Split Photosystem Protein, Linear-Mapping Topology, and Growth of Structural Complexity in the Plastid Genome of Chromera velia

Jan Janouškovec,*†,1 Roman Sobotka,‡,2,3 De-Hua Lai,‡,4,† Pavel Flegontov,‡ Peter Koník,3 Josef Komenda,2,3 Shahjahan Ali,⁵ Ondřej Prášil,2,3 Arnab Pain,⁶ Miroslav Oborník,2,3,4 Julius Lukeš,3,4 and Patrick J. Keeling*,1

1Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada
2Institute of Microbiology, Czech Academy of Sciences, Třeboň, Czech Republic
3Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic
4Biology Centre, Institute of Parasitology, Czech Academy of Sciences, České Budějovice, Czech Republic
5Biosciences Core Laboratory-Genomics, King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia
6Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia

†These authors contributed equally to this work.
‡Present address: School of Life Sciences, Sun Yat-Sen University, Guangzhou, China
*Corresponding author: E-mail: janjan.cz@gmail.com; pkeeling@mail.ubc.ca.
Associate editor: Charles Delwiche

Abstract

The canonical photosynthetic plastid genomes consist of a single circular-mapping chromosome that encodes a highly conserved protein core, involved in photosynthesis and ATP generation. Here, we demonstrate that the plastid genome of the photosynthetic relative of apicomplexans, Chromera velia, departs from this view in several unique ways. Core photosynthesis proteins PsaA and AtpB have been broken into two fragments, which we show are independently transcribed, oligoU-tailed, translated, and assembled into functional photosystem I and ATP synthase complexes. Genome-wide transcription profiles support expression of many other highly modified proteins, including several that contain extensions amounting to hundreds of amino acids in length. Canonical gene clusters and operons have been fragmented and reshuffled into novel putative transcriptional units. Massive genomic coverage by paired-end reads, coupled with pulsed-field gel electrophoresis and polymerase chain reaction, consistently indicate that the C. velia plastid genome is linear-mapping, a unique state among all plastids. Abundant intragenomic duplication probably mediated by recombination can explain protein splits, extensions, and genome linearization and is perhaps the key driving force behind the many features that defy the conventional ways of plastid genome architecture and function.

Key words: plastid genome evolution, Chromera velia, split protein, linear-mapping genome.

Introduction

Chromera velia is an autotrophic alveolate that was discovered during a survey of zooxanthellae in Australian coral reefs (Moore et al. 2008). The dominant reef symbionts are dinoflagellates from the genus Symbiodinium, but C. velia was found to be related to the sister group of dinoflagellates, the apicomplexan parasites (Moore et al. 2008; Oborník et al. 2009). Because apicomplexans include a number of medically and economically important pathogens (e.g., the malaria parasite Plasmodium, as well as Toxoplasma, Cryptosporidium, and Babesia), and because of the interest in the cryptic, non-photosynthetic plastid that has now been found in many of these apicomplexans, the nature of the plastid in C. velia was of immediate interest. Accordingly, the complete C. velia plastid genome was characterized and has proven critical in elucidating the origin of the apicomplexan plastid and its relationship to that of dinoflagellates (Janouškovec et al. 2010), and other plastid-related metabolic pathways have also already been compared with those of apicomplexans (Botté et al. 2011; Köreny et al. 2011).

Although these questions have certainly directed much attention to C. velia and its plastid in particular, they have also overshadowed the intrinsic interest in this organism. Aside from being a key to understanding apicomplexan plastids, C. velia is itself potentially interesting and important both ecologically and evolutionarily (Janouškovec et al. 2012). This is because C. velia is one of the few known photosynthetic lineages in the alveolates (the others being Vitrella and dinoflagellates; Oborník et al. 2012), so its plastid represents new breadth in the study of plastid diversity never before
accessible for comparison with other plastids. Indeed, the initial description of the C. velia plastid genome noted several unusual features with implications for plastid evolution and function (Janouškovec et al. 2010), but because they were typically not shared with either apicomplexans or dinoflagellates (and therefore not comparable to them), they have not been characterized further.

