J. Phycol. **53,** 415–424 (2017)
© 2017 Phycological Society of America
DOI: 10.1111/jpy.12515

COMPARATIVE DNA SEQUENCE ANALYSES OF *PYRAMIMONAS PARKEAE* (PRASINOPHYCEAE) CHLOROPLAST GENOMES[1]

Anchittha Satjarak,[2] *and Linda E. Graham*

Department of Botany, University of Wisconsin-Madison, 430 Lincoln Drive, Madison, Wisconsin, USA

Prasinophytes form a paraphyletic assemblage of early diverging green algae, which have the potential to reveal the traits of the last common ancestor of the main two green lineages: (i) chlorophyte algae and (ii) streptophyte algae. Understanding the genetic composition of prasinophyte algae is fundamental to understanding the diversification and evolutionary processes that may have occurred in both green lineages. In this study, we sequenced the chloroplast genome of *Pyramimonas parkeae* NIES254 and compared it with that of *P. parkeae* CCMP726, the only other fully sequenced *P. parkeae* chloroplast genome. The results revealed that *P. parkeae* chloroplast genomes are surprisingly variable. The chloroplast genome of NIES254 was larger than that of CCMP726 by 3,204 bp, the NIES254 large single copy was 288 bp longer, the small single copy was 5,088 bp longer, and the IR was 1,086 bp shorter than that of CCMP726. Similarity values of the two strains were almost zero in four large hot spot regions. Finally, the strains differed in copy number for three protein-coding genes: *ycf20*, *psaC*, and *ndhE*. Phylogenetic analyses using 16S and 18S rDNA and *rbcL* sequences resolved a clade consisting of these two *P. parkeae* strains and a clade consisting of these plus other *Pyramimonas* isolates. These results are consistent with past studies indicating that prasinophyte chloroplast genomes display a higher level of variation than is commonly found among land plants. Consequently, prasinophyte chloroplast genomes may be less useful for inferring the early history of Viridiplantae than has been the case for land plant diversification.

Key index words: **chloroplast DNA variation; chloroplast genome; intraspecific variation; prasinophyte; *Pyramimonas parkeae***

***Abbreviations*: NIES, Microbial Culture Collection at the National Institute for Environmental Studies; SNPs, single nucleotide polymorphisms**

Prasinophytes are early diverging green algae that show heterogeneity in morphology and the potential to elucidate the traits of the last common ancestor of the main two green (Viridiplantae) lineages: 1) chlorophyte algae, the clade representing the majority of present green algal diversity and 2) streptophyte algae, a smaller, paraphyletic algal assemblage known to be closely related to land plants. The independent lineages currently recognized are: Pyramimonadales (clade I), Mamiellophyceae (clade II), Nephroselmidophyceae (clade III), Chlorodendrales (clade IV), Pycnococcaceae (clade V), Prasinococcales (clade VI, Palmophyllophyceae class nov., which was recently hypothesized to be closest to the divergence of chlorophyte and streptophyte algae, Leliaert et al. 2016), clade VII, clade VIII, and clade IX (Lemieux et al. 2014).

Chloroplast genome sequences have been sources of data for plant phylogenetics because chloroplasts exhibit uniparental inheritance and a slow rate of mutation (Wolfe et al. 1987, Allender et al. 2007). Comparative studies of the majority of plant plastid genome architectures show only low variation; gene order and essential gene content are highly conserved in plant plastid genomes (De Las Rivas et al. 2002). However, by comparison to plants, green algal chloroplast genomes seem to evolve in a much less conservative fashion. Green algal plastid genomes show variation in gene order, genome length, and presence of quadripartite structure (Brouard et al. 2010, Turmel et al. 2015, 2016, Lemieux et al. 2016) Comparisons of 12 chloroplast genomes from six prasinophyte clades revealed that prasinophyte chloroplast genomes likewise display variability in organization, gene content, and gene order (Turmel et al. 1999, 2009, Robbens et al. 2007, Worden et al. 2009, Lemieux et al. 2014).

The presence of this high variation among chloroplast genomes of prasinophyte clade II (Mamiellophyceae) suggests that variability may also occur at the intra-specific level. Comparison of 13 *Ostreococcus tauri* strains showed intra-specific variation in single nucleotide polymorphisms (SNPs), and presence of large insertion/deletion regions (Blanc-Mathieu et al. 2013). These observations indicate that similar surprising levels of intra-specific variation in chloroplast sequences might occur in other prasinophyte clades, but so far that possibility has not been investigated. We evaluated the level of intra-specific variation in chloroplast sequence in *Pyramimonas parkeae* (R.E. Norris & B.R. Pearson) representing

[1] Received 6 May 2016. Accepted 13 December 2016.
[2] Author for correspondence: e-mail anchittha.satjarak@gmail.com.
Editorial Responsibility: M. Wood (Associate Editor)

prasinophyte clade I, which has a conserved quadripartite structure and is closely related to divergence of streptophytes.

The complete chloroplast genome sequence of *P. parkeae* CCMP726 was released in 2009 by Turmel et al. For comparative analyses, we assembled the complete *P. parkeae* NIES254 chloroplast genome, a closely related strain obtained from National Institute for Environmental Studies (NIES), Japan. Our comparison showed that the two chloroplast genomes have identical gene content and similar arrangement. However, we found evidence for four large hotspot regions with nearly zero similarity value, movement of inverted repeat (IR) boundaries, and gene copy number differences.

MATERIALS AND METHODS

DNA extraction. A culture of *P. parkeae* Norris and Pearson (NIES254) was acquired from NIES, Japan. The culture was propagated in Alga-Gro® seawater medium (Carolina Biological Supply Company, Burlington, NC, USA), and was maintained in a walk-in growth room with 16:8 daily light:dark cycle at 20°C. Cells were harvested during the exponential phase. Total DNA was prepared by using FastDNA® SPIN Kit for Soil (MP Biomedicals, Solon, OH, USA) and sequenced by Illumina Miseq technologies at the University of Wisconsin-Madison Biotechnology Center.

Data pre-processing and genome construction. The raw paired-end Illumina data consisted of 13,232,998 reads with average read length of 251 bp. The data were trimmed by Trimmomatic v 0.33 (Bolger et al. 2014) to obtain the quality score of at least 28 on the phred 64 scale. The chloroplast genome was initially constructed using de novo sequence assembly, which proved challenging because repeat regions were longer than individual reads. Therefore, we employed a baiting and iterative mapping method described in Satjarak et al. (2016) using MIRA v 4.0.2 and MITObim v 1.8 (Hahn et al. 2013). Protein-coding sequences of *P. parkeae* CCMP726 chloroplast genome (Turmel et al. 2009) available in GenBank (accession number: FJ493499.1) were used as baits.

