RESEARCH ARTICLES

The Complete Chloroplast Genome of the Chlorarachniophyte *Bigelowiella natans*: Evidence for Independent Origins of Chlorarachniophyte and Euglenid Secondary Endosymbionts

Matthew B. Rogers, Paul R. Gilson,† Vanessa Su,‡ Geoffrey I. McFadden,‡ and Patrick J. Keeling**

*Botany Department, University of British Columbia, British Columbia, Canada; †The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia; and ‡School of Botany, University of Melbourne, Victoria, Australia

Chlorarachniophytes are amoeboflagellate cercozoans that acquired a plastid by secondary endosymbiosis. Chlorarachniophytes are the last major group of algae for which there is no completely sequenced plastid genome. Here we describe the 69.2-kbp chloroplast genome of the model chlorarachniophyte *Bigelowiella natans*. The genome is highly reduced in size compared with plastids of other photosynthetic algae and is closer in size to genomes of several nonphotosynthetic plastids. Unlike nonphotosynthetic plastids, however, the *B. natans* chloroplast genome has not sustained a massive loss of genes, and it retains nearly all of the functional photosynthesis-related genes represented in the genomes of other green algae. Instead, the genome is highly compacted and gene dense. The genes are organized with a strong strand bias, and several unusual rearrangements and inversions also characterize the genome; notably, an inversion in the small-subunit rRNA gene, a translocation of 3 genes in the major ribosomal protein operon, and the fragmentation of the cluster encoding the large photosystem proteins PsaA and PsaB. The chloroplast endosymbiont is known to be a green alga, but its evolutionary origin and relationship to other primary and secondary green plastids has been much debated. A recent hypothesis proposes that the endosymbionts of chlorarachniophytes and euglenids share a common origin (the Cabozoa hypothesis). We inferred phylogenies using individual and concatenated gene sequences for all genes in the genome. Concatenated gene phylogenies show a relationship between the *B. natans* plastid and the ulvophyte–trebouxiophyte–chlorophyte clade of green algae to the exclusion of *Euglena*. The *B. natans* plastid is thus not closely related to that of *Euglena*, which suggests that plastids originated independently in these 2 groups and the Cabozoa hypothesis is false.

Introduction

Chlorarachniophytes are marine amoeboflagellates belonging to the recently recognized and diverse assemblage of protists called phylum Cercozoa (Bhattacharya et al. 1995; Keeling 2001; Cavalier-Smith and Chao 2003a). Unlike the vast majority of cercozoans, chlorarachniophytes are photosynthetic, having acquired a plastid by secondary endosymbiosis of a green alga. Secondary endosymbiotic events have occurred on multiple occasions in the course of eukaryotic evolution and have involved hosts and endosymbionts from several different eukaryotic groups. Chromalveolates with plastids (cryptomonads, heterokonts, haptophytes, apicomplexa, and dinoflagellates) have secondary plastids derived from a red algal endosymbiont, and these have been hypothesized to trace back to a single endosymbiosis (Cavalier-Smith 1999; Fast et al. 2001; Patron et al. 2004). In contrast, euglenids and chlorarachniophytes have plastids derived from green algal endosymbionts (Gibbs 1978; Ludwig and Gibbs 1989; McFadden et al. 1995; Van de Peer et al. 1996); however, there is no clear indication of what kind of green alga gave rise to either plastid. Chlorarachniophyte and euglenid endosymbionts are most commonly believed to be derived from 2 independent endosymbiotic events (Delwiche 1999; Archibald and Keeling 2002), but they have also been hypothesized to have arisen from a single common secondary endosymbiosis, the so-called Cabozoa hypothesis (Cavalier-Smith 1999; Cavalier-Smith and Chao 2003b). The secondary endosymbiont of chlorarachniophytes is also noteworthy because it has retained its nucleus and genome in a highly reduced form known as a nucleomorph (Gilson et al. 2006). The discovery of nucleomorphs, and the demonstration that they were degenerated algal nuclei, clinched the argument that plastids spread between eukaryotic lineages by secondary endosymbiosis (Whatley 1981), and they are still the source of important clues as to how secondary endosymbiosis works (Douglas et al. 2001; Cavalier-Smith 2002; Gilson and McFadden 2002). In addition to chlorarachniophytes, nucleomorphs are only found in one other group of algae, the cryptomonads, where they are derived from a red alga, (Ludwig and Gibbs 1987; Douglas et al. 1991). The photosynthetic organelle of chlorarachniophytes (and cryptomonads), therefore, contains 2 genomes, a highly reduced eukaryotic genome located within the periplastid space (between the 2 outer eukaryotic-derived membranes and the inner 2 membranes making up the plastid envelope), and a plastid genome within the stroma (McFadden et al. 1997). Proteins that function in the plastid of *Bigelowiella natans* are, by extension, encoded in 3 separate genomes: the nucleus of the host (Deane et al. 2000; Archibald et al. 2003), the nucleomorph (Gilson et al. 2006), and the plastid itself. Ironically, of these genomes, plastid proteins encoded in the nucleomorph and nucleus have been more intensively studied than those encoded in the plastid itself, and overall, the least is known about the plastid genome in general.

Indeed, representative plastid genomes have been sequenced from all major groups of algae, except chlorarachniophytes. Complete genomes are known from at least one member of all 3 groups of primary plastids: glaucophytes, red algae, and green algae (including plants and charophytes). Similarly, plastid genomes have been sequenced from secondary plastids of euglenids, cryptomonads, heterokonts,

Key words: plastid, genome, chlorarachniophyte, phylogeny, endosymbiosis.

E-mail: pkeeling@interchange.ubc.ca.

Mol. Biol. Evol. 24(1):54–62. 2007
doi:10.1093/molbev/msl129
Advance Access publication September 21, 2006

haptophytes, and apicomplexa. In addition, a great deal of data is known from the unusual genome of dinoflagellates, which is difficult to define as a genome because genes are encoded on single-gene minicircles (Zhang et al. 1999).