Here, we describe three intriguing features of the C. velia plastid genome: the presence of split genes encoding split proteins, divergent characteristics of gene organization and expression, and the physical structure of the chromosome. We show that two functionally important proteins involved in photosynthesis, PsaA representing a core subunit of photosystem I (PSI) and AtpB representing the beta subunit of the ATP synthase, are expressed in two discrete fragments at both the RNA and protein levels, which has interesting implications for the structure and function of these otherwise highly conserved proteins. We also show that the plastid chromosome is highly divergent in structure with a pronounced strand polarity, altered gene order, and large extensions in many coding sequences, which appear transcribed. Finally, we significantly expand the depth of sequence coverage at the DNA level, and show that the coverage pattern at the chromosome ends, polymerase chain reaction (PCR) experiments, genome migration on pulse field gels, and presence of two long terminal inverted repeats (TIRs) all suggest that the genome is linear in structure. This would represent the first documented case of a linear-mapping plastid genome. These unique characteristics substantially expand our current understanding of plastid genome diversity, much of which we hypothesize is due to high levels of intragenomic duplication in this lineage.

**Results**

**Fragmented Genes Encode Fragmented Proteins Expressed from Polycistronic mRNAs**

The complete C. velia plastid genome contained a number of small fragments of genes, including psaA, atpB, psaB, psbB, rpl13, and tufA. In most of these cases, intact homologs were also present, but in psaA and atpB only two fragments were found that were widely separated in the genome and that together would account for the entire gene (psaA-1, psaA-2, atpB-1, and atpB-2). Both gene products are critical for photosynthesis and ATP generation, suggesting three possible explanations, all of which are unusual: the plastid fragments are pseudogenes and intact proteins are imported from the cytosol, the plastid gene products are trans-spliced at the RNA or protein levels, or the proteins function as unique two-subunit forms.

Expression of psaA-1 and psaA-2 fragments was examined by transcript mapping by circularizing mRNAs and Northern analysis (see Materials and Methods). Reverse transcriptase RT-PCR on circularized mRNA produced a product from each of the fragments comprising the coding region, flanking sequence, and a short 6–12 nucleotide-long oligouridylylated (oligoU) tail (fig. 1). The consistent failure to connect the fragments by RT-PCR using multiple primer sets suggested they are not spliced at the RNA level. Both psaA-1 and psaA-2 are surrounded by several genes on the same strand, so their co-ordinated expression was analyzed by hybridization to probes corresponding to psaA-1, psaA-2, the upstream genes, and the downstream noncoding region. All probes hybridized with a fragment of the expected size of a stand-alone oligoU mRNA corresponding to the fragment in question: no evidence for a spliced RNA form was found. However, probes did hybridize to larger fragments (fig 1 and supplementary fig. S1, Supplementary Material online) corresponding in size and hybridization pattern to a single dicistronic (psaA-1 + psbB) or multiple large polycistronic mRNAs (psaA-2 + psbE + psaB, and surrounding genes). Linkages between psaA-1 and psbB and psaA-2 and psaB were confirmed by RT-PCR. Hybridization patterns around psaA-2 suggested processing of large mRNAs into the dicistronic psaA-2 + psbE and subsequently to single gene transcripts (fig. 1). Similar to psaA, all attempts to link the two atpB fragments by RT-PCR yielded no products. The Northern analysis of atpB-1 and atpB-2 revealed single bands corresponding in size to each expected fragment and no evidence was found for a spliced mRNA, again suggesting independent expression (fig. 2A). The transcriptional profiles of both atpB fragments were supportive of this conclusion (fig. 2B). All four psaA and atpB fragments contained structurally conserved domains (fig. 1 and 2A) and were among the top 22 most abundantly transcribed plastid genes (table 1, and see later). Evidence at the genomic and transcriptomic levels therefore consistently suggested that all psaA and atpB fragments are independently transcribed, translated, and code for functional products.

