region of *Neochloris* to that of a close relative,

Chlorotetraedron, which has a 13 kb IR with only one intron in *rrl* and one in *rrs*, versus four in each gene in *Neochloris*. Our results thus suggest a dynamic evolution of IRs, where genes can be added to or removed from the region, spacers shrink and/or expand, and introns can invade rapidly and in large numbers. This may not be surprising given the considerable variation in IR structure previously reported across green algae (e.g., in Trebouxiophyceae, Turmel et al., 2015), but has not yet been documented on a densely sampled ordinal scale.

Perhaps most striking is the small cp genome of *Mychonastes homosphaera*, a species reported to have an extremely compact, intron-less mitochondrial genome as well (Fučíková et al., 2014a). Its cp genome is only 102.7 kb in size, which makes it even smaller than the recently reported, also extremely small genomes of the ulvophyceans *Bryopsis* and *Tydemania* (Leliaert and Lopez-Bautista, 2015). It is impossible to determine the cause behind such a streamlined genome structure with the currently available data, but it can be hypothesized that the same forces are driving the evolution of both organellar genomes towards smaller size and low intron content. In addition to compaction in the non-coding portions of the genome, the pronounced nucleotide and codon bias in the coding regions of *Mychonastes* (Fig. 4) complicate the efforts to place this morphologically simple yet evolutionarily most unusual taxon in the phylogeny of Sphaeropleales (see sections 4.2.2 and 4.2.3 below).

4.2. Phylogenetic analyses

4.2.1. Previous Studies on the Phylogeny of Sphaeropleales

Unlike other green algal classes, the monophyly of the class Chlorophyceae has not been contested and the group is a long-accepted taxon in the Chlorophyta (e.g., Fučíková et al., 2014c). However, within the class, the reciprocal monophyly of Sphaeropleales and Volvocales has been subject to considerable debate (e.g., Tippo et al., 2012) and the relationships within Sphaeropleales have yet to be confidently resolved (Fučíková et al., 2014b).

A previous study focusing on sphaeroplealean mitochondrial genomes (Fučíková et al., 2014a) presented a moderately to well-supported phylogenetic tree based on 13 mitochondrial protein-coding genes. The topology of this mitochondrial tree conflicted with that resulting from analyses of 4 chloroplast genes and of 3 nuclear ribosomal genes, and is not entirely consistent with either of the topologies recovered in the present study (Fig. 2) – the main difference is in the grouping, albeit weakly supported, of *Bracteacoccus*, *Pseudomuriella*, and *Mychonastes* in the mt tree. Such conflicts suggest that more data and analyses are needed to pinpoint the evolutionary history within the order.

4.2.2. Conflict Among and Within Data Sets

The present study approached the phylogeny of Sphaeropleales with newly acquired chloroplast data (56 protein-coding genes) and a similar taxon sampling as Fučíková et al. (2014a). In contrast to the mitochondrial tree presented by Fučíková et al. (2014a), the present cp nt tree (Fig. 2) shows absolute (or nearly so) Bayesian support for all nodes, which may be simply due to the larger size of the concatenated cp data set (10,728 mt characters vs. 35,712 cp characters). Despite obtaining a fully resolved, well supported tree from our nt ML and Bayesian analyses, several nodes in this tree were in stark conflict with the topology from the corresponding

amino-acid analyses (Fig. 2). Clearly, at least one of the two topologies is an incorrect representation of the order's evolution, despite receiving high statistical support for all nodes. It is also notable that the mt analyses of Fučíková et al. (2014a) did not exhibit such a nt-aa conflict.