Here we describe the complete chloroplast genome from the model chlorarachniophyte *B. natans*. At 69.2 kbp, the *B. natans* chloroplast genome is the smallest chloroplast genome known from any photosynthetic eukaryote. Indeed, the *B. natans* plastid genome falls in the size range of plastid genomes from some nonphotosynthetic organisms. However, unlike *B. natans*, these genomes have lost a large number of genes relating to photosynthesis. The *B. natans* plastid genome has lost or transferred a few of the larger genes to the nucleus, but for the most part its reduced size is a result of compaction: small intergenic spaces and the absence of introns. This genome also allowed us to carry out the first phylogenetic analysis with representatives of all major plastid groups and to test the Cabozoa hypothesis. A phylogeny of concatenated plastid proteins was conducted to test the relationship of the *B. natans* chloroplast to those of green algae, in particular euglenids. These analyses placed *B. natans* within the ulvophyte-trebouxiophyte-chlorophyte (UTC) group of green algae, at face value rejecting the Cabozoa hypothesis and supporting 2 independent origins for chlorarachniophyte and euglenid plastids.

Materials and Methods

Genome Sequencing and Annotation

Clones encoding chloroplast genomic DNA sequences were identified from the *B. natans* nucleomorph genome-sequencing project (Gilson et al. 2006) by similarity to genes known to be plastid encoded in most algae and plants. Assembly of these sequences resulted in 61 kbp of plastid sequence in 9 individual fragments. Gaps were filled by polymerase chain reaction amplification from one fragment end to all possible ends until a single, circular mapping contig was acquired. All amplified fragments were cloned into pCR 2.1 vector by TOPO TA cloning (Invitrogen, Carlsbad, CA) and sequenced on both strands. Additional regions of ambiguous sequence were also amplified, cloned, and resequenced. One clone was found to contain a short, repeat-rich region resistant to sequencing in the intergenic region between *psb*E and *atp*I (a *trna*H gene exists in the same intergenic space, but it was not part of the unsequenced region). The region in question was mapped by restriction digestion and determined to be approximately 100 bp in length. It was reamplified and subcloned as progressively smaller fragments, but no additional sequence was obtained, and we concluded the region likely has a highly stable structure making it difficult to sequence and is too small to encode a gene.

All open reading frames larger than 100 bp were identified, and their similarity to known genes was determined by BlastX searches (Altschul et al. 1990). RNA-encoding genes were sought by BlastN searches. tRNA genes were identified using the tRNAscan online server (http://lowelab.ucsc.edu/tRNAscan-SE/). Because most of the genome consists of genes with a high degree of similarity to homologues in other green algal plastids, very few regions of ambiguous annotation remained, but all unassigned regions were searched by BlastN and BlastX.

Pulsed Field Gels

Bigelowiella natans was grown in nutrient supplemented seawater (f/2) bubbled with filter-sterilized air in continuous lighting at 24 °C. The algae were harvested by centrifugation (3000 × g), and the cell pellet was resuspended in 10 mM Tris-HCl, 100 mM ethylenediaminetetraacetic acid (EDTA), 200 mM NaCl, and 0.5% molten low gelling temperature agarose at 37 °C. The mixture was poured into a prechilled plug mold, and once set, the cell plugs were digested in 10 mM Tris-HCl, 400 mM EDTA, 1% *N*-lauryl sarkosyl and 1 mg/ml Pronase E (Sigma, St Louis, MO) for 48 h at 50 °C. The digested chromosome plugs were loaded into 1% agarose gels that were electrophoresed in a CHEF DRIII apparatus (BioRad, Hercules, CA) in 0.5× Tris-borate-EDTA buffer at 14 °C. To separate large chromosomes, the pulse time was 100 s at 100 V for 3 h. This was then ramped over 36 h from 60 to 120 s at 200 V. Smaller chromosomes were separated with a pulse time of 20 s for 16 h at 175 V.

Phylogenetic Analyses

Protein alignments were constructed for all protein-coding sequences identified in the *B. natans* genome using ClustalX (Thompson et al. 1997) and edited in MacClade 4.07. One exception was the *ycf*1 gene, which is highly divergent and proved to be too difficult to align for a meaningful analysis. Phylogenetic trees were generated for all alignments individually using PhyML 2.4.4 (Guindon and Gascuel 2003) with the Dayhoff substitution matrix and rates across sites modeled on a discrete gamma distribution with 8 variable site categories and 1 category of invariable sites. Concatenated data sets were analyzed using PhyML 2.4.4 with the WAG substitution matrix and site-to-site rate variation modeled on a discrete gamma distribution with 4 categories of variable sites and 1 category of invariable sites. Maximum likelihood analyses were also carried out using ProML 3.6 (Felsenstein 1993) with the JTT correction matrix and no gamma correction. Bootstraps for both methods were carried out in the same way. Bayesian analyses of the concatenated data were constructed using MrBayes 3.0b4 (Ronquist and Huelsenbeck 2003) from 300,000 generations with sampling every 100 generations using the WAG substitution model, 4 gamma categories and 1 category of invariable sites. Branch lengths for Bayesian trees were inferred using ProML 3.6 with the JTT correction matrix and site rate variation modeled on a discrete gamma distribution with 4 rate categories with the alpha parameter and invariable sites obtained from PhyML 2.4.4 as above. All analyses were performed on the full data set consisting of 56 proteins and 11,296 characters and a data set with ribosomal proteins excluded resulting in 38 proteins and 9,108 characters. For the data set with ribosomal proteins excluded, 50 burn-in trees were removed from Bayesian analysis, whereas 60 were removed from the full data set.

Approximately unbiased (AU) tests (Shimodaira 2002) were performed on the concatenated data set excluding ribosomal proteins to compare several alternative positions of *B. natans*. Test trees were constructed by optimizing the phylogeny with *B. natans* excluded using MrBayes 3.0b4 with