To obtain convincing evidence that psaA and atpB gene fragments are expressed as separate polypeptides, and that no intact version of the proteins is present, we analyzed membrane protein complexes by combination of two-dimensional (2D) electrophoresis and mass spectrometry (MS). Membrane fraction isolated from C. velia cells was solubilized by dodecyl-β-maltoside and protein complexes were separated on Clear-Native gel in the first dimension and on sodium dodecyl sulfate (SDS) gel in the second dimension (fig. 3). The most abundant spots were subjected to MS analysis of their tryptic peptides which were correlated with predicted peptides of plastid-encoded genes and available expressed sequence tag (EST) sequences. This allowed us to distinguish photosystem II (PSII), cytochrome b_{6}f, and PSI and PSII supercomplexes with bound antennas (fig. 3). Fragments corresponding to PsaA-1, PsaA-2, AtpB-1, and AtpB-2 were all identified, but no spot with the expected mass/charge properties of intact PsaA and AtpB was found (see supplementary tables S1 and S2, Supplementary Material online, for a list of peptides assigned to PsaA/B and AtpA/B proteins). Identification of spots was consistent with chlorophyll fluorescence detected in the native gel. All three PSII complexes exhibited strong fluorescence, which contrasted with the minimal fluorescence of chlorophyll bound within the PSI supercomplex. The minimal fluorescence emission from the PSI complex indicated it was well preserved before it was separated into subunits in the second dimension.
**FIG. 1.** Expression model for the split PsA. PsA fragments are separated at the genomic, transcriptomic, and protein level (top to bottom). The top graph shows the transcriptional profile (total RNA in black upper line and polyA RNA in gray lower line) of the genomic region below. In the genomic region, boxes represent genes and gray arrows show the coding DNA strand. Consistent with the transcriptional profile, Northern hybridization blots (boxes below) reveal that both psaA fragments are transcribed within larger polycistronic transcripts (P) and further processed into dicistronic (D) and monocistronic (M) units. The monocistronic psaA transcripts contain oligoU tails and translate into independent peptides both of which participate in PSI (thylakoid trans-membrane domains are shown by vertical boxes and numbered).

**FIG. 2.** Expression model for the split AtpB. (A) Northern analysis revealed a monocistronic transcript for each of the two atpB fragments. (B) Transcriptome coverage of the two genomic regions is shown for total RNA (black upper line) and polyA RNA (gray lower line). Genes are shown by boxes and gray arrows indicate coding strands. Black arrows below indicate predicted transcripts based on the combination of transcript mapping and Northern analysis. (C) Alignment of plastid, cyanobacterial, and yeast mitochondrial AtpBs reveals that the split of *Chromera velia* AtpB occurred within a 24 amino acid insertion that is also present in the sister taxon *Vitrella* (the most probable positions of the split are boxed).
This, together with the lack of evidence for intact genes or transcripts for either gene, suggested strongly that the two separate subunits of both PsaA and AtpB proteins are assembled into functional PSI and ATP synthase, respectively. Modeling the two-subunit forms shows that the point at which both genes were split corresponds to a loop spanning structural domains (fig. 1). In the case of PsaA, the position of this breakpoint corresponds to the largest loop in the protein, which separates the first two pairs of peripheral helices from the rest of the protein including the photochemistry-performing core (fig. 1; Jordan et al. 2001; Green 2003). The split has also driven a considerable change in divergence rate between the two peptides: PsaA-1 at the photosynthetic antenna periphery is significantly less conserved than PsaA-2 at the antenna core. By calculating maximum likelihood-corrected genetic distances (see Materials and Methods), we estimated that PsaA-1 is about 4.6 times more divergent than PsaA-2 relative to other plastid homologs, suggesting that it is, not surprisingly, under weaker functional constraint. Maximum likelihood phylogenies using a wide sampling of plastid and bacterial homologs (see Materials and Methods) also excluded a secondary bacterial origin for PsaA-1 as suggested previously (Mazor et al. 2012): although fast-evolving, both C. velia PsaA-1 and PsaA-2 were most closely related to the sister taxon Vitrella and other plastids.

