Molecular Phylogenetics and Evolution 127 (2018) 248–255

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Improving phylogenetic inference of core Chlorophyta using chloroplast sequences with strong phylogenetic signals and heterogeneous models

Ling Fang[a], Frederik Leliaert[b,c], Phil M. Novis[d], Zhenhua Zhang[a], Huan Zhu[e], Guoxiang Liu[e], David Penny[f], Bojian Zhong[a,*]

[a] College of Life Sciences, Nanjing Normal University, Nanjing 210023, China
[b] Botanic Garden Meise, 1860 Meise, Belgium
[c] Phycology Research Group, Biology Department, Ghent University, 9000 Ghent, Belgium
[d] Allan Herbarium, Manaaki Whenua-Landcare Research, Lincoln 7640, New Zealand
[e] Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China
[f] Institute of Fundamental Sciences, Massey University, Palmerston North 4442, New Zealand

ARTICLE INFO

Keywords:
Green algae
Phylogenomics
GC-heterogeneous sites
Heterogeneous models
Phylogenetic signals

ABSTRACT

Phylogenetic relationships within the green algal phylum Chlorophyta have proven difficult to resolve. The core Chlorophyta include Chlorophyceae, Ulvophyceae, Trebouxiophyceae, Pedinophyceae and Chlorodendrophyceae, but the relationships among these classes remain unresolved and the monophyly of Ulvophyceae and Trebouxiophyceae are highly controversial. We analyzed a dataset of 101 green algal species and 73 protein-coding genes sampled from complete and partial chloroplast genomes, including six newly sequenced ulvophyte genomes (*Blidingia minima* NIES-1837, *Ulothrix zonata*, *Halochlorococcum* sp. NIES-1838, *Scotinosphaera* sp. NIES-154, *Caulerpa brownii* and *Cephaleuros* sp. HZ-2017). We applied the Tree Certainty (TC) score to quantify the level of incongruence between phylogenetic trees in chloroplast genomic datasets, and show that the conflicting phylogenetic trees of core Chlorophyta stem from the most GC-heterogeneous sites. With removing the most GC-heterogeneous sites, our chloroplast phylogenomic analyses using heterogeneous models consistently support monophyly of the Chlorophyceae and of the Trebouxiophyceae, but the Ulvophyceae was resolved as polyphyletic. Our analytical framework provides an efficient approach to reconstruct the optimal phylogenetic relationships by minimizing conflicting signals.

1. Introduction

Green algae represent an ancient lineage and one of the most abundant groups of photosynthetic eukaryotes (Leliaert et al., 2012). They are divided into two main clades: the charophyte algae which include the closest living relatives of land plants, and the Chlorophyta including ecologically, morphologically and cytologically diverse green algae. The phylogenetic relationships of charophyte algae have been relatively well resolved (Zhong et al., 2013, 2015; Cooper, 2014; Wickett et al., 2014), however the evolutionary history of the Chlorophyta remains ambiguous (Fučíková et al., 2014; Lemieux et al., 2014a; Turmel et al., 2016; Leliaert and Lopez-Bautista, 2015). The early diverging lineages of Chlorophyta are mainly composed of unicellular planktonic algae (known as prasinophytes), forming a paraphyletic group with a wide variety of cell shapes and flagellar numbers (Lemieux et al., 2014a; Leliaert et al., 2016). The monophyletic core Chlorophyta currently consists of five classes: Chlorophyceae, Ulvophyceae, Trebouxiophyceae, Pedinophyceae and Chlorodendrophyceae (Fučíková et al., 2014; Turmel et al., 2016, 2017; Sun et al., 2016). A reliable and stable phylogenetic tree of the core Chlorophyta is important to elucidate the relationships among the ancient lineages and understand the early evolutionary history of green algae. The class Chlorophyceae was recovered as a monophyletic group with strong support based on chloroplast and nuclear genes (Buchheim et al., 2001; Fučíková et al., 2014; Lemieux et al., 2015). The classes Ulvophyceae and Trebouxiophyceae were originally defined based on ultrastructural data and their circumscriptions refined by 18S data (Mattox and Stewart, 1984; Lewis and McCourt, 2004; Leliaert et al., 2012). However, monophyly of the Ulvophyceae and of the Trebouxiophyceae are still contentious. Although some previous phylogenetic analyses based on 18S data suggested monophyly of these classes, this was never with strong support (Kantz et al., 1990; Krienitz et al., 2003; O'kelly et al., 2004). Cocquyt et al. (2010b) recovered the Ulvophyceae as a monophyletic group using eight nuclear and two plastid genes, but recent