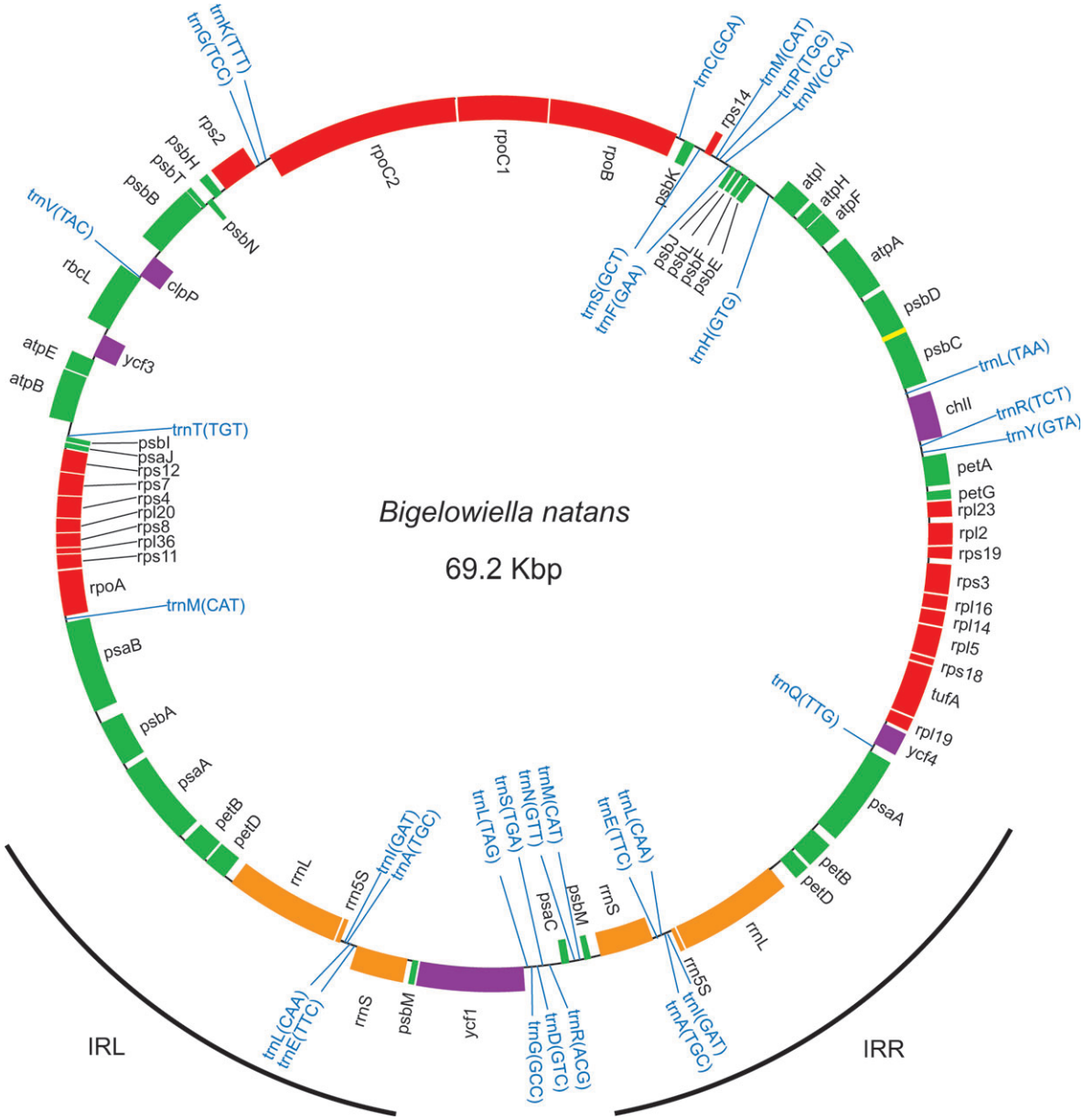

FIG. 1.—Chloroplast genome of *Bigelowiella natans*. Genes on the outside are transcribed in the clockwise direction, and those on the inside are transcribed in the counterclockwise direction. Genes are color-coded according to their function in photosynthesis (green), transcription/translation(red), or miscellaneous (purple). Transfer RNAs are indicated by their anticodon and the amino acid they decode.

same parameters as above (which resulted in an identical topology with the exception of *B. natans* being absent). *B. natans* was then added to 21 alternate positions, including as sister to *Euglena gracilis*. Site likelihoods for each tree were calculated using Tree-Puzzle 5.2 (Schmidt et al. 2002) using the −wsl command with site-to-site rate variation modeled using the parameters from the original data set. AU tests were carried out on site likelihoods using CONSEL 1.19 (Shimodaira and Hasegawa 2001).

Results and Discussion

Genome Structure

The *B. natans* chloroplast genome maps as a circle of 69,166 bp (fig. 1). This is consistent with results from pulsed field gel electrophoresis, from which the size is estimated to be ~70 kbp and which also show the genome exists in complex concatenates (fig. 2). One small (100 bp) region between *psb*E and *atp*I could not be sequenced, but the sequence that was obtained from this intergenic region has several direct and inverted repeats, suggesting the possibility of a stable secondary structure. The overall GC composition of the genome is 30.2%, whereas coding sequences (protein-coding and RNAs) are 32.3% and noncoding is 16.1%, which is not unusual for a plastid genome. The genome includes inverted repeats of 9,380 bps comprising the small-subunit (SSU), long-subunit (LSU), and 5S rRNA genes, several tRNAs, and genes encoding PsbM, PetD, PetB, and the large photosystem I apoprotein PsaA. With the exception of 2 protein-coding genes and

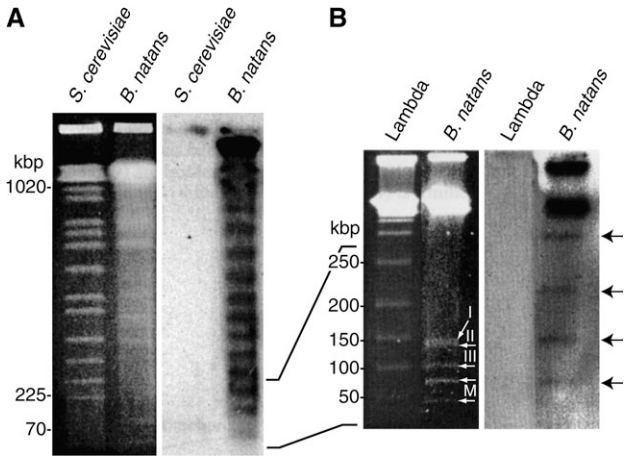

FIG. 2.—The chloroplast chromosome of *Bigelowiella natans* occurs as 69 kbp concatamers. (*A*) An ethidium bromide–stained gel of the chromosome-sized DNA molecules of *B. natans* electrophoresed beside the chromosomes of *Saccharomyces cerevisiae* (left). These were blotted and probed with the *B. natans rbc*L gene (right). The chloroplast chromosome migrates as linear 69 kbp concatamers that are probably the products of the breakage of chloroplast DNA circles. (*B*) The sizes of the chloroplast DNA molecules as shown by Southern blot (right and indicated by arrows) can be seen in comparison to the ethidium bromide–stained mitochondrial (M) and nucleomorph chromosomes (I, II, and III) when the pulsed field gel was run under different conditions.