Sequence analyses. To determine the chloroplast genome coverage, we aligned the trimmed reads against the newly constructed NIES254 chloroplast genome using BWA non-model species alignment v 0.7.4 (Li and Durbin 2009) and calculated the coverage of every position in the plastid genome using Bedtools Genome Coverage BAM v 2.19.1 (Quinlan 2014) implemented in iPlant Collaborative (Goff et al. 2011). The functions of the open reading frames (ORFs) with length of at least 100 bp were predicted using BLAST search against the NCBI non-redundant protein databases accessed in February 2016 (http://blast.ncbi.nlm.nih.gov/Blast.cgi). tRNAs and rRNAs were predicted using tRNAscan-SE v 1.21 (Schattner et al. 2005) and RNAmmer v 1.2 (Lagesen et al. 2007). Base frequencies, amino acid frequencies, and codon usage were calculated using statistics option in Geneious v 9.0.4 (Kearse et al. 2012). The circular genome was drawn using OGDraw v 1.2 (Lohse et al. 2013). The resulting annotated sequence has been deposited at the GenBank under accession number KX013546.

Relationship between P. parkeae *NIES254 and CCMP726.* We used a phylogenetic approach to assess the relationship between *P. parkeae* NIES254 and CCMP726. The 18S rDNA of *P. parkeae* NIES254 was constructed from paird-end Mi-Seq reads sequenced from whole genomic DNA of *P. parkeae* NIES254 using methods described in Satjarak et al. (2016). The gene was assembled using 18S rDNA of *P. parkeae* Hachijo (accession number AB017124, Nakayama et al. 1998) as a bait. The average coverage of the sequence was estimated using BWA non-model species alignment v 0.7.4 (Li and Durbin 2009) and Bedtools Genome Coverage BAM v 2.19.1 (Quinlan 2014) implemented in iPlant Collaborative (Goff et al. 2011). 18S rDNA was predicted using RNAmmer v 1.2 (Lagesen et al. 2007). The final 18S rDNA construct was a linear molecule of 1,802 bp. The average coverage of every position of the gene was 144-fold. The resulting annotated sequence has been deposited at the GenBank under accession number KX611141.

To assess the relationship between the two strains, we performed phylogenetic analyses of 3 genes: 18S rDNA, 16S rDNA, and *rbcL*. *Pyramimonas* 18S rDNA, 16S rDNA, and *rbcL* sequences publicly available in GenBank (accessed in July 2016) were used in the analyses. *Cymbomonas tetramitiformis* DNA sequences of corresponding genes were used as outgroups.

The accession numbers of 18S rDNA sequences used in the phylogenetic analyses were FN562438, AB017126, AB052289, AJ404886, FN562440, AB017121, HQ111511, HQ111509, HQ111510, KF422615, FN562442, AB017122, KF615765, FN562443, KT860881, AB017124, AB017123, FN562441, AB999994, AB853999, AB854000, AB854001, AB854002, KF899837, AB854003, AB854004, AB854006, AB854005, AB854007, AB854008, AB854009, AB854010, AB854011, AB854012, AB854013, AB854014, AB854016, AB854015, AB854017, AB854018, AB854021, AB854020, AB854019, AB854022, AB854023, AB854024, AB854025, JN934670, JF794047, JF794048, JN934689, KT860923, AB854026, AB854027, AB854028, AB854029, AB854030, AB854031, AB854032, AB854033, AB854034, AB854035, AB854036, AB854037, AB854038, AB854039 and KX611141 (Nakayama et al. 1998, Moro et al. 2002, Suda 2004, Marin and Melkonian 2010, Balzano et al. 2012, Suda et al. 2013, Duanmu et al. 2014, Bhuiyan et al. 2015). The accession numbers of 16S rDNA sequences used in the phylogenetic analyses were AF393608, L34687, LK391817, LK391818, K391819, LK391820, LN735316, LN735321, LN735377, LN735378, LN735435, KX013545.1, FJ493499.1, and KX013546 (Daugbjerg et al. 1994, Turmel et al. 2002, 2009, Decelle et al. 2015, Satjarak et al. 2016). The accession numbers of *rbcL* sequences used in the phylogenetic analyses were AB052290, L34776, L34814, L34819, L34779, L34810, L34812, L34811, L34817, L34815, L34816, L34813, L34777, L34833, L34778, LC015748, LC015747, L34834, KP096399, L34818, KX013545.1, FJ493499.1, and KX013546 (Daugbjerg et al. 1994, Suda 2004, Bhuiyan et al. 2015, Satjarak et al. 2016).

We aligned the sequences using Geneious; setting free end gaps and identity to (1.0/0.0) resulted in 1,922 bp unambiguously aligned sequences of 18S rDNA, 706 bp of 16S rDNA, and 1,089 bp of *rbcL*. For each gene, the nucleotide substitution model was computed using jModelTest2 (Darriba et al. 2012). Maximum-Likelihood (ML) analysis was performed using RAxML (v 8.2.8) (Stamatakis 2014) on the CIPRES XSEDE Portal (Miller et al. 2010) using a GTR + I + F substitution model, employing the rapid bootstrapping method with 1,000 replications for bootstrap analyses. Baysian analyses were performed with MrBayes v 3.2.6 (Ronquist and Huelsenbeck 2003) using a GTR + I + F substitution model. Four independent chains were run for 1,100,000 cycles and the consensus topologies were calculated after the burn-in of 100,000 cycles.

Comparative analysis of P. parkeae *chloroplast genomes.* The analysis of syntenic conservation between *P. parkeae* NIES254 and CCMP726 was performed using progressiveMauve alignment v 2.4.0 (Darling et al. 2010). The two genomes were also aligned using LAST (Kiełbasa et al. 2011), with the following parameters: maximum score, max multiplicity for initial matches = 10, minimum length for initial matches = 1, step-size along reference sequences = 1, step-size along query sequences = 1, query letters per random alignment = 1e6. SNPs within the whole genome and within the protein coding regions were identified using Geneious alignments v 9.0.4 (Kearse et al. 2012). Synonymous (Ks) and nonsynonymous (Ka) substitution sites as well as the Ka/Ks ratio were calculated using MEGA6 v 6.06 (Tamura et al. 2013). To compare variability at the intraspecific level, we also calculated Ka/Ks ratios of protein coding sequences of *O. tauri*.

RESULTS

Chloroplast genome of P. parkeae *NIES254.* We sequenced the chloroplast genome of *P. parkeae* NIES254 for comparison to *P. parkeae* CCMP726 to investigate intra-specific genetic diversity. The newly sequenced *P. parkeae* NIES254 chloroplast genome was observed to have quadripartite structure of a 104,809 bp-long mapping circular molecule. The genome featured two copies of the IR (11,971 bp encompassing 22.84% of the genome), which separated the large single copy region (LSC; 65,441 bp) from the small single copy region (SSC; 15,426 bp; Fig. 1). The coverage of every position of the chloroplast genome ranged from 286- to 939-fold. GC content was 34.2%. The coding capacity of NIES254 was the same as that of CCMP726. This NIES254 chloroplast genome encoded 112 conserved genes including two rRNAs, 26 tRNAs, and 81 protein coding genes plus three ORFs (Fig. 1). The genome of NIES254 was longer than that of CCMP726 by 3,204 bp. NIES254 LSC was 288 bp longer, the SSC was 5,088 bp longer, and the IR was 1,086 bp shorter than that of CCMP726.