Besides concatenation, another approach to summarizing phylogenetic information from multiple genes is constructing a consensus tree of the individual gene trees. We adopted the method of Salichos et al. (2014) and constructed the Majority Rule Extended (MRE) consensus and mapped the internode certainty (IC) scores on it for both the nt and the aa data (only aa results are shown in Fig. 3). This tree is consistent with the concatenation results in most of the deep nodes and in recovering monophyletic Sphaeropleales, Volvocales, and OCC. However, it presents a third hypothesis for the topology within Sphaeropleales in addition to the trees obtained from analysis of concatenated nt and aa data. Unlike either of these two trees, the MRE tree places *Mychonastes* as sister to *Kirchneriella* and *Ourococcus* (members of Selenastraceae). This placement receives very low to no IC support, but this relationship had been recovered in a previous study (Fučíková et al., 2014a – results of a 4-chloroplast gene analysis). Overall the MRE tree showed low IC support, indicating that most consensus nodes are only recovered in a handful of single-gene analyses. It appears that the coccoid alga *Mychonastes* might be the problematic taxon and further, the positions of other deeply diverging coccoid taxa – *Bracteacoccus*, *Pseudomuriella*, and *Chromochloris* – are also inconsistently recovered. In some analyses including mitochondrial analyses of Fučíková et al., (2014a), *Chromochloris* is sister to colony-forming families Hydrodictyaceae (*Pediastrum*), Neochloridaceae (*Chlorotetraedron* and *Neochloris*), Rotundellaceae (*Rotundella*), Scenedesmaceae (*Acutodesmus*), and Selenastraceae (*Kirchneriella* and *Ourococcus*). In nt analyses of chloroplast data (Fig. 2) it is part of a well-supported clade with

Bracteacoccus and *Pseudomuriella*. All these taxa are subtended by long branches and it is therefore not unexpected to encounter problems with resolving their relationships with confidence.

One of the more recent divergences also appears problematic: the relationships among *Pediastrum*, *Chlorotetraedron*, and *Neochloris* have been resolved in all three possible ways by different data analyses (Fučíková et al., 2014a, present study), and always with good support. This conflict might improve in the future with increased taxon sampling within Hydrodictyaceae and Neochloridaceae, but may also result from systematic error or even horizontal gene transfer. A closer look at the genome architecture in these taxa might provide better insight into the evolution of their genomes and shed light on this peculiar “conflict zone”, where different, yet highly supported topologies can be derived even from the same data set.

4.2.3. Possible Sources of Conflict

As suggested by Novis et al. (2013), expansion of particular genes may mislead phylogenetic inference. However, our study is unlikely to be riddled with the same issues as seen in that study because we carefully removed the expanded and other hypervariable regions from our alignments, not only in the *rpo* genes implicated in Novis et al. (2013), but also other fast evolving genes. Some genes were even excluded from our analyses altogether. More importantly, removal of the *rpo* genes and *rps3*, as suggested by Novis et al. (2013), did not affect the results of our concatenated cp analyses, and neither did progressive removal of fast-evolving sites (Dryad study #XXX).

Among-taxon heterogeneity in nucleotide composition can be a major problem for correct inference of phylogenetic relationships, as it directly violates the stationarity assumptions of commonly used models of evolution (Cooper, 2014). Aside from simple GC nt bias, e.g., Rota-Stabelli et al. (2013) pointed out that codon usage bias may affect aa data as well. We therefore investigated possible biases in the aa concatenated data sets and across individual genes using KL divergence. We calculated the divergence of codon and aa usage from the values expected based on nt composition and determined that isoleucine and arginine exhibit the most variation among taxa (Fig. 4). However, Cox et al. (2014) demonstrated that synonymous nt substitutions in Arg, Leu, and Ser codons (the three amino acids coded for by six different codons) can be responsible for incorrect topology inference, and that ambiguous recoding of these amino acids can rectify the problem.

Merely removing the highly codon-biased *Mychonastes homosphaera* from the pre-trimmed non-ambiguous data set did not ameliorate the aa-nt conflict (alignments and trees available in Dryad). Thus, codon bias across multiple taxa is a plausible culprit for misleading the nt phylogeny. Following Cox et al. (2014), we recoded Arg, Leu, and Ser as ambiguous codons and recovered a topology much more consistent with the aa tree. The only remaining conflict is in the clade containing *Chlorotetraedron*, *Neochloris*, and *Pediastrum*, for which not only did the ambiguation fail to yield the aa topology, but in fact resulted in a topology different from either of the topologies seen in Fig. 2. Clearly, resolving the phylogenetic puzzle among members of this clade will require additional examination, once a full cp genome of *Pediastrum* becomes available.

At least in the case of *Mychonastes*, it is possible that nt composition bias similar to some of the outgroups may be “pulling” this long-branched taxon towards the base of the tree, and that this issue is corrected by using aa data. However, our results (Fig. 4)

show that such biases exist on aa level as well, and therefore it would be unwise to assume that the aa topology is also not riddled with systematic error. Other approaches have been implemented in previous studies; for example Zwick et al. (2012) treated the two Serine codon families as separate amino acids, and by doing so resolved the conflict between nt and aa topologies. However, the results of such an analysis in our case did not differ from the regular aa analysis (not shown). Also, having a full genome sequence of *Pediastrum* (which is the focus of another study) might allow us to resolve the phylogenetic puzzle within the *Chlorotetraedron*, *Neochloris*, and *Pediastrum* clade, and will be pursued in a forthcoming study focusing on this group of taxa.