5 tRNA-encoding genes, the remainder of the genome is contained in the large single-copy (LSC) region. The small single-copy (SSC) region is reduced in comparison to that of plants and other algae, being only 4,124 bps (compared with typical SSCs which are between 10 and 20 kbp), and encoding only genes for PsaC and Ycf1. Many genes that are typically present in the SSC region are missing from the *B. natans* genome, including protein-coding genes such as *rpl*32, *cys*T, and the NDH cluster.

Many unusual or unique rearrangements in gene order are also found in the *B. natans* chloroplast genome. Genes for photosystem proteins PsaA and PsaB are contiguous in all plastid genomes with the exception of *Chlamydomonas reinhardtii* and *Pseudendoclonium akinetum*, and this cluster has also been separated in *B. natans*. Similarly, an inversion has occurred in what would normally constitute the rRNA operon, so that the SSU and 5S rRNA genes are on the opposite strand from the LSU gene. The rRNA operon is a characteristic of nearly all genomes, but inversions breaking up the operon are a feature of only a few plastid genomes, for instance, the apicomplexa (Wilson et al. 1996; Cai et al. 2003), zygnematales (Turmel et al. 2005), and the trebouxiophyte *Helicosporidium* (de Koning and Keeling 2006). In the ulvophyte *P. akinetum*, the entire operon has also inverted, so it is transcribed in the direction of the LSC region (Pombert et al. 2005).

A more unusual rearrangement has occurred in the major ribosomal protein operon. This cluster is conserved in many plastids and cyanobacteria, although losses have occurred in several lineages as well as lineage-specific fission and fusion events (Stoebe and Kowallik 1999). In most green algae and plants, a cluster of 12 genes between *rpl*23 and *rpo*A and a smaller cluster of *rps*12, *rps*7, and *tuf*A is all that remains of the original cyanobacterial operon. In *E. gracilis, rpo*A has been transferred to the nucleus, and the genes for the remaining proteins have moved to other parts of the plastid genome, whereas in *C. reinhardtii* the operon ends at *rps*8. In *B. natans*, the operon has been split in a way similar to that seen in *C. reinhardtii*, except *rps*8 has been translocated as well and the operon ends with *rpl*5. The gene order at the 3′ end of the operon (*rps*8 to *rpo*A) remains conserved but has also been translocated.

Both the SSC and LSC exhibit some degree of strand bias that centers around the inversion of the rRNA genes. The rRNA genes are transcribed convergently and so are the protein-coding genes flanking them in the SSC and many of the protein-coding genes in the LSC. In the right half of the LSC in figure 1, all genes are transcribed toward the SSU rRNA, as are the block of genes proximal to the SSU rRNA on the left half of the LSC. The overall pattern of the genome has 2 points of divergence, one between *psb*E and *atp*I (the repeat-rich region that could not be sequenced) and another between *psa*C and *ycf*1, and 2 points of convergence between SSU and 5S rRNA. Strand bias has been observed in other plastid genomes where genes tend to be transcribed away from the origin of replication, for example, *E. gracilis* and *Helicosporidium* sp. (Hallick et al. 1993; de Koning and Keeling 2006), although other explanations seem to apply to other genomes (Cui et al. 2006). We have no direct evidence for a putative origin of replication in the *B. natans* plastid genome, but the strand bias and the existence of several direct and inverted repeats in the region between *psb*E and *atp*I (the unsequenceable region and also one of the 2 regions where transcription tends to diverge), all suggest this intergenic space is a good candidate.

Gene Content, Loss, and Compaction

The *B. natans* plastid genome is considerably smaller than that of other photosynthetic eukaryotes and marginally smaller than that of the nonphotosynthetic parasitic plant *Epifagus virginiana* (Wolfe et al. 1992). This reduction in size can be attributed to both gene loss and gene compaction (fig. 3).

In terms of gene loss, the genome contains more genes than any nonphotosynthetic plastid but fewer than any other photosynthetic plastid. We identified 57 protein-coding genes, 4 of which are duplicated in the inverted repeat (giving a total of 61). Although this is less than any of the photosynthetic plastids, it is comparable with the 66 genes in the much larger genome of *E. gracilis* or the 69 in *C. reinhardtii*, the largest completely sequenced chloroplast genome. Overall, there has been some gene loss in the *B. natans* genome, but not much—compared with all other publicly available green algal genomes, 8 genes common to all these algae are absent in *B. natans* (Supplementary Table 1, Supplementary material online). When this comparison is expanded to include *E. gracilis*, only 3 genes common to this group are absent in *B. natans*, and if this is further expanded to include photosynthetic plant genomes, only one loss is unique to *B. natans*, *psb*Z. All genes encoding photosystem proteins found in any other green algae have been retained, with the exception of *psa*I, *psa*M, and *psb*Z. Additionally, all the cytochrome components found in other green algae with the exception of *pet*L have

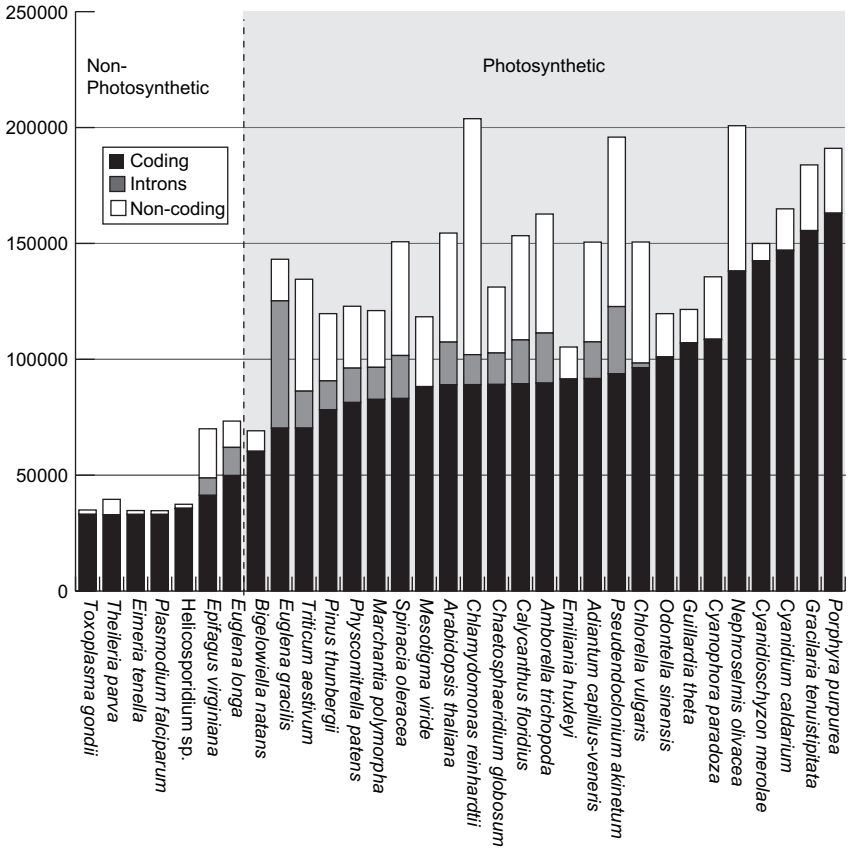

FIG. 3.—Histogram representing plastid genome size and coding capacity ranked by coding capacity. Bars are divided into protein- and RNA-coding sequence (black, bottom), intron content (gray, middle), and intergenic spacers (white, top). Genomes are ranked according to the amount of coding DNA and ordered from left to right by increasing amount of coding sequence.