Relationship between P. parkeae *NIES254 and CCMP726.* ML and Baysian assessments of 18S and 16S rDNA and *rbcL* sequences to infer the relationship between *P. parkeae* NIES254 and CCMP726 resulted in monophyletic clades of *P. parkeae* strains (Figs. 2–4).

Comparative analyses of P. parkeae *chloroplast genomes.* Mauve alignment analysis of synteny of the two chloroplast genomes—*P. parkeae* strains NIES254 and CCMP726 showed that these genomes exhibited a collinear relationship, as only one syntenic block from each strain was present (Fig. 5). Although the genomes were collinear, the Mauve alignment showed four large hotspot regions where similarity values were almost zero. Such regions could be classified into three categories: (i) the 6 kb intergenic region between *psbA-trnS* and *ndhB* in LSC, (ii) 2 kb intron of *atpB*, and (iii) boundaries of IR and SSC, 5.7 kb at IRB-SSC and 6.8 kb at SSC-IRA (Fig. 2).

The size of the first large hotspot region (located between *psbA-trnS* and *ndhB* in LSC) was 6,848 bp in NIES254 and 6,240 bp in CCMP726. This intergenic region, which represented the largest intergenic region, contained different ORFs. However, we did not find orthologous protein products of the ORFs by similarity searches between the two genomes and against the non-redundant protein databases. A second hotspot region was the intron of *atpB* gene. The region was 2,144 bp long in NIES254 and 2,757 bp long in CCMP726. Both were group II introns with conserved region of reverse transcriptases of group II intron origin. The third and the fourth large hotspots occurred at the border between IRB-SSC and SSC-IRA. These hotspot regions resulted from boundary movement. The shift observed at the IRB-LSC boundary was minor, but the shift at the IRA-SSC boundary was greater. These movements and nucleotide variation at the boundaries resulted in difference in length and presence of ORFs and genes within IRB-SSC-IRA region of the two algae. Only one of the hypothetical ORFs (orf 454, 1,365 bp) in the CCMP726 IRs was present in those of NIES254, but in the latter, it was fragmented into three separated ORFs having lengths of 126, 402, and 186 bp. Also, the boundary movement caused re-positioning and change in copy number of three genes: *ycf20, psaC,* and *ndhE.* In CCMP726, these three genes were present on both copies of the IR region, whereas in NIES254 they were present on one end of the SSC region.

Variability between the two chloroplast genomes was present in all of three informative regions: (i) protein coding regions, (ii) intronic regions, and (iii) intergenic regions. Alignment of NIES254 and CCMP726 chloroplast using LAST resulted in 93 similar regions due to the high variability present in the intergenic regions. This high variability made it challenging to identify the variable positions throughout the whole genome. Therefore, we were only able to perform comparative analyses of protein coding regions.

The total number of polymorphic sites in protein coding genes and ORF (*orf91*) was 3,111 positions including 2,684 SNPs and 44 indels. *ftsH* exhibited the highest number of SNPs and indels: 246 SNPs (seven positions per 100 bpbase pair) and six indels. The number of substitutions per 100 nucleotides of protein coding genes and *orf91* showed that mutations were randomly distributed across the chloroplast genomes (Fig. 6). Among plastid coding sequences, *psbT* possessed the highest Ks (57 positions per 100 nucleotides), *ftsH* possessed the highest Ka (16 positions per 100 nucleotides), and *petA* exhibited the highest Ka/Ks ratio (1.00; Fig. 6).

DISCUSSION

The availability of a sequenced chloroplast genome for *P. parkeae* NIES254 provided the opportunity for comparative analysis of chloroplast genome structure between *P. parkeae* strains NIES254 and CCMP726. These *P. parkeae* chloroplast genomes were similar in gene content. Of three ORFs (*orf91,*