<table>
<thead>
<tr>
<th>Region</th>
<th>Number/Gene</th>
<th>Length (kb)</th>
<th>Cov/bp</th>
<th>Rel. Cov. vs. Med.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intergenic regions</td>
<td>105</td>
<td>22.4</td>
<td>3444</td>
<td></td>
</tr>
<tr>
<td>Genes</td>
<td>107</td>
<td>94.1</td>
<td>33990</td>
<td></td>
</tr>
<tr>
<td>Ribosomal RNA</td>
<td>3</td>
<td>4.8</td>
<td>348584</td>
<td>475.3</td>
</tr>
<tr>
<td>Transfer RNA</td>
<td>29</td>
<td>2.4</td>
<td>699</td>
<td>1.0</td>
</tr>
<tr>
<td>Protein-coding</td>
<td>75</td>
<td>86.9</td>
<td>12033</td>
<td>16.4</td>
</tr>
<tr>
<td>Photosystem II (psb)</td>
<td>11</td>
<td>6.6</td>
<td>52244</td>
<td>71.2</td>
</tr>
<tr>
<td>Photosystem I (psa)</td>
<td>4</td>
<td>5.7</td>
<td>22740</td>
<td>31.0</td>
</tr>
<tr>
<td>Cytochrome b6/f (pet)</td>
<td>4</td>
<td>2.1</td>
<td>18301</td>
<td>25.0</td>
</tr>
<tr>
<td>ATP synthase (atp)</td>
<td>6</td>
<td>5.1</td>
<td>14762</td>
<td>20.1</td>
</tr>
<tr>
<td>Ribosomal proteins large subunit (rfl)</td>
<td>10</td>
<td>5.4</td>
<td>1141</td>
<td>1.6</td>
</tr>
<tr>
<td>Ribosomal proteins small subunit (rrs)</td>
<td>12</td>
<td>16.8</td>
<td>1328</td>
<td>1.8</td>
</tr>
<tr>
<td>Thylakoid import (sec, tat)</td>
<td>3</td>
<td>4.5</td>
<td>817</td>
<td>1.1</td>
</tr>
<tr>
<td>RNA polymerase (rpo)</td>
<td>4</td>
<td>15.4</td>
<td>293</td>
<td>0.4</td>
</tr>
<tr>
<td>Other function (acsF, ccsA, clpC, tufA, ycf3)</td>
<td>7</td>
<td>11.9</td>
<td>3617</td>
<td>4.9</td>
</tr>
<tr>
<td>Unknown function (orf)</td>
<td>14</td>
<td>13.5</td>
<td>1339</td>
<td>1.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Highly expressed protein-coding genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>psbA</td>
</tr>
<tr>
<td>psaA-2</td>
</tr>
<tr>
<td>psbE</td>
</tr>
<tr>
<td>atpH1</td>
</tr>
<tr>
<td>psbD</td>
</tr>
<tr>
<td>psbC</td>
</tr>
<tr>
<td>petD</td>
</tr>
<tr>
<td>psbB</td>
</tr>
<tr>
<td>psbV</td>
</tr>
<tr>
<td>psaC</td>
</tr>
<tr>
<td>atpA</td>
</tr>
<tr>
<td>petA</td>
</tr>
<tr>
<td>atpB-1</td>
</tr>
<tr>
<td>psbH</td>
</tr>
<tr>
<td>petB</td>
</tr>
<tr>
<td>psaB</td>
</tr>
<tr>
<td>tufA</td>
</tr>
<tr>
<td>petG</td>
</tr>
<tr>
<td>atpB-2</td>
</tr>
<tr>
<td>psbK</td>
</tr>
<tr>
<td>psaA-1</td>
</tr>
<tr>
<td>psbJ</td>
</tr>
</tbody>
</table>

Note.—Genes encoding split proteins are underlined.

aAverage coverage per base pair.

bRelative coverage compared with median gene coverage depth per base pair (=733.4).

cGene/intergenic regions in the TIRs were counted once.
also present in the AtpB of the related dinoflagellate *Amphidinium* (data not shown). Although unrelated in sequence, the notorious divergence of dinoflagellate plastid proteins and their greater distance to *C. velia* and *Vitrella* could easily account for the insertion divergence pointing to a possibly even deeper origin.

**Many Other Plastid Open Reading Frames are Extraordinary and Transcribed**

PsA and AtpB splits stand out as unprecedented cases but represent only a fraction of peculiarities in *C. velia* plastid open reading frames (ORFs). The genome contains many unusually modified ORFs, including several genes that are expected to be present in plastid genomes, but which are unusually large due to the presence of long extensions that bear no recognizable similarity to known genes. In three genes (*rps7, 8*, and *17*), the ORF region has been extended by 126 to 273 amino acids toward the C-terminus effectively tripling them in size (supplementary fig. S2, Supplementary Material online). In four other genes (*rps3, 4, 8*, and *11*), the region sharing homology with other plastid proteins is found at the end of an unusually large ORF (supplementary fig. S3, Supplementary Material online). The most obvious example of this is Rps4, which is encoded within a 1,998 aa long ORF, compared with 201 aa in the red alga *Porphyra purpurea*. In these cases, start codons may be located near the beginning of the homologous region, but it is still puzzling why the homologous region would be preceded by such a long stretch of sequence uninterrupted by stop codons. Canonical Shine-Dalgarno sequences are absent (the predicted 3′-end of the 16S rRNA is highly divergent in *C. velia*) and therefore cannot support start codon prediction. ORFs of two additional genes, *rpoC1* and *rpoC2*, contain long insertions (507 and 971 aa) in their variable regions (none of them appears to represent an intein; supplementary fig. S4, Supplementary Material online). The most obvious example of this is Rps4, which is encoded within a 1,998 aa long ORF, compared with 201 aa in the red alga *Porphyra purpurea*. In these cases, start codons may be located near the beginning of the homologous region, but it is still puzzling why the homologous region would be preceded by such a long stretch of sequence uninterrupted by stop codons. Canonical Shine-Dalgarno sequences are absent (the predicted 3′-end of the 16S rRNA is highly divergent in *C. velia*) and therefore cannot support start codon prediction. ORFs of two additional genes, *rpoC1* and *rpoC2*, contain long insertions (507 and 971 aa) in their variable regions (none of them appears to represent an intein; supplementary fig. S4, Supplementary Material online). The most obvious example of this is Rps4, which is encoded within a 1,998 aa long ORF, compared with 201 aa in the red alga *Porphyra purpurea*.