been retained, as have all the adenosine triphosphate synthase genes found in other green algae. The complement of ribosomal protein genes is also similar to that of green algae. *rpl*12, *rpl*32, and *rps*9 are absent, but none of these losses are unique among green algae or streptophytes. Nucleus-encoded genes for plastid-targeted *rpl*12 and *rps*9 have already been reported from *B. natans* (Archibald et al. 2003). Interestingly, these 2 genes are contiguous in other green algae with the exception of *C. reinhardtii* and *Mesostigma viride*, raising the possibility that these 2 genes may have been transferred to the host nucleus together.

Much of the reduction in gene content in *B. natans* comes from genes with unidentified function (*ycf* genes) and genes with miscellaneous functions in chlorophyll biosynthesis, (*chl*B, *chl*L, and *chl*N), cytochrome biogenesis (*ccs*A), fatty acid metabolism (*acc*D), and cell division (*fts*I, *fts*H, *fts*W, *min*D, and *min*E). Also absent are all genes for NDH proteins, which are absent from green algae with the exception of *Nephroselmis olivacea*. Many of these genes are often found in the SSC region, which is considerably reduced in *B. natans*. Parallel gene loss has been shown to be relatively common in plastid genome evolution (Martin et al. 1998). We plotted gene losses on tree topologies inferred by Bayesian and likelihood methods (see below) and found that 23 losses are predicted to have occurred in the lineage leading to *B. natans* because it diverged from its last common chlorophyte ancestor (Supplementary Figure 1, Supplementary material online). This is more than most lineages but comparable with *E. gracilis* and *C. reinhardtii*.

As alluded to earlier, gene loss only partly accounts for the small size of the *B. natans* plastid; the *B. natans* plastid genome is also unusually gene dense. To illustrate this point, compare the *B. natans* genome with those of *E. gracilis* and *C. reinhardtii*, both of which encode similar numbers of genes but are much larger (fig. 3). Whereas the *E. gracilis* genome is characterized by large numbers of introns, the *B. natans* genome contains no introns whatsoever, not even the typically conserved tRNALeu intron also found in cyanobacteria (Kuhsel et al. 1990). The *C. reinhardtii* genome has an average intron content but has large intergenic spaces. In contrast, the intergenic spaces in the *B. natans* genome are severely reduced. Average intergenic space is only 91 bp, which is comparable with the apicoplast genomes of parasites such as *Theileria parva* and the large gene-rich plastids of red algae and the heterokont *Odontella sinensis*.

tRNA Genes

The *B. natans* plastid genome encodes 27 tRNAs. One species of tRNA is found for each amino acid, except for serine and glycine, which have 2 each, and leucine and methionine, which have 3 each. The 3 methionyl-tRNAs

correspond to the initiator-tRNA (f-Met), the elongation methionyl-tRNA, and the modified isoleucyl-tRNA. Intriguingly, the euglenids *E. gracilis* and *Euglena longa* possess the exact same complement of tRNAs. With wobble rules considered, this complement of tRNAs is near but not exactly the minimum set of tRNAs (de Koning and Keeling 2006), so why do they share the same set? Although the tRNA content between *B. natans* and euglenids is the same, chloroplasts genomes have descended from an already limited subset of tRNAs, so such convergence may not be unlikely when considering the limited subset of tRNAs present in all plastid genomes.

Phylogenetic Relationship to Other Plastid Genomes

To investigate the origin of the *B. natans* endosymbiont, we have inferred phylogenetic trees from concatenated data sets of nearly all protein-coding genes in the chloroplast genome, as well as individual gene phylogenies for each protein-coding gene. Individual phylogenies were reconstructed for 56 of the 57 protein-coding genes in the *B. natans* plastid genome. One protein, Ycf1, was not included in the analysis as it proved too divergent and difficult to align. Overall, most of the individual phylogenies place *B. natans* within the Chlorophyta with good support but without any consensus as to which group of green algae is sister to *B. natans* (not shown).

Analyses of concatenated genes were also carried out. Several previous studies have used concatenated plastid proteins to address a variety of questions, and one issue that has emerged is the divergent nature of the ribosomal proteins and their potentially misleading contribution to the phylogeny. This was recently shown relating to the monophyly of the chromists (Hagopian et al. 2004). We have accordingly inferred phylogenies using both the full set of 56 proteins (11,296 characters) and a slightly reduced set excluding the ribosomal proteins (38 proteins and 9,108 characters). The tree of concatenated proteins excluding ribosomal proteins is shown in figure 4. Overall, the tree resembles other analyses of similar data (Martin et al. 2002; Hagopian et al. 2004; Matsuzaki et al. 2004), with well-supported groups for the red plastid lineage (with a monophyletic and well-supported chromist subgroup), and distinct streptophyte and chlorophyte groups. The glaucophyte *Cyanophora paradoxa* branches as sister to green algae and plants, a topology that has been recovered in similar analyses with ribosomal proteins excluded (Hagopian et al. 2004). *B. natans* branches definitively within the Chlorophyta and, more specifically, within the clade consisting of ulvophytes, trebouxiophytes, and chlorophytes (collectively the UTC group), although the position of *B. natans* with regard to specific members of the UTC clade is equivocal. The branching order within the UTC has previously been shown to differ between analyses—in figure 4, the chlorophyte *C. reinhardtii* branches first, in accordance with recent analyses based on concatenated chloroplast-encoded genes from green algae (Pombert et al. 2005). Significantly, *E. gracilis* was never observed to branch within the well-supported UTC/chlorarachniophyte clade.