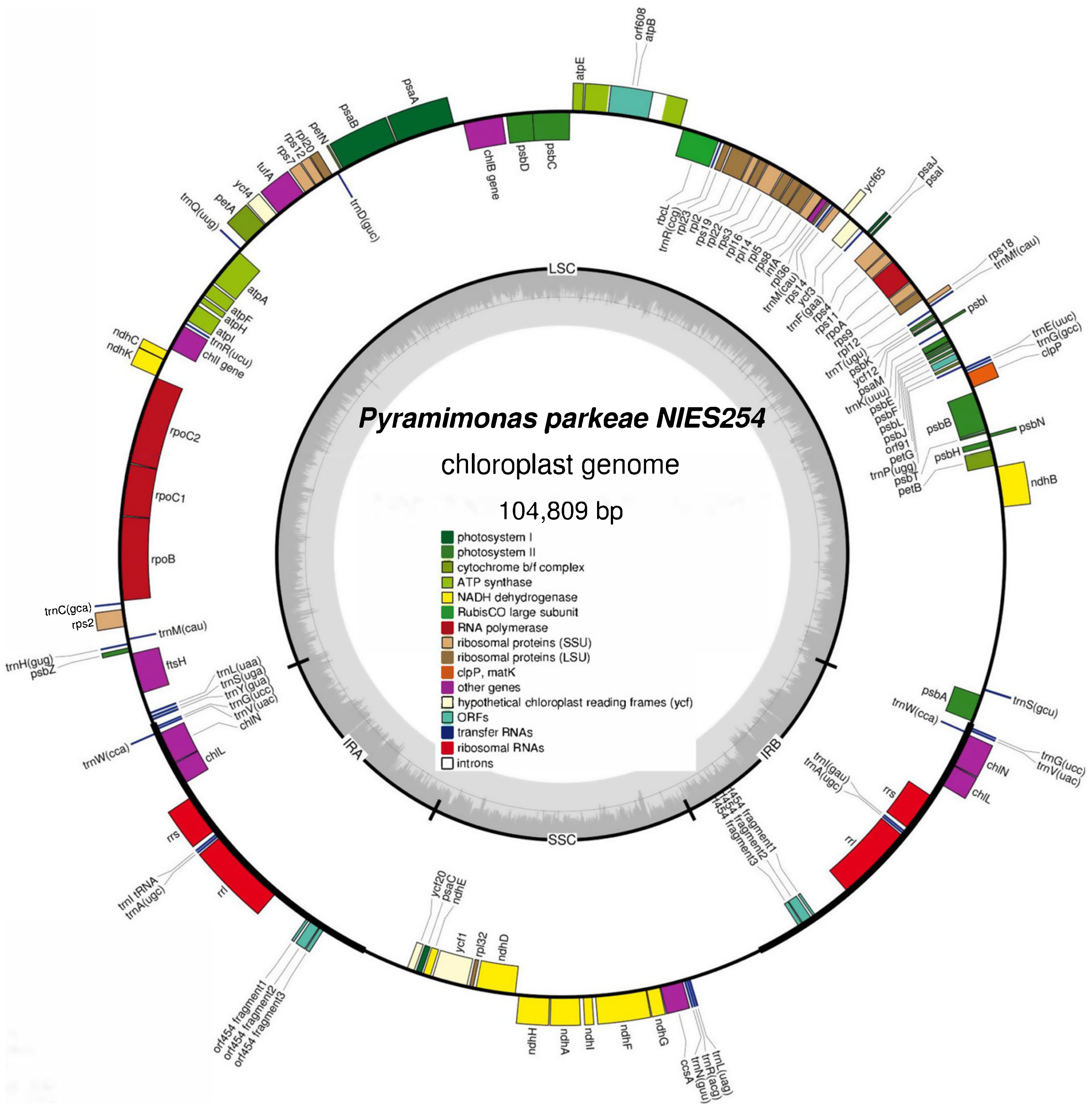

FIG. 1. Map of *Pyramimonas parkeae* NIES254 chloroplast genome. The thick lines indicate the extent of the inverted repeat regions (IRA and IRB), which separate the genome into LSC and SSC regions. Genes outside the map are transcribed counterclockwise and those inside the map are transcribed clockwise. [Color figure can be viewed at wileyonlinelibrary.com]

orf454, *orf608*) present in CCMP726, all were also present in NIES254, though *orf454* present as a single unit in CCMP726 was fragmented into three separate pieces in NIES254, and *orf608* present in the *atpB* introns of both genomes differed in nucleotide sequence. These differences in non-coding sequences may reflect lower constraint than experienced by coding regions. In the prasinophyte species *O. tauri* a group II intron similarly evolved rapidly, resulting in sequence loss in some strains (Blanc-Mathieu et al. 2013).

Comparison of these two *P. parkeae* plastid genomes also indicated IRB-SSC and SSC-IRA boundary movement. The expansion and contraction of the IR regions at the inter-specific level is not uncommon (Goulding et al. 1996). Comparative plastome studies in embryophyte families showed that boundaries between the IR and single copy regions are not static, but rather have been subjected to dynamic and random processes that allow the conservative expansion and contraction of IR regions. Movement of IR boundaries is likely to be unique

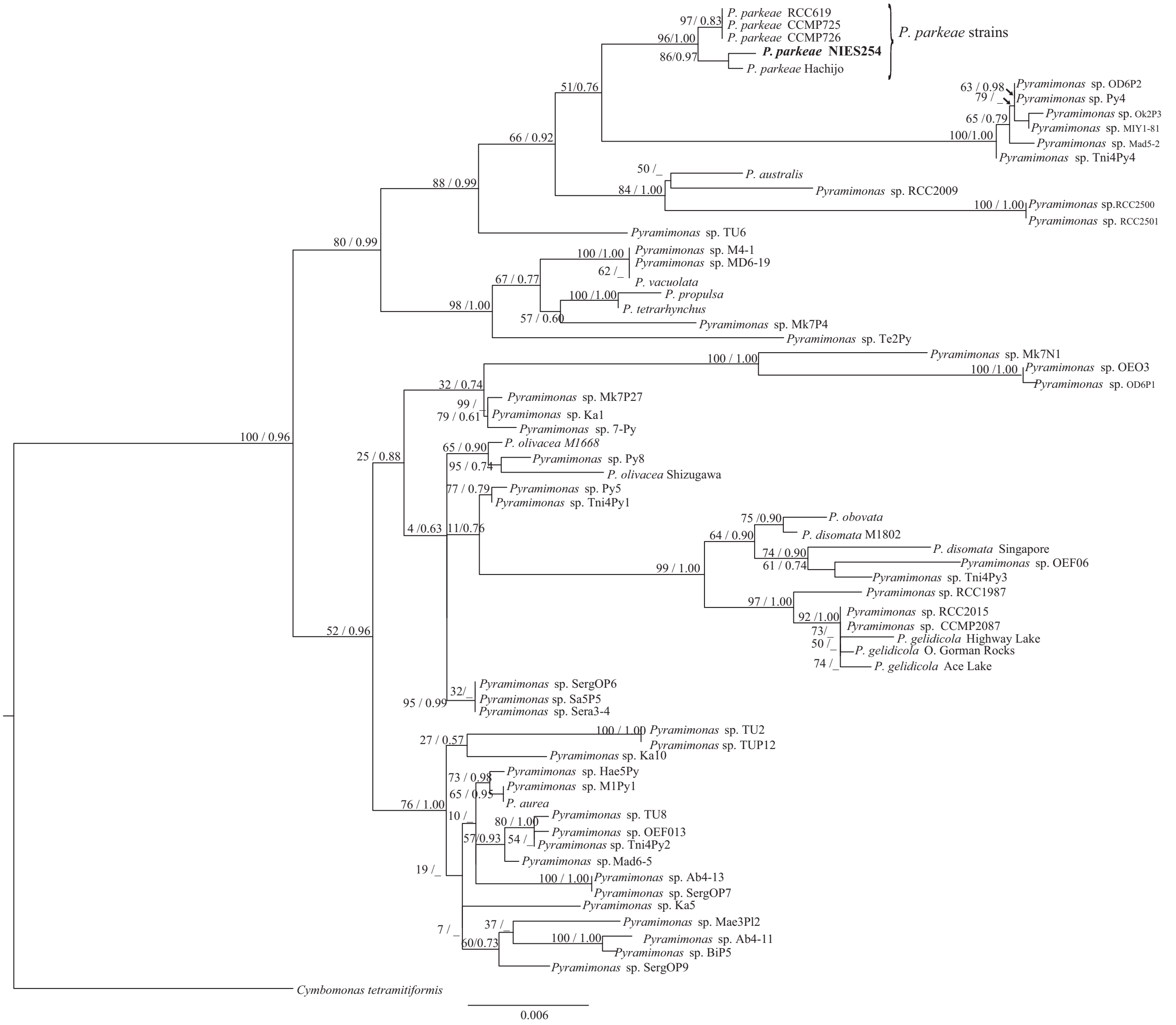

FIG. 2. Maximum-likelihood tree inferred from 18S rDNA sequences of 65 *Pyramimonas* spp. using a GTR+I+F model. The bootstrap and posterior probability values are reported at the respective nodes. The scale bar represents the estimated number of nucleotide substitutions per site. The bracket indicates the monophyletic relationship of *P. parkeae* strains. *Cymbomonas tetramitiformis* was used as an outgroup.

for each species, and hypothesized to reflect relationship among embryophyte families (Zhu et al. 2015, Wang et al. 2016) and contribute to the expansion of the genome (Dugas et al. 2015, Zhu et al. 2016).