Altogether, ORFs of various functionally unrelated proteins contain unusual modifications. Many of the proteins are indispensable and exclusively plastid-encoded in algae and plants suggesting that they are unlikely to be in the process of degeneration, loss, or relocation to the nucleus in *C. velia*. Accordingly, none of these proteins can be identified in the available nuclear transcriptomes from *C. velia* (Woehle et al. 2011). To provide first insights into expression of the unusual proteins, we analyzed the plastid transcriptome using total RNA sequence uninterrupted by stop codons. Canonical Shine-Dalgarno sequences are absent (the predicted 3′-end of the 16S rRNA is highly divergent in *C. velia*) and therefore cannot support start codon prediction. ORFs of two additional genes, *rpoC1* and *rpoC2*, contain long insertions (507 and 971 aa) in their variable regions (none of them appears to represent an intein; supplementary fig. S4, Supplementary Material online). The most obvious example of this is Rps4, which is encoded within a 1,998 aa long ORF, compared with 201 aa in the red alga *Porphyra purpurea*.

When compared with PsA, the AtpB split is suggestive of a different, but equally interesting process. The region surrounding the split is conserved in plastid, mitochondrial, and bacterial homologs, but interrupted by a 24 amino acid (aa) insertion in *Vitrella*, *C. velia*’s closest photosynthetic relative (fig. 2C). A detailed sequence alignment of this region shows that the split in *C. velia* AtpB occurred within this shared insertion, suggesting that the ancestral acquisition of this insertion might have provided an evolutionary opportunity for the split. A 31 aa long insertion at the same site is
(supplementary figs. S2 and S3, Supplementary Material online). Although this approach cannot accurately pinpoint the position of start codons (many rps genes are expressed relatively little and intergenic regions are often transcribed), the consistent coverage by transcripts suggests that the ORFs may really have expanded in size at the protein level. Much of the long insertions in rpo genes is also equally represented in the transcriptome data; however, in this case a drop in coverage is found within each insertion creating a 200bp gap (supplementary fig. S4, Supplementary Material online). This raises the possibility that RpoC1 and RpoC2 may be split much like PsaA and AtpB; however, both are in frame and low coverage, so more evidence is required to distinguish between the two alternatives.

Unusual changes to gene presence have also been noted: two genes are present in two (atpH) and three (clpC) full-length paralogs, despite that a single type is found in other plastid genomes. All three clpC paralogs are highly divergent in sequence from each other and clpCs in other plastids. They overlap only partially and the largest and most complete of the paralogs (clpC3) is an order of magnitude less expressed than the other two (supplementary fig. SSA, Supplementary Material online). Likewise, the two atpH paralogs are different in sequence (93% nucleotide similarity), length, and expression. AtpH1 is both highly expressed and conserved, but the expression of the duplicated atpH2 located on the TIRs is nearly 100-fold lower (supplementary fig. SSB, Supplementary Material online). The atpH2 gene contains an unusual C-terminal extension, which doubles it in size. A closer look at the surrounding sequence reveals that atpH2 originated by a duplication event of the atpH1 region (including trnK; discussed later). This, together with its low expression, suggests that atpH2 is most likely a non-functional pseudogene. More generally, however, this example illustrates how duplication can drive gene paralogy, ORF expansion, and altered gene order; all features observed throughout the C. velia plastid genome.