Analyses using the entire 56-gene data set were also performed, and no difference was found in most of the well-supported branches, with the exception of chromists, which did not emerge as a monophyletic clade using the full data set (supplementary Figure 2, Supplementary material online). The UTC clade including *B. natans* was recovered with similar bootstrap support. Because plastid genomes encode slightly different repertoires of proteins, in particular *B. natans* and *E. gracilis*, we also constructed PhyML trees from concatenated data sets with all gaps removed. These trees were based on 9,107 characters, and they produced similar topologies. In particular, PhyML support for the UTC clade including *B. natans* remained 98% (not shown).

AU tests were performed on the data set excluding ribosomal proteins to compare several different positions of *B. natans*, including a sister relationship to *E. gracilis* (i.e., the Cabozoa hypothesis) and a basal relationship to all Chlorophyta. All alternative topologies were rejected at the 5% confidence level, except 2 topologies, a sister relationship between *B. natans* and the entire UTC clade and a sister relationship between *B. natans* and chlorophytes.

Origin of Chlorarachniophyte Plastids

With the aim of explaining plastid diversity with as few endosymbiotic events as possible, the Cabozoa hypothesis suggested that the green algal endosymbionts of chlorarachniophytes and euglenids shared a common origin (Cavalier-Smith 1999, 2003). Chlorarachniophytes and euglenids are thought to belong to 2 different supergroups of eukaryotes that are principally nonphotosynthetic, the chlorarachniophytes to the Rhizaria and the euglenids to the Excavata (see Keeling et al. 2005 for review). Excavates include a diversity of nonphotosynthetic groups like diplomonads, retortamonads, parabasalids, oxymonads, and jakobids. Euglenids are the only photosynthetic excavates and are known to be specifically related to a subgroup of nonphotosyntehtic excavates, kinetoplastids and diplonemids. Rhizaria comprises foraminiferans, cercozoans, and some radiolarians and heliozoans. Like excavates, Rhizaria are primarily nonphotosynthetic. Chlorarachniophytes are the only Rhizaria known to have secondary endosymbionts, though the thecate filose amoeba *Paulinella chromatophora* has a cyanobacteria-derived photosynthetic organelle unrelated to the primary plastids of other eukaryotes (Marin et al. 2005). Like euglenids, chlorarachniophytes are derived cercozoans; recent phylogenies of the Cercozoa suggest that they are a sister group to filosa (Bass et al. 2005). Taken to its necessary conclusion, the Cabozoa hypothesis predicts that excavates and Rhizaria share a common photosynthetic ancestor and therefore that the majority of both excavates and rhizarians have lost photosynthesis.

The improbability of these multiple losses of photosynthesis is, in the Cabozoa hypothesis, counterbalanced by the improbability of secondary symbioses occurring twice, given the difficulties implicit in the *de novo* evolution of targeting machinery in independent lineages (Cavalier-Smith 1999). This is a difficult argument to sustain because we have no appreciation of the relative probabilities of these 2 events. Indeed, "plastid loss" is arguably more difficult than gain because an organism could become dependent on nonphotosynthetic metabolic pathways such

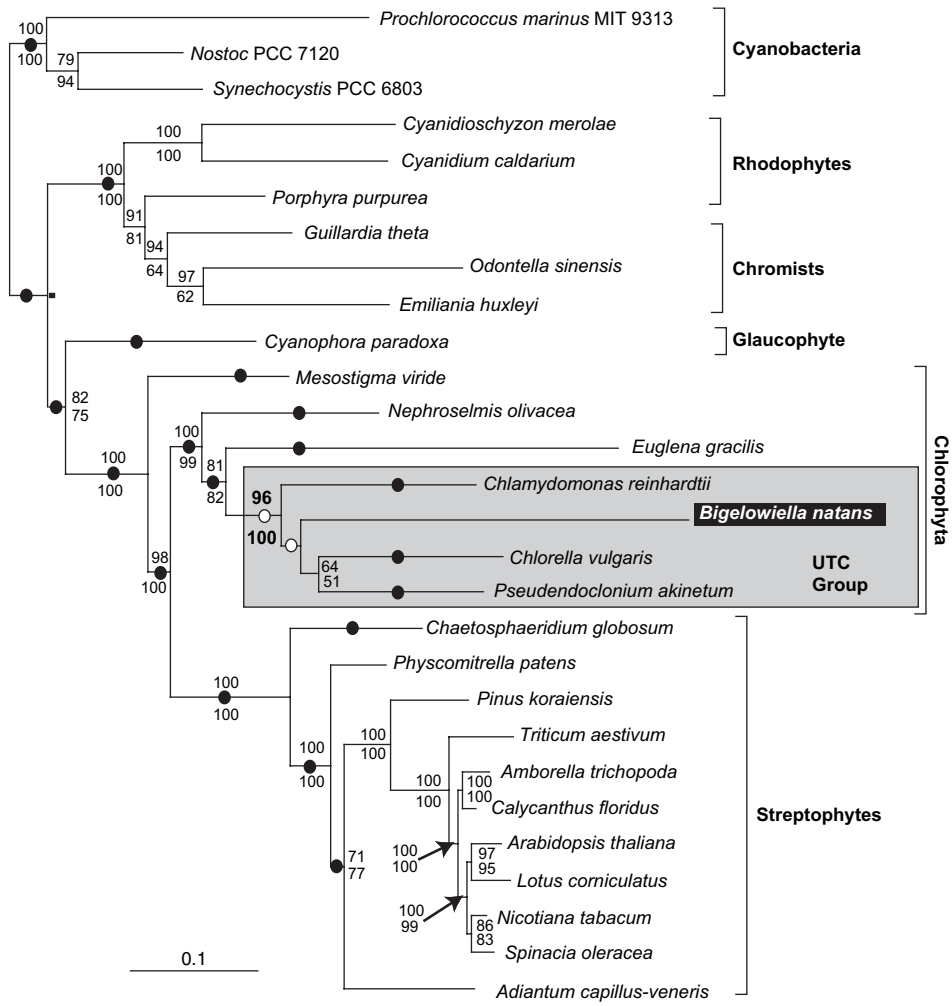

FIG. 4.—Protein maximum likelihood tree of concatenated plastid-encoded genes. The tree was constructed from 38 proteins amounting to 9,108 amino acid positions (complete set excluding ribosomal proteins). The tree topology was inferred using Bayesian analysis with maximum likelihood branch lengths. Numbers at nodes correspond to bootstrap support from ProML (top) and PhyML (bottom). Distance analyses carried out on the same alignment with missing data removed recovered the *Bigelowiella natans* plus UTC clade with 100% support. Filled circles correspond to alternate topologies that failed AU tests at a 5% confidence level, and open circles indicate topologies that cannot be rejected at a 5% confidence level.

as fatty acid, isoprenoid, and heme biosynthesis that plastids can bring with them. Even making the important distinction between plastid loss and "photosynthetic loss," the Cabozoa hypothesis demands many plastid loss events in organisms with complete or near complete genome sequences from which no plastid data are in evidence (e.g., typanosomes, trichomonads, and diplomonads).