Most observations of boundary movements have involved boundary shifts at IRA-LSC and LSC-IRB, which have been hypothesized to be lineage-specific and tend to be minor; across the embryophytes, most such shifts resulted in loss and gain of a few nucleotides or partial genes, which often gave rise to a pseudogene at one end of the borders. While movements of IR-LSC boundaries have been known to be evolutionary markers, IR-SSC boundaries of closely related embryophyte species tend to be static, with shifts involving only a few nucleotides (Zhu et al. 2015). Extreme cases included (i) the medicinal plant *Eucommia ulmoides*, where the IR was expanded by 5 kb in comparison to other angiosperms, and (ii) the legumes *Acacia* and *Inga*, where the IR was expanded by 13 kb. The boundary shifts in *Eucommia ulmoides* were hypothesized to be the result of genome rearrangement, whereas the shifts in the legumes were accompanied by the presence of an increased number of tandem repeats in the genome (Dugas et al. 2015, Wang et al. 2016).

The chloroplast genome IR regions of *P. parkeae* genomes are similar to those of other green algae and embryophytes in clustering *rrl* and *rrs*. However, the IR boundaries of prasinophytes seem less stable

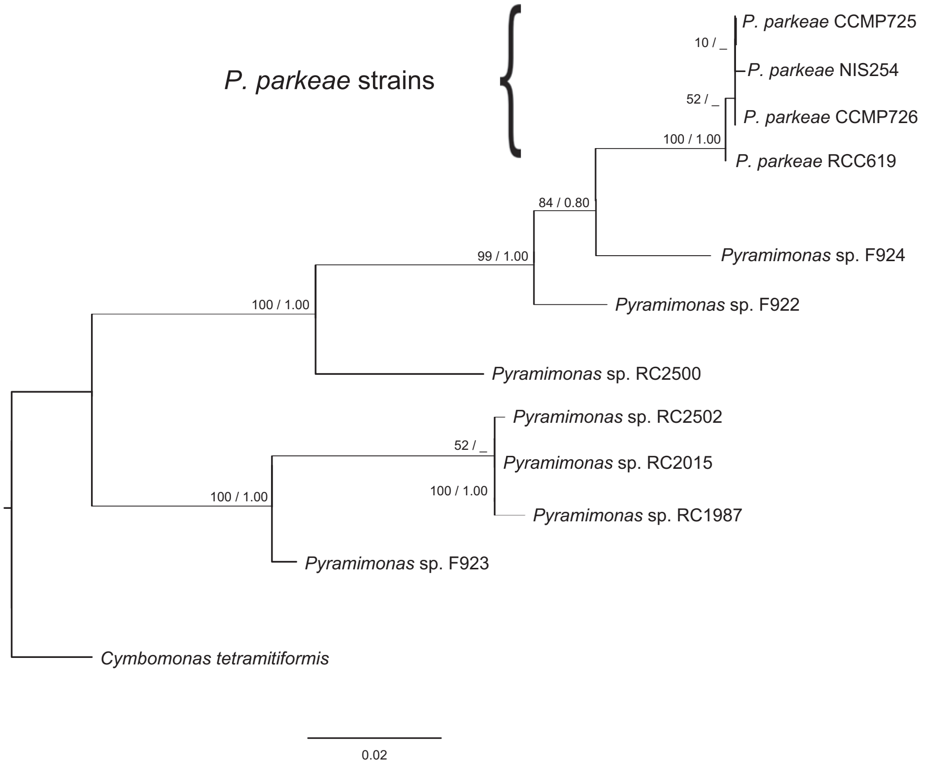

Fig. 3. Maximum-likelihood tree inferred from 16S rDNA sequences of 11 *Pyramimonas* spp. using a GTR+I+F model. The bootstrap and posterior probability values are reported at the respective nodes. The scale bar represents the estimated number of nucleotide substitutions per site. The bracket indicates the monophyletic relationship of *P. parkeae* strains. *Cymbomonas tetramitiformis* was used as an outgroup.

(e.g., Turmel et al. 2009) than in of most land plants (Zhu et al. 2015, Zhu et al. 2016). Our study provides an example of such instability in the form of evidence for boundary movement that has affected both length and copy number of some genes.

A more complete understanding of the mechanism underlying intra-specific plastid genome contraction/expansion will require analysis of additional *Pyramimonas* strains. However, we can speculate about processes that may have been involved in their origin. Contraction of the IRs might be as simple as DNA deletion in one IR copy. This deletion would leave one copy of the IR nucleotides on either LSC or SSC. A more complicated scenario would be IR expansion, which might arise from repair after a double-strand DNA break (Goulding et al. 1996).

Synonymous (Ks) and nonsynonymous (Ka) substitution sites as well as the Ka/Ks ratio calculated from protein coding sequence and a common ORF from *P. parkeae* NIES254 and CCMP726 suggested that mutation in these chloroplast genomes occurred in a random fashion. This contrasts with results of some other studies (Ogihara and Tsunewaki 1988, Birky and Walsh 1992, Zhu et al. 2016), where the observed substitution rates in IR regions were lower than in single copy regions. However, the NIES254 and CCMP726 *P. parkeae* IRs contain *rrl*, *rrs*, and tRNAs clusters that are highly conserved. Therefore, if we include rRNAs and tRNAs in the analyses, the mutation rate will be relatively lower in the IR regions. This depressed substitution rate in the IR regions is hypothesized to provide copy-dependent repair mechanism during the D-loop replication of the chloroplast genome (Zhu et al. 2015, Zhu et al. 2016).

These nucletotide substitutions may alter nucleotide sequences, resulting in change in GC content that if occurring in coding regions, may alter amino acid frequencies and codon usage. Given that no RNA editing processes have as yet been found in the green algae (Stern et al. 2010), we deduced the frequency of amino acid and codon usage based on protein coding sequences and tRNAs. Our results showed that the amino acid frequencies and codon usage differed slightly between the two strains, but the GC content remained the same (data not shown).

Nucleotide substitution rate varies within genes, among genes, and across lineages (Wolfe et al. 1987). Knowing the extent of this variation aids understanding the mode of evolution of protein coding plastid genes in prasinophytes. The observed disproportional increases in Ka/Ks suggest a history of relaxed purifying selection and/or increase in positive selection acting on a subset of plastid genes.