Analysis of the total RNA and polyA fraction (fig. 4) transcription profiles revealed more unexpected features. Transcripts for all plastid genes were present in the total RNA fraction and consistently also found in the polyA fraction at two to three magnitudes lower abundance, probably due to a carryover of non-polyA transcripts (fig. 4). The psbA gene stood out in two ways, however. First, it was highly expressed, being represented at almost an order of magnitude higher levels than other protein-coding genes (and similar levels as the rRNA operon, fig. 4 and table 1). Second, the representation of psbA in the polyA fraction was about equal to the total RNA fraction, which would be the expected result for a polyadenylated transcript (fig. 4). Only two additional regions corresponding to rps19 and a psbB fragment (part of the annotated orf157) were relatively overrepresented in the polyA fraction, but only rps19 contained A-rich genomic sequence at the 3’-end that could explain the bias. Whether the psbA gene and psbB fragment are uniquely polyadenylated or a carryover into the polyA fraction occurred due to
A Linear-Mapping Plastid Genome in *C. velia*

One of the most significant features of the previously published plastid genome from *C. velia* is that it does not map as a circle as do all other plastid genomes completed to date. The gap in the sequence falls between the two copies of the TIR containing *psbA* (Janouškovec et al. 2010). We attempted to close this gap using multiple approaches. We started by massively increasing the depth of sequence coverage. The genome was originally sequenced by 454 pyrosequencing to a 13-fold depth of coverage, so we used an independent approach and assembled the genome from paired-end Illumina sequence reads to an average depth of almost 600-fold. This approach extended the two TIRs by 300 bp on each side, which was verified by PCR (supplementary fig. S6, Supplementary Material online). The very end of each TIR was followed by a 15 bp sequence further followed by an extension into the complementary TIR at a very low depth of coverage (fig. 5). These extensions created an overlap between the ends; however, several lines of evidence suggested that this is not due to a circular chromosome, but rather a linear topology with structured ends. First, mapping the depth of coverage revealed a pattern expected for a linear molecule. The depth of coverage was generally consistent across the large single copy (LSC) region and most of the TIRs, but reduced linearly beginning at about 600 bp from the end, followed by an exponential reduction and partial recovery between 300 to 50 bp from the end, followed by a terminal drop (fig. 5A). The size and shape of the exponential reduction/recovery would be predicted by the mean DNA fragment size in the sequencing library (327 bp) and linear topology, as a consequence of fragmentation bias near the chromosome end. Therefore, both the overall decrease in coverage depth and the exponential decrease/recovery close to the ends support a linear chromosome conformation (if it were circular, the whole region should not differ from the rest of the genome). Second, if the genome was circular-mapping and the gap was hard to sequence, we would expect some paired ends to span the gap. However, when more than 650,000 paired ends are mapped to the genome, not a single paired end spanning the gap was identified (see Materials and Methods). Instead, both ends of the chromosome consisted exclusively of reverse-mapping reads (fig. 5A and B). Third, PCR experiments using six primers (all shown to successfully amplify products from one side of the gap) and all their nested combinations failed to fill the gap between the two TIRs (supplementary fig. S6, Supplementary Material online). Fourth, the size of the plastid genome estimated directly by pulsed-field gel electrophoresis (PFGE) and by *psbA* probe hybridization was consistent with abundant presence of linear monomers, 120 kb in size (supplementary fig. S7, Supplementary Material online, and Janouškovec et al. 2010). Subgenomic-sized material and unresolved DNA in the well were also abundantly represented, whereas putative linear dimers were rare and linear concatamers were not detected. The PFGE experiment was reproduced by using probes to three plastid genes encoded in the LSC region, *tufA*, *petA*, and *atpA* (see Materials and Methods), and consistently led to the same result, suggesting that linear monomers are an important constituent of plastid DNA in *C. velia*. Subgenomic-sized molecules could represent plastid DNA fragmented during PFGE preparation (note that intact *C. velia* cells were digested in agarose plugs, however), nucleus-encoded plastid DNA...
fragments, plastid DNA in the process of replication, or other forms of genuine plastid DNA (Ellis and Day 1986; Oldenburg and Bendich 2004; Scharff and Koop 2006). Incompletely digested cells of *C. velia* in the agarose plugs (see Materials and Methods) could be responsible for much of the signal in the well, although high-molecular-weight branched DNA forms could also be present and cannot be distinguished using the current approach (Bendich 1991; Bendich 2004). Altogether, because the plastid genome was not closed by deep sequencing using either 454 or Illumina, no paired-end linkage could be established, no linkage could be established by PCR, and because the size of the plastid genome estimated by PFGE and Southern blot hybridizations was consistent with the existence of linear monomers, we conclude that the genome is principally linear in structure and circular molecules and linear concatemers are rare or absent.