* Corresponding author.
E-mail address: bjzhong@gmail.com (B. Zhong).

https://doi.org/10.1016/j.ympev.2018.06.006
Received 22 March 2018; Received in revised form 26 May 2018; Accepted 4 June 2018
Available online 08 June 2018
1055-7903/

chloroplast phylogenomic analyses have supported non-monophyly of the Ulvophyceae (Fučíková et al., 2014; Leliaert and Lopez-Bautista, 2015; Melton et al., 2015). The Trebouxiophyceae was supported as a monophyletic class with nuclear ribosomal data (Neustupa et al., 2011; Bock et al., 2013; Neustupa et al., 2013), while analyses of chloroplast genomic data suggested that organisms currently grouped into Trebouxiophyceae form two distinct clades (Chlorellales and core Trebouxiophyceae) that are not sisters (Fučíková et al., 2014; Lemieux et al., 2014b; Turmel et al., 2015, 2017). Therefore, there is an urgent need to re-evaluate the unresolved classification of core Chlorophyta.

With the rapid advances of high-throughput sequencing technologies, the availability of chloroplast genomes of green algae are dramatically expanding, providing an unprecedented large-scale data set to investigate the evolution of green algae (Fang et al., 2017). Identifying loci with reliable phylogenetic signals among large amounts of data is becoming one of the most important issues in phylogenomic analyses (Fučíková et al., 2014; Zhong et al., 2014; Romiguier et al., 2015). It is worth noting that most chloroplast genomes of green algae are AT-rich, with unusual exceptions such as *Coccomyxa* sp. C-169 (Smith et al., 2011), *Trebouxiophyceae* sp. MX-AZ01 (Servíngarcidueñas and Martínezromero, 2012) and *Paradoxia multiseta* (Lemieux et al., 2014b). The level of GC bias results in phylogenetic conflicts within genome-scale datasets (Romiguier et al., 2013; Weber et al., 2014; Romiguier et al., 2015; Bossert et al., 2017). It has been widely reported that GC-heterogeneous genes (genes showing a wide variance of GC% among sequences in an alignment) have a negative effect on tree reconstruction, and phylogenomic analyses have demonstrated that most GC-heterogeneous loci experienced fast rates of evolution and may exacerbate model misspecification issues, which considerably reduces the accuracy of phylogenetic reconstructions (Delsuc et al., 2003; Betancur-R et al., 2013; Romiguier et al., 2015).

There have been two main approaches to model the evolutionary history of sequences: homogeneous model and heterogeneous model. The widely used homogeneous models can model rate heterogeneity across sites using a gamma distribution for among-site rate variation (Yang, 1996). Numerous homogeneous models have been applied to phylogenomic datasets (Cocquyt et al., 2010b; Fučíková et al., 2014; Zhong et al., 2014; Lemieux et al., 2014b). However, homogeneous models are often too simple to fully consider compositional heterogeneity. The violation of base compositional homogeneity can induce erroneous topologies due to model misspecifications (Phillips et al., 2004; Romiguier and Roux, 2017). In contrast, heterogeneous models can model exchange rate variation and compositional variation among species or across sites (Foster, 2004; Lartillot and Philippe, 2004; Boussau and Gouy, 2006). It has been shown that the heterogeneous models have better fitness to data than homogeneous models (Morgan et al., 2013, Moran et al., 2015). These models could increase the accuracy of phylogenetic inference (Cox et al., 2014; Zhong et al., 2014; Romiguier et al., 2015).

Bootstrap support values have been widely used to measure the robustness of inference when the data are limited (Felsenstein, 1985). However, for large-scale data sets, bootstrapping will often result in strong support, while masking conflicting phylogenetic signals (Rokas and Carroll, 2006; Kumar et al., 2012; Salichos and Rokas, 2013). A recently developed approach to calculate internode certainty (IC) and tree certainty (TC) values can quantify conflicts in phylogenetic signals for a single internode and the whole tree, and has proven useful to select the optimal dataset with strong phylogenetic signals (Salichos and Rokas, 2013; Salichos et al., 2014; Kobert et al., 2016). Combining this approach with heterogeneous modeling is novel in phylogenetic analysis of the Chlorophyta.