The Cabozoa hypothesis makes no predictions about what kind of green alga gave rise to the chlorarachniophyte and euglenid endosymbionts, it does require that they are related to the exclusion of other green algae. It is of course possible that rhizarians and excavates do share a common ancestor (there are currently no data supporting or refuting this), but the Cabozoa hypothesis also requires that their common ancestor already had a plastid. Therefore, plastid sequence data can potentially disprove the Cabozoa hypothesis by showing one or both of chlorarachniophyte or euglenid plastids is more closely related to any other green algal plastid than they are to one another. Our concatenated analyses support a close relationship between *B. natans* and UTC green algal plastids to the exclusion of *E. gracilis*, arguing against a single secondary endosymbiosis of green plastids and the Cabozoa hypothesis.

Conclusions

The chloroplast genome of *B. natans* is the first chloroplast genome from a chlorarachniophyte, the last major algal lineage for which a chloroplast genome has not been sequenced. It is also the smallest chloroplast genome known to date from a photosynthetic eukaryote, although it encodes most of the genes found in other photosynthetic green algae and plants. Chloroplast genomes of photosynthetic green algae display a large variation in size (Simpson and Stern 2002), those completely sequenced range between 150 and 200 kbp, but this may be only a small subset of the diversity that exists. Restriction digests suggest that the chloroplast genome of the ulvophyte, *Acetabularia mediterranea*, is larger than 400 kbp (Tymms and Schweiger 1985), and physical maps of the plastid genome of the ulvophyte *Codium fragilis* suggest that it is only 89 kbp

(Manhart et al. 1989). Although the genome of *B. natans* is smaller than any chloroplast genome yet reported, it is possible that the discrepancy in size between the genome of *B. natans* and that of other green algae may not be so dramatic and that *B. natans* simply lies at the lower end of a diverse spectrum of algal chloroplast genome sizes.

The origin of the chlorarachniophyte endosymbiont has been a topic of controversy since its discovery. Pigment composition was used to suggest a prasinophyte origin of the endosymbiont (Sasa et al. 1992). In contrast, molecular data has suggested a trebouxiophyte (Van de Peer et al. 1996) and, more recently, an ulvophyte origin of the *B. natans* endosymbiont. (Ishida et al. 1997, 1999). Our analyses do not distinguish between an ulvophyte, trebouxiophyte, or chlorophyte origin for the endosymbiont, but they do preclude a prasinophyte, streptophyte, or deeper chlorophyte origin of the chlorarachniophyte plastid. Similarly, we recover no support for a clade of chlorarachniophytes and euglenids, arguing against the Cabozoa hypothesis. Taken together, our data suggests that the plastids of chlorarachniophytes are related to a derived group of green algae and that the plastids of euglenids and chlorarachniophytes are of distinct and independent origin.

Supplementary Material

Supplementary Table 1 and Supplementary Figures 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada to P.J.K. (227301) and the National Health and Medical Research Council to G.I.M. P.J.K. is a Fellow of the Canadian Institute for Advanced Research (CIAR), a New Investigator of the Canadian Institutes for Health Research, and Senior Scholar of the Michael Smith Foundation for Health Research. G.I.M. is a Howard Hughes Medical Institute International Scholar, an associate of the CIAR, and an Australian Research Council Professorial Fellow. Nucleotide sequences reported have been deposited in Genbank under accession number DQ851108.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Archibald JM, Keeling PJ. 2002. Recycled plastids: a green movement in eukaryotic evolution. Trends Genet. 18:577–584.

Archibald JM, Rogers MB, Toop M, Ishida K, Keeling PJ. 2003. Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*. Proc Natl Acad Sci USA. 100:7678–7683.

Bass D, Moreira D, Lopez-Garcia P, Polet S, Chao EE, von der Heyden S, Pawlowski J, Cavalier-Smith T. 2005. Polyubiquitin insertions and the phylogeny of Cercozoa and Rhizaria. Protist. 156:149–161.

Bhattacharya D, Helmchen T, Melkonian M. 1995. Molecular evolutionary analyses of nuclear-encoded small subunit ribosomal RNA identify an independent rhizopod lineage containing the Euglyphidae and the Chlorarachniophyta. J Eukaryot Microbiol. 42:64–68.

Cai X, Fuller AL, McDougald LR, Zhu G. 2003. Apicoplast genome of the coccidian *Eimeria tenella*. Gene. 321:39–46.

Cavalier-Smith T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J Eukaryot Microbiol. 46:347–366.

Cavalier-Smith T. 2002. Nucleomorphs: enslaved algal nuclei. Curr Opin Microbiol. 5:612–619.

Cavalier-Smith T. 2003. Protist phylogeny and the high-level classification of Protozoa. Eur J Protistol. 39:338–348.

Cavalier-Smith T, Chao EE. 2003a. Phylogeny and classification of phylum Cercozoa (Protozoa). Protist. 154:341–358.

Cavalier-Smith T, Chao EE. 2003b. Phylogeny of choanozoa, apusozoa, and other protozoa and early eukaryote megaevolution. J Mol Evol. 56:540–563.

Cui L, Leebens-Mack J, Wang LS, Tang J, Rymarquis L, Stern DB, dePamphilis CW. 2006. Adaptive evolution of chloroplast genome structure inferred using a parametric bootstrap approach. BMC Evol Biol. 6:13.

Deane JA, Fraunholz M, Su V, Maier UG, Martin W, Durnford DG, McFadden GI. 2000. Evidence for nucleomorph to host nucleus gene transfer: light-harvesting complex proteins from cryptomonads and chlorarachniophytes. Protist. 151:239–252.

de Koning AP, Keeling PJ. 2006. The complete plastid genome sequence of the parasitic green alga, *Helicosporidium* sp. is highly reduced and structured. BMC Biol. 4:12.

Delwiche CF. 1999. Tracing the thread of plastid diversity through the tapestry of life. Am Nat. 154 (Suppl):S164–S177.