**FIG. 5.** Schematic of plastid chromosome ends. (A) Coverage depth (y axis) by forward-mapping (yellow), reverse-mapping (purple), and total (red) genomic reads is projected onto the TIR. Position of genes is shown by gray boxes at the bottom (K = trnK, P = trnP). Total coverage depth starts dropping gradually at about 800 bp from the chromosome end, and goes through a U-shaped minimum at about 150 bp from the end. The calculated mean size of the end-sequenced DNA fragments (horizontal black bar, red bits correspond to sequenced end pairs) and coverage by forward/reverse read pairs in this region suggest that the U-shaped minimum results from unequal DNA fragmentation near the chromosome end. A steep drop in total coverage depth occurs at the very end of the chromosome (B). Here, the TIRs on each of the chromosome ends (green and blue, respectively) diverge into a shared 15 bp region followed by a short sequence of the complementary TIR (DNA ends marked as 3′ and 5′). The total depth of coverage by DNA reads (y axis) steeply decreases in the 15 bp region. The 15 bp region is also enriched in nucleotide discrepancies (sequence logos in bold). The last 130 bp at each chromosome end is exclusively assembled from reverse-oriented paired ends (horizontal colored arrows) and no paired-end spanning this region can be identified, suggesting that the genome cannot be genuinely circularized.
plastid genome in silico using the cumulative GC-skew analysis. This region (68–69 kb from one end of the linear contig, supplementary fig. S8A, Supplementary Material online) is characterized by a minimum in cumulative GC-skew and loosely conserved tandem repeats (positions 68404–69029). In contrast, both ends of the chromosome are at a cumulative GC-skew maximum, suggesting that they most likely correspond to replication termini. Another feature commonly associated with origins of replication is the major shift in gene orientation, which is located approximately 18 kb away from the predicted replication origin in C. velia (supplementary fig. S8B, Supplementary Material online). The genomic region surrounding the two features is relatively overrepresented in the Illumina data, whereas both chromosome ends are relatively underrepresented (this holds after correction for base composition; supplementary fig. S8C, Supplementary Material online), which is consistent with the existence of replicating molecules in the Illumina library. Altogether, these three characteristics are consistent with linear-mapping topology and suggest that replication starts at about two-thirds of the way from one end of the molecule and proceeds bidirectionally toward the ends. The copy number for two C. velia plastid genes (tufA and atpH) was determined using dot blot hybridization to total genomic DNA and a synthetic construct (see Materials and Methods). The inferred copy number of both plastid genes was 9 times higher compared with a single-copy nuclear gene, topoisomerase II, and confirmed that genes encoded in the LSC region and TIR are equally represented (atpH is present in three closely related paralogs, all of which hybridized to the construct) (supplementary fig. S9, Supplementary Material online).

Discussion

Significance of Split Proteins for Photosynthesis and ATP Generation

Both PsA and AtpB have never been observed to be fragmented like those of C. velia. Although the psaA gene is split in Chlamydomonas reinhardtii, RNA trans-splicing produces a full-length transcript, which is translated to a full-length protein (Merendino et al. 2006). The structure of the PSI core is highly conserved in all known oxygenic phototrophs (Nelson and Yocum 2006; Busch and Hippler 2011) and it is formed by a heterodimer of structurally similar PsA and PsB proteins, which bind a remarkably high number of cofactors: about 100 chlorophylls, 14 carotenoids, and 2 phytoquinones (Amunts et al. 2010). The algal PSI is expected to be structurally similar to the plant PSI including several Lhca antenna proteins bound in a half-circle around the PSI (fig. 6A).

PsA in C. velia is split between domains 4 and 5 (figs. 1 and 6A). This suggests that the split was not simply a reversal of the proposed ancient formation of psaA by the gene fusion of a psbB/psbC-like antenna component and a psbA/psbD-like core of the reaction centre (domains 1–6 and 7–11, respectively, in fig. 1; Schubert et al. 1998). The position of the split and the faster evolutionary