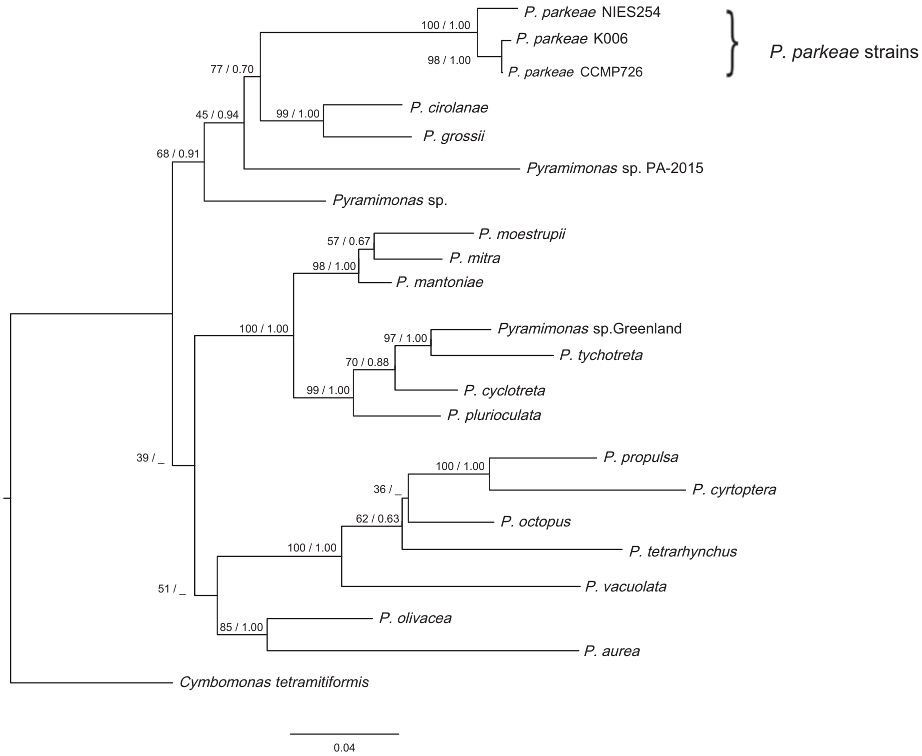

FIG. 4. Maximum-likelihood tree inferred from *rbcL* sequences of 21 *Pyramimonas* spp. using a GTR+I+F model. The bootstrap and posterior probability values are reported at the respective nodes. The scale bar represents the estimated number of nucleotide substitutions per site. The bracket indicates the monophyletic relationship of *P. parkeae* strains. *Cymbomonas tetramitiformis* was used as an outgroup.

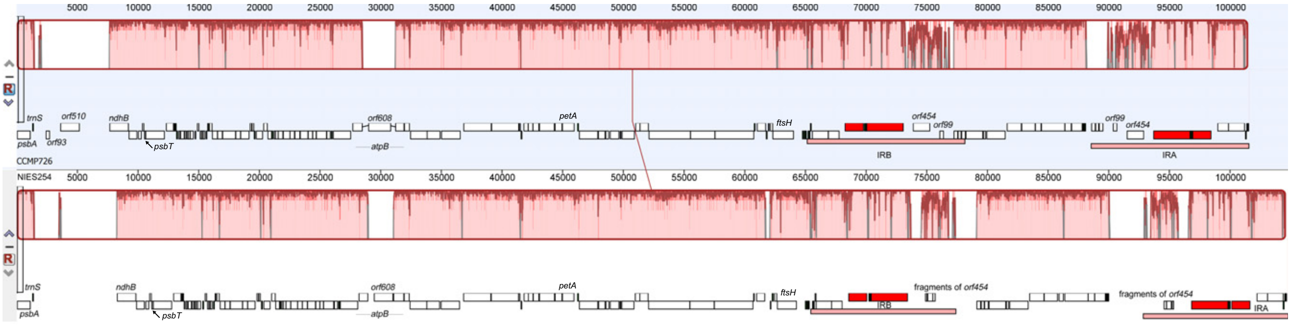

FIG. 5. Mauve alignment of *Pyramimonas parkeae* NIES254 and CCMP726 chloroplast genomes showing shared synteny. The vertical line connecting the two syntenic regions between NIES254 and CCMP726 represents the collinear synteny of the two chloroplast genomes. The histogram inside each block represents pair-wise nucleotide sequence identity. The large four areas where the heights of the histograms are almost equal to zero represent the four large hotspot regions: (1) the 6 kb intergenic region between *psbA-trnS* and *ndhB* in LSC, (2) 2 kb intron of *atpB*, and (3) 5.7 and 6.8 kb located at the boundaries of IR and SSC. [Color figure can be viewed at wileyonlinelibrary.com]

The ratio differences also suggest that changes in selection pressure may be associated with specific biochemical pathways or functions rather than across the entire genome (Magee et al. 2010).

One explanation for observed high variability in prasinophyte chloroplast genomes at the intra-specific level may be long divergence time. It is known that divergence time is correlated with the number of substitutions, because nucleotide substitutions accumulate over time in independent populations. When compared to the prasinophyte *O. tauri* (Blanc-Mathieu et al. 2013), at the intra-specific level, *P. parkeae* chloroplast genomes contained fewer variable positions overall (37,873 positions in *O. tauri* and ~16,700 positions in *P. parkeae* estimated using whole genome Geneious alignment). However, the variability within protein coding sequences of *P. parkeae* was much higher. 3,111 variable positions (2,684 SNPs and 44 indels) were present in protein coding genes of *P. parkeae* while only 153 SNPs were present in that of *O. tauri* (Table S1 in the Supporting Information).