Acknowledgements

Genomic data collection and analyses were supported with funding from NSF grants DEB-1036448 and DEB-1354146 awarded to LAL and POL at University of Connecticut. Analyses were carried out at the Bioinformatics Facility of the University of Connecticut Biotechnology and Bioservices Center (BBC). The authors would like to thank Dr. Jill Wegrzyn for valuable help with genome assemblies.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol., 215, 403–10.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSIBLAST: a new generation of protein database search programs.

Nucleic Acids Res., 25, 3389–402.

Bélanger, A. S., Brouard, J.-S., Charlebois, P., Otis, C., Lemieux, C., Turmel, M., 2006. Distinctive architecture of the chloroplast genome in the chlorophycean green alga *Stigeoclonium helveticum*. Mol. Genet. Genomics, 276, 464-477.

Brouard, J.-S., Otis, C., Lemieux, C., Turmel, M., 2008. Chloroplast DNA sequence of the green alga *Oedogonium cardiacum* (Chlorophyceae): Unique genome architecture, derived characters shared with the Chaetophorales and novel genes acquired through horizontal transfer. BMC Genomics, 9, 290.

Brouard, J.-S., Otis, C., Lemieux, C., Turmel, M., 2010. The exceptionally large chloroplast genome of the green alga *Floydiella terrestris* illuminates the evolutionary history of the Chlorophyceae. Genome Biol. Evol., 2, 240-256.

Brouard, J.-S., Otis, C., Lemieux, C., Turmel, M., 2011. The chloroplast genome of the green alga *Schizomeris leibleinii* (Chlorophyceae) provides evidence for bidirectional DNA replication from a single origin in the Chaetophorales. Genome Biol. Evol., 3, 505-515.

Cooper, E.D., 2014. Overly simplistic substitution models obscure green plant phylogeny. Trends Plant Sci., 19, 576-582.

Cox, C.J., Li, B., Foster, P.G., Embley, T.M., Civián, P., 2014. Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. Syst. Biol., 63, 272-279.

Cui, L., Leebens-Mack, J., Wang, L.-S., Tang, J., Rymarkis, L., Stern, D.B., dePamphilis, C.W., 2006. Adaptive evolution of chloroplast genome structure inferred using a parametric bootstrap approach. BMC Evol. Biol., 6, 13.

- De Cambiaire, J.-C., Otis, C., Lemieux, C, Turmel, M., 2006. The complete chloroplast genome sequence of the chlorophycean green alga *Scenedesmus obliquus* reveals a compact gene organization and a biased distribution of genes on the two DNA strands. *BMC Evol. Biol.*, 6, 37.
- Farwagi, A., Fučíková, K., McManus, H. A., 2015. Phylogenetic patterns of gene rearrangements in four mitochondrial genomes from the green algal family Hydrodictyaceae (Sphaeropleales, Chlorophyceae). *BMC Genomics*, 16, 826.
- Fučíková, K., Lewis, P.O., González-Halphen, D., Lewis, L.A., 2014a. Gene arrangement convergence, diverse intron content, and genetic code modifications in mitochondrial genomes of Sphaeropleales (Chlorophyta). *Genome Biol. Evol.*, 6, 2170-2180.
- Fučíková, K., Lewis, P.O., Lewis, L.A., 2014b. Putting incertae sedis taxa in their place: a proposal for ten new families and three new genera in the Sphaeropleales (Chlorophyta). *J. Phycol.*, 50, 14-25.
- Fučíková, K., Leliaert, F., Cooper, E.D., Škaloud, P., D'Hondt, S., De Clerck, O., Gurgel, C.F.D., Lewis, L.A., Lewis, P.O., Lopez-Bautista, J.M., Delwiche, C.F., Verbruggen, H., 2014c. New phylogenetic hypotheses for the core Chlorophyta based on chloroplast sequence data. *Frontiers in Ecology and Evolution*, 2, 63.
- Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics*, 17, 754-755.
- Kullback, S., Leibler, R.A. 1951. On information and sufficiency. *Ann. Math. Stat.*, 22, 79–86.
- Leliaert, F., Lopez-Bautista, J.M., 2015. The chloroplast genomes of *Bryopsis plumosa* and *Tydemania expeditiones* (Bryopsidales, Chlorophyta): compact genomes and genes of bacterial origin. *BMC Genomics*, 16, 204.