Reconstruction of phylogenetic relationships among the main clades of the core Chlorophyta has been difficult due to the small number of genes available, model misspecification and uneven taxon sampling. The classes Chlorophyceae and Trebouxiophyceae have been rather well sampled. As of April 2017, 60 complete chloroplast genomes in the Chlorophyceae and Trebouxiophyceae were available, which almost cover all orders and families in two classes (Lemieux et al., 2014b; Fučíková et al., 2016). However, the class Ulvophyceae is poorly sampled, with 15 representatives of only four orders containing complete chloroplast genomes (Fučíková et al., 2014; Lemieux et al., 2014a; Turmel et al., 2016; Leliaert and Lopez-Bautista, 2015). In this study, we sequenced chloroplast protein-coding genes of six ulvophycean species (*Blidingia minima* NIES-1837, *Ulothrix zonata*, *Halochlorococcum* sp. NIES-1838, *Scotinosphaera* sp. NIES-154, *Caulerpa brownii* and *Cephaleuros* sp. HZ-2017) from six orders. We successively removed most GC-heterogeneous sites to select optimum datasets with less conflicting phylogenetic signals, applied two heterogeneous models to reconstruct the phylogenetic relationships of the core Chlorophyta, and tested the monophyly of the Trebouxiophyceae and Ulvophyceae.

2. Materials and methods

2.1. Algal strains and culture conditions

Blidingia minima (NIES-1837) (Hamana et al., 2013), *Ulothrix zonata* (NIES-537) (Takamura et al., 1989; Mori et al., 2002), *Halochlorococcum* sp. (NIES-1838), and *Scotinosphaera* sp. (NIES-154) (Watanabe, 1983; Mori et al., 2002) strains were obtained from the Microbial Culture Collection at the National Institute for Environmental Studies, Tsukuba, Japan. *Caulerpa brownii* was collected from Brinn's Point (−45.670013°S, 170.652462°E), Puketeraki, New Zealand (archived specimen CHR 640974), and *Cephaleuros* sp. HZ-2017 was obtained from the Freshwater Algae Culture Collection at the Institute of Hydrobiology, Chinese Academy of Sciences (FACHB-collection). *Ulothrix zonata*, *Scotinosphaera* sp. and *Cephaleuros* sp. were grown in C medium (Ichimura, 1971), *Blidingia minima* in IMK and *Halochlorococcum* sp. in f/2 (Guillard and Ryther, 1962). All strains were cultured at 20 °C under alternating 12 h-light/12 h-dark periods.

2.2. Chloroplast genome sequencing, assembly and annotation

Total genomic DNA was extracted using a QIAGEN DNEasy Plant Mini Kit (QIAGEN, Valencia, CA, USA). High-throughput sequencing was performed using Illumina HiSeqTM2000 technology (150 bp paired-end reads, supplementary Table S2). De novo assembly of paired-end reads was performed using SPAdes 3.6.2 (Bankevich et al., 2012) with the parameter "–careful". The contigs of the chloroplast genomes were identified by BLASTN (Altschul et al., 1997) similarity search with an E-value cutoff of $1e^{-5}$ against a custom-built database including genes from all published chloroplast genomes of Chlorophyta. The paired-end reads were used to iteratively increase the length of contigs using the PRICE assembler (Ruby et al., 2013) and the extended contigs were checked by Bowtie2 (Langmead and Salzberg, 2012) with the paired-end reads. Contigs encoding chloroplast genes were annotated with Geneious v9.0.4 (Kearse et al., 2012).

2.3. Phylogenomic analyses

Phylogenomic analyses were based on complete and partial chloroplast genome data from 101 green algae, including 33 Chlorophyceae (five orders), 15 Ulvophyceae (seven orders), 30 Trebouxiophyceae (11 major clades), three Pedinophyceae, two Chlorodendrophyceae, 16 prasinophytes and two streptophytes (*Chlorokybus atmophyticus* and *Mesostigma viride*) as outgroups. GenBank accession numbers of genes used and determined in the present study are listed in supplementary Table S1. The amino acid sequences were aligned for each gene using MUSCLE (Edgar, 2004) with the default settings, and the sequences back-translated to nucleotides. Poorly aligned sites were removed using Gblocks 0.91b (Castresana, 2000) with −b5 = h and other default parameters. A full dataset including 39,123 sites was produced by concatenating 73 chloroplast genes of 99 chlorophytes and two streptophytes.