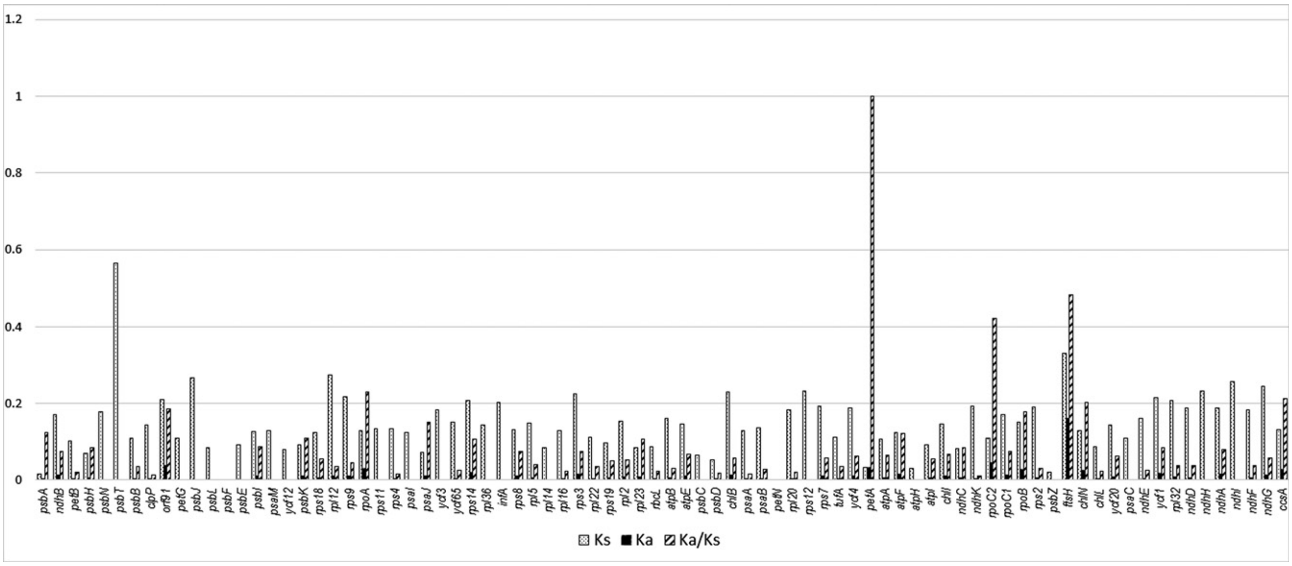

FIG. 6. Comparison of NIES254 and CCMP726 shows that substitution in *Pyramimonas parkeae* chloroplast genomes is unevenly distributed. The *x*-axis shows protein coding genes present in the chloroplast genomes, in genomic order. The *y*-axis is the value for the substitution rate (per 100 bp) and the value for Ka/Ks ratio. The dotted bar indicates synonymous substitution (Ks), the solid bar indicates nonsynonymous substitution (Ka) and the diagonal bar indicates the ratio of nonsynonymous substitution to synonymous substitution (Ka/Ks).

It is also possible that *P. parkeae* chloroplast genomes contain a trait that allows the chloroplast genomes to evolve at a higher rate when compared to those of other organisms in the green lineage. This hypothesis is supported by presence of high intra-specific variability in some euglenoid chloroplast genomes (Bennett and Triemer 2015), which were inherited from a *Pyramimonas*-like chloroplast donor (Palmer 1987, Turmel et al. 2009). Similar to *P. parkeae* chloroplast genomes, those euglenoid chloroplast genomes exhibit intra-specific variability, however, with a higher mutation rate (Bennett and Triemer 2015). It might be possible that a *Pyramimonas*-like chloroplast genome progenitor had a trait that favors mutation and was passed on to its descendants.

Another potential explanation for observed high variability in prasinophyte chloroplast genomes at the intra-specific level is recombination of bi-parentally inherited chloroplast genomes. Evidence for chloroplast DNA recombination has been reported for the prasinophyte *O. tauri* (Blanc-Mathieu et al. 2013). These observations indicate that earliest diverging green algae may display bi-parental chloroplast genome inheritance. If so, uni-parental chloroplast inheritance may have evolved independently in chlorophyte and streptophyte lineages.

Last, but not least, our observation of greater than expected variability between the chloroplast genomes might indicate that NIES254 and CCMP726 are actually different species of *Pyramimonas.* However, phylogenetic analysis of publically available *Pyramimonas* 18S rDNA, 16S rDNA, and *rbcL* sequences were consistent with previous studies (Balzano et al. 2012, Suda et al. 2013) in resolving all *P. parkeae* strains known to date as a monophyletic clade. Additional *Pyramimonas* strains and molecular data may clarify diversification patterns for this ecologically and evolutionarily important genus.

CONCLUSIONS

The availability of a newly sequenced chloroplast genome for *P. parkeae* NIES254 made it possible to examine intra-specific variation in chloroplast genomes in early diverging green algae. Although plastid genomes of CCMP726 and NIES254 have identical gene content, these genomes exhibited some of the highest variability known to occur at the intra-specific level in the green lineage: (i) the NIES254 chloroplast genome is longer than that of CCMP726 by 3,024 bp; (ii) there are four large hot-spot regions where the similarity value between the two studied strains is close to zero; (iii) IR boundaries have shifted and (iv) boundaries of the IR at the IR-SSC junction have undergone contraction or expansion for not just a few nucleotides, but for about 2.5 kb, resulting in differences in copy number for the three protein coding genes *ycf20, psaC,* and *ndhE.*

This study was supported in part by NSF grant DEB-1119944. We thank the Laboratory of Genetics for access to computational infrastructure and support.

Allender, C. J., Allainguillaume, J., Lynn, J. & King, G. J. 2007. Simple sequence repeats reveal uneven distribution of genetic diversity in chloroplast genomes of *Brassica oleracea* L. and (n = 9) wild relatives. *Theor. Appl. Genet.* 114:609–18.

Balzano, S., Gourvil, P., Siano, R., Chanoine, M., Marie, D., Lessard, S., Sarno, D. & Vaulot, D. 2012. Diversity of cultured photosynthetic flagellates in the northeast Pacific and Arctic Oceans in summer. *Biogeosciences* 9:4553–71.

Bennett, M. S. & Triemer, R. E. 2015. Chloroplast genome evolution in the Euglenaceae. *J. Eukaryot. Microbiol.* 62:773–85.

Bhuiyan, M. A. H., Faria, D. G., Horiguchi, T., Sym, S. D. & Suda, S. 2015. Taxonomy and phylogeny of *Pyramimonas vacuolata* sp. nov. (Pyramimonadales, Chlorophyta). *Phycologia* 54:323–32.

Birky, C. W. & Walsh, J. B. 1992. Biased gene conversion, copy number, and apparent mutation rate differences within chloroplast and bacterial genomes. *Genetics* 130:677–83.

Blanc-Mathieu, R., Sanchez-Ferandin, S., Eyre-Walker, A. & Piganeau, G. 2013. Organellar inheritance in the green lineage: insights from *Ostreococcus tauri*. *Genome Biol. Evol.* 5:1503–11.

Bolger, A. M., Lohse, M. & Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–20.

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.* 2:240–56.

Darling, A. E., Mau, B. & Perna, N. T. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* 5:e11147.

Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* 9(8):772.

Daugbjerg, N., Moestrup, Ø. & Arctander, P. 1994. Phylogeny of the genus *Pyramimonas* (Prasinophyceae, Chlorophyta) inferred from the *rbcL* gene. *J. Phycol.* 30:991–9.

De Las Rivas, J., Lozano, J. J. & Ortiz, A. R. 2002. Comparative analysis of chloroplast genomes: functional annotation, genome-based phylogeny, and deduced evolutionary patterns. *Genome Res.* 12:567–83.

Decelle, J., Romac, S., Stern, R. F., Bendif, E. M., Zingone, A., Audic, S., Guiry, M. D., Guillou, L., Tessier, D., Le Gall, F. & Gourvil, P. 2015. PhytoREF: a reference database of the plastidial 16S *rRNA* gene of photosynthetic eukaryotes with curated taxonomy. *Mol. Ecol. Resour.* 15:1435–45.

Duanmu, D., Bachy, C., Sudek, S., Wong, C. H., Jiménez, V., Rockwell, N. C., Martin, S. S., Ngan, C. Y., Reistetter, E. N., van Baren, M. J. & Price, D. C. 2014. Marine algae and land plants share conserved phytochrome signaling systems. *Proc. Natl. Acad. Sci. USA* 111:15827–32.

Dugas, D. V., Hernandez, D., Koenen, E. J., Schwarz, E., Straub, S., Hughes, C. E., Jansen, R. K., Nageswara-Rao, M., Staats, M., Trujillo, J. T. & Hajrah, N. H. 2015. Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in *clpP*. *Sci. Rep.* 5:16958.