- Lemieux, C., Otis, C., Turmel, M., 2014a. Chloroplast phylogenomic analysis resolves deep-level relationships within the green algal class Trebouxiophyceae. *BMC Evol. Biol.*, 14, 211.
- Lemieux, C., Otis, C., Turmel, M., 2014b. Six newly sequenced chloroplast genomes from prasinophyte green algae provide insights into the relationships among prasinophyte lineages and the diversity of streamlined genome architecture in picoplanktonic species. *BMC Genomics*, 15, 857.
- Lewis, P.O., Xie, W., Chen, M.-H., Fan, Y., Kuo, L. 2014. Posterior predictive Bayesian phylogenetic model selection. *Syst. Biol.*, 63, 309–321.
- Maul, J.E., Lilly, J.W., Cui, L., dePamphilis, C.W., Miller, W., Harris, E.H., Stern, D.B., 2002. The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. *Plant Cell*, 14, 2659-2679.
- Novis, P.M., Smissen, R., Buckley, T.R., Gopalakrishnan, K., Visnovsky, G., 2013. Inclusion of chloroplast genes that have undergone expansion misleads phylogeny reconstruction in the Chlorophyta. *Am. J. Bot.*, 100, 2194-2209.
- Ronquist, F., Huelsenbeck, J.P., 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19, 1572-1574.
- Rota-Stabelli, O., Lartillot, N., Philippe, H., Pisani, D., 2013. Serine codon-usage bias in deep phylogenomics: Pancrustacean relationships as a case study. *Syst. Biol.* 62, 121–133.
- Ruhfel, B., Gitzendanner, M., Soltis, P., Soltis, D., Burleigh, J., 2014. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evol. Biol.*, 14, 23.

Salichos, L., Stamatakis, A., Rokas, A., 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol. Biol. Evol.*, 31, 1261-71.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I., 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res.*, 19, 1117-1123.

Smith, D.R., Lee, R.W., Cushman, J.C., Magnuson, J.K., Tran, D., Polle, J.E.W., 2010. The *Dunaliella salina* organelle genomes: large sequences, inflated with intronic and intergenic DNA. *BMC Plant Biol.*, 10, 83.

Smith, D.R., Hamaji, T., Olson, B.J.S.C., Durand, P.M., Ferris, P., Michod, R.E., Featherston, J., Nozaki, H., Keeling, P., 2013. Organelle genome complexity scales positively with organism size in volvocine green algae. *Mol. Biol. Evol.*, 30, 793-797.

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

Swofford, D.L., 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Tipperry, N.P., Fučíková, K., Lewis, P.O., Lewis, L.A., 2012. Probing the monophyly of the Sphaeropleales (Chlorophyceae) using data from five genes. *J. Phycol.*, 48, 1482-1493.

Turmel, M., Otis, C., Lemieux, C., 2013. Tracing the evolution of streptophyte algae and their mitochondrial genome. *Genome Biol. Evol.*, 5, 1817-1835.

Turmel, M., Bellemare, G., Lemieux, C., 1987. Physical mapping of differences between the chloroplast DNAs of the interfertile algae *Chlamydomonas eugametos* and *Chlamydomonas moewusii*. *Curr. Genet.*, 11, 543-552.

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, 2062-2082.

Zwick, A., Regier, J.C., Zwickl, D.J., 2012. Resolving discrepancy between nucleotides and amino acids in deep-level arthropod phylogenomics: differentiating serine codons in 21-amino-acid models. *PLoS One*, 7, e47450.

Figure legends:

Fig. 1. Overall size (in kb) of chloroplast genomes of green algae in Sphaeropleales. Proportions of the length of coding (bottom portion of the column), intronic (middle), and spacer (top) DNA in chloroplast genomes of Chlorophyceae. Numbers above columns indicate numbers of introns present in each genome. Examples of morphologies of sphaeroplealean algae are shown in micrographs above diagram. Scale bar applies to all micrographs.