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG. 2001. The highly reduced genome of an enslaved algal nucleus. Nature. 410:1016–1091.

Douglas SE, Murphy CA, Spencer DF, Gray MW. 1991. Cryptomonad algae are evolutionary chimaeras of two phylogenetically distinct unicellular eukaryotes. Nature. 350:148–151.

Fast NM, Kissinger JC, Roos DS, Keeling PJ. 2001. Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. Mol Biol Evol. 18:418–426.

Felsenstein J. 1993. PHYLIP (phylogeny inference package). Distributed by the author. Seattle(WA): University of Washington.

Gibbs SP. 1978. The chloroplasts of *Euglena* may have evolved from symbiotic green algae. Can J Bot. 56:2883–2889.

Gilson PR, McFadden GI. 2002. Jam packed genomes—a preliminary, comparative analysis of nucleomorphs. Genetica. 115:13–28.

Gilson PR, Su V, Slamovits CH, Reith M, Keeling PJ, McFadden GI. 2006. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. Proc Natl Acad Sci USA. 103:9566–9571.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Hagopian JC, Reis M, Kitajima JP, Bhattacharya D, de Oliveira MC. 2004. Comparative analysis of the complete plastid genome sequence of the red alga *Gracilaria tenuistipitata* var. liui provides insights into the evolution of rhodoplasts and their relationship to other plastids. J Mol Evol. 59:464–477.

Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, Spielmann A, Stutz E. 1993. Complete sequence of *Euglena gracilis* chloroplast DNA. Nucleic Acids Res. 21:3537–3544.

Ishida K, Cao Y, Hasegawa M, Okada N, Hara Y. 1997. The origin of chlorarachniophyte plastids, as inferred from phylogenetic

comparisons of amino acid sequences of EF-Tu. J Mol Evol. 45:682–687.
Ishida K, Green BR, Cavalier-Smith T. 1999. Diversification of a chimaeric algal group, the chlorarachniophytes: phylogeny of nuclear and nucleomorph small-subunit rRNA genes. Mol Biol Evol. 16:321–331.
Keeling PJ. 2001. Foraminifera and Cercozoa are related in actin phylogeny: two orphans find a home? Mol Biol Evol. 18:1551–1557.
Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW. 2005. The tree of eukaryotes. Trends Ecol Evol. 20:670–676.
Kuhsel MG, Strickland R, Palmer JD. 1990. An ancient group I intron shared by eubacteria and chloroplasts. Science. 250:1570–1573.
Ludwig M, Gibbs SP. 1987. Are the nucleomorphs of cryptomonads and *Chlorarachnion* the vestigial nuclei of eukaryotic endosymbionts? Ann N Y Acad Sci. 503:198–211.
Ludwig M, Gibbs SP. 1989. Evidence that nucleomorphs of *Chlorarachnion reptans* (Chlorarachniophyceae) are vestigial nuclei: morphology, division and DNA-DAPI fluorescence. J Phycol. 25:385–394.
Manhart JR, Kelly K, Dudock BS, Palmer JD. 1989. Unusual characteristics of *Codium fragile* chloroplast DNA revealed by physical and gene mapping. Mol Gen Genet. 216:417–421.
Marin B, Nowack EC, Melkonian M. 2005. A plastid in the making: evidence for a second primary endosymbiosis. Protist. 156:425–432.
Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA. 99:12246–12251.
Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. Nature. 393:162–165.
Matsuzaki M, Misumi O, Shin IT, Maruyama S, et al. (41 co-authors). 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature. 428:653–657.
McFadden GI, Gilson PR, Douglas SE, Cavalier-Smith T, Hofmann CJ, Maier UG. 1997. Bonsai genomics: sequencing the smallest eukaryotic genomes. Trends Genet. 13:46–49.
McFadden GI, Gilson PR, Waller RF. 1995. Molecular phylogeny of chlorarachniophytes based on plastid rRNA and *rbcL* sequences. Arch Protistenkd. 145:231–239.
Patron NJ, Rogers MB, Keeling PJ. 2004. Gene replacement of fructose-1,6-bisphosphate aldolase (FBA) supports a single photosynthetic ancestor of chromalveolates. Eukaryot Cell. 3:1169–1175.
Pombert JF, Otis C, Lemieux C, Turmel M. 2005. The chloroplast genome sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae) reveals unusual structural features and new insights into the branching order of chlorophyte lineages. Mol Biol Evol. 22:1903–1918.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19: 1572–1574.
Sasa T, Takaichi S, Hatakeyama N, Watanabe MM. 1992. A novel carotenoid ester, loroxanthin dodecenoate, from *Pyramimonas-parkeae* (Prasinophyceae) and a chlorarachniophycean alga. Plant Cell Physiol. 33:921–925.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18: 502–504.
Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 51:492–508.
Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 17: 1246–1247.
Simpson CL, Stern DB. 2002. The treasure trove of algal chloroplast genomes. Surprises in architecture and gene content, and their functional implications. Plant Physiol. 129:957–966.
Stoebe B, Kowallik KV. 1999. Gene-cluster analysis in chloroplast genomics. Trends Genet. 15:344–347.
Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25:4876–4882.
Turmel M, Otis C, Lemieux C. 2005. The complete chloroplast DNA sequences of the charophycean green algae *Staurastrum* and *Zygnema* reveal that the chloroplast genome underwent extensive changes during the evolution of the Zygnematales. BMC Biol. 3:22.
Tymms MJ, Schweiger HG. 1985. Tandemly repeated nonribosomal DNA sequences in the chloroplast genome of an *Acetabularia mediterranea* strain. Proc Natl Acad Sci USA. 82: 1706–1710.
Van de Peer Y, Rensing SA, Maier U-G. 1996. Substitution rate calibration of small subunit ribosomal RNA identifies chlorarachniophyte endosymbionts as remnants of green algae. Proc Natl Acad Sci USA. 93:7732–7736.
Whatley JM. 1981. Chloroplast evolution—ancient and modern. Ann N Y Acad Sci. 361:154–165.
Wilson RJMI, Denny PW, Preiser DJ, et al. (11 co-authors). 1996. Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum.* J Mol Biol. 261:155–172.
Wolfe KH, Morden CW, Palmer JD. 1992. Function and evolution nbsp;of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc Natl Acad Sci USA. 89:10648–10652.
Zhang Z, Green BR, Cavalier-Smith T. 1999. Single gene circles in dinoflagellate chloroplast genomes. Nature. 400: 155–159.

Charles Delwiche, Associate Editor

Accepted September 18, 2006