Goff, S. A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A. E., Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A. & Muir, A. 2011. The iPlant collaborative: cyberinfrastructure for plant biology. *Front. Plant Sci.* 2:34.

Goulding, S. E., Wolfe, K. H., Olmstead, R. G. & Morden, C. W. 1996. Ebb and flow of the chloroplast inverted repeat. *Mol. Gen. Genet.* 252:195–206.

Hahn, C., Bachmann, L. & Chevreux, B. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res.* 41:e129–e129.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C. & Thierer, T. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28:1647–9.

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21:487–93.

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H. H., Rognes, T. & Ussery, D. W. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35:3100–8.

Leliaert, F., Tronholm, A., Lemieux, C., Turmel, M., DePriest, M. S., Bhattacharya, D., Karol, K. G., Fredericq, S., Zechman, F. W. & Lopez-Bautista, J. M. 2016. Chloroplast phylogenomic analyses reveal the deepest-branching lineage of the Chlorophyta, Palmophyllophyceae class. nov. *Sci. Rep.* 6:25367.

Lemieux, C., Otis, C. & Turmel, M. 2014. 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 Genom.* 15:1.

Lemieux, C., Otis, C. & Turmel, M. 2016. Comparative chloroplast genome analyses of streptophyte green algae uncover major structural alterations in the Klebsormidiophyceae, Coleochaetophyceae and Zygnematophyceae. *Front. Plant Sci.* 7:6971.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–60.

Lohse, M., Drechsel, O., Kahlau, S. & Bock, R. 2013. OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res.* 41:W575–81.

Magee, A. M., Aspinall, S., Rice, D. W., Cusack, B. P., Semon, M., Perry, A. S., Stefanović, S., Milbourne, D., Barth, S., Palmer, J. D. & Gray, J. C. 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Res.* 20:1700–10.

Marin, B. & Melkonian, M. 2010. Molecular phylogeny and classification of the Mamiellophyceae class. nov. (Chlorophyta) based on sequence comparisons of the nuclear-and plastid-encoded rRNA operons. *Protist* 161:304–36.

Miller, M. A., Pfieffer, W. & Schwartz, T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop (GCE), New Orleans, Louisiana.

Moro, I., La Rocca, N., Valle, L. D., Moschin, E., Negrisolo, E. & Andreoli, C. 2002. *Pyramimonas australis* sp. nov. (Prasinophyceae, Chlorophyta) from Antarctica: fine structure and molecular phylogeny. *Eur. J. Phycol.* 37:103–14.

Nakayama, T., Marin, B., Kranz, H. D., Surek, B., Huss, V. A., Inouye, I. & Melkonian, M. 1998. The basal position of scaly green flagellates among the green algae (Chlorophyta) is revealed by analyses of nuclear-encoded SSU rRNA sequences. *Protist* 149:367–80.

Ogihara, Y. & Tsunewaki, K. 1988. Diversity and evolution of chloroplast DNA in *Triticum* and *Aegilops* as revealed by restriction fragment analysis. *Theor. Appl. Genet.* 76:321–32.

Palmer, J. D. 1987. Chloroplast DNA evolution and biosystematic uses of chloroplast DNA variation. *Am. Nat.* 130:S6–29.

Quinlan, A. R. 2014. BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* 47:11.12.1–12.34.

Robbens, S., Derelle, E., Ferraz, C., Wuyts, J., Moreau, H. & Van de Peer, Y. 2007. The complete chloroplast and mitochondrial DNA sequences of *Ostreococcus tauri*: organelle genomes of the smallest eukaryote are examples of compaction. *Mol. Biol. Evol.* 24:956–68.

Ronquist, F. & Huelsenbeck, J. P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–4.

Satjarak, A., Paasch, A. E., Graham, L. E. & Kim, E. 2016. Complete chloroplast genome sequence of phagomixotrophic green alga *Cymbomonas tetramitiformis*. *Genome Announc.* 4: e00551–16.

Schattner, P., Brooks, A. N. & Lowe, T. M. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 33:W686–9.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–13.

Stern, D. B., Goldschmidt-Clermont, M. & Hanson, M. R. 2010. Chloroplast RNA metabolism. *Annu. Rev. Plant Biol.* 61:125–55.

Suda, S. 2004. Taxonomic characterization of *Pyramimonas aurea* sp. nov. (Prasinophyceae, Chlorophyta). *Phycologia* 43:682–92.

Suda, S., Bhuiyan, M. A. H. & Faria, D. G. 2013. Genetic diversity of *Pyramimonas* from Ryukyu Archipelago, Japan (Chlorophyceae, Pyramimonadales). *J. Mar. Sci. Technol.* 21:285–96.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30:2725–9.

Turmel, M., de Cambiaire, J. C., 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:e0148934.

Turmel, M., Ehara, M., Otis, C. & Lemieux, C. 2002. Phylogenetic relationships among streptophytes as inferred from chloroplast small and large subunit rRNA gene sequences. *J. Phycol.* 38:364–75.

Turmel, M., Gagnon, M. C., O'Kelly, C. J., Otis, C. & Lemieux, C. 2009. The chloroplast genomes of the green algae *Pyramimonas*, *Monomastix*, and *Pycnococcus* shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. *Mol. Biol. Evol.* 26:631–48.

Turmel, M., Otis, C. & Lemieux, C. 1999. The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes. *Proc. Natl. Acad. Sci. USA* 96:10248–53.

Turmel, M., Otis, C. & Lemieux, C. 2015. Dynamic evolution of the chloroplast genome in the green algal classes Pedinophyceae and Trebouxiophyceae. *Genome Biol. Evol.* 7:2062–82.

Wang, L., Wuyun, T. N., Du, H., Wang, D. & Cao, D. 2016. Complete chloroplast genome sequences of *Eucommia ulmoides*: genome structure and evolution. *Tree Genet. Genomes* 12:1–15.

Wolfe, K. H., Li, W. H. & Sharp, P. M. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. USA* 84:9054–8.

Worden, A. Z., Lee, J. H., Mock, T., Rouzé, P., Simmons, M. P., Aerts, A. L., Allen, A. E., Cuvelier, M. L., Derelle, E., Everett, M. V. & Foulon, E. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas. Science* 324:268–72.

Zhu, A., Guo, W., Gupta, S., Fan, W. & Mower, J. P. 2015. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytol.* 209:1747–56.

Zhu, A., Guo, W., Gupta, S., Fan, W. & Mower, J. P. 2016. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytol.* 209(4):1747–56.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web site:

Table S1. Variable positions present within the protein coding regions of *Pyramimonas parkeae* and *Ostreococcus tauri* identified using Geneious alignments v 9.0.4 (Kearse et al. 2012). Synonymous (Ks) and nonsynonymous (Ka) substitution sites as well as the Ka/Ks ratio using MEGA6 v 6.06 (Tamura et al. 2013).