Fig. 2. Contrast of tree topologies resulting from concatenated analyses of 56 chloroplast genes: aa (left) and nt (right). The first number at each branch indicates Bayesian posterior probability (PP), the second number ML bootstrap support (BS) and the third, if present, indicates PP resulting from Bayesian analysis of the concatenated nucleotide data set analyzed under a codon model (all except one are indicated on the aa topology and a large asterisk indicates the only relationship that is supported by the codon analysis but not the aa analysis – therefore in this one case the codon support value is indicated on the nt tree).

Fig. 3. MRE consensus tree calculated by RAxML from best ML trees resulting from individual single-gene aa analyses. Numbers at branches indicate GSF (first number, expressed as a percentage) and IC (second number) support. Negative numbers indicate that another split, not included in the consensus tree, received higher support.

Fig. 4. Amino acid bias across Chlorophyceae measured using KL divergence. Amino acids exhibiting the highest variance are highlighted with bold lines, as are the three amino acids implicated as problematic by Cox et al. (2014), who found compositional biases in synonymous Arg, Leu, and Ser substitutions to confound phylogenetic inference.

Table 1. Overview of quantitative characteristics of known chloroplast genomes in Chlorophyceae. Newly obtained genomes are highlighted with boldface species names. Numbers of genes do not include unidentified ORFs in previously published genomes.

Numbers of IR genes do not include genes only partially placed in the IR.

Species	Strain	GenBank	Size (bp)	Percent GC	Percent coding	Genes	Sided blocks	Sidedness index	Introns	IR size	IR genes
<i>Acutodesmus obliquus</i>	UTEX 393	DQ396875	161452	26.9	56.0	106	11	0.90	10	12022	6
<i>Ankyra judai</i>	SAG 17.84	KT199255	157224	28.3	57.0	109	23	0.80	2	8247	6
<i>Bracteacoccus aerius</i>	UTEX 1250	KT199254	165732	31.7	54.9	103	14	0.87	2	7271	5
<i>Bracteacoccus minor</i>	UTEX B 66	KT199253	192761	31.9	48.4	104	14	0.87	3	9577	6
<i>Chlorotetraedron incus</i>	SAG 43.81	KT199252	193197	27.1	46.7	106	14	0.88	10	13490	8
<i>Chromochloris zofingiensis</i>	UTEX 56	KT199251	188937	30.9	47.2	106	17	0.85	2	6375	6
<i>Kirchneriella aperta</i>	SAG 2004	KT199250	207516	34.1	42.3	106	13	0.89	27	35503	7
<i>Mychonastes homosphaera</i>	CAUP H 6502	KT199249	102718	39.8	80.1	105	12	0.89	1	6472	6
<i>Neochloris aquatica</i>	UTEX 138	KT199248	166767	30.3	50.9	102	12	0.89	32	18217	8
<i>Pseudomuriella schumacherensis</i>	SAG 2137	KT199256	220357	31.2	41.6	109	10	0.92	8	22004	7
<i>Chlamydomonas reinhardtii</i>	CC-503 cw92	FJ423446	204159	34.5	36.2	107	14	0.88	11	22213	8
<i>Dunaliella salina</i>	CCAP 19/18	GQ250046	269044	32.1	41.7	105	16	0.86	36	14409	5
<i>Gonium pectorale</i>	K3-F3-4	AP012494	222582	29.8	43.3	107	8	0.93	3	14750	8

<i>Pleodorina starii</i>	NIES-1363	JX977846	269857	34.9	37.9	103	16	0.85	20	16608	6
<i>Volvox carteri</i>	UTEX 2908	GU084820	461064	42.4	20.9	105	18	0.84	8	15948	8
<i>Schizomeris leiblienii</i>	UTEX LB 1228	HQ700713	182759	27.2	51.6	101	4	0.97	34	0	0
<i>Stigeoclonium helveticum</i>	UTEX 441	DQ630521	223902	31.2	48.6	98	8	0.93	22	0	0
<i>Floydiella terrestris</i>	UTEX 1709	GU196268	521168	34.5	18.6	99	10	0.91	27	0	0
<i>Oedogonium cardiacum</i>	SAG 575-1b	EU677193	196547	29.5	55.8	115	27	0.77	34	35492	11

Figure

Fig 1 grayscale

Figure 1

Figure 2

Figure 3

well supported by all concatenated analyses

well supported by some concatenated analyses

not supported by concatenated analyses

Fig 4 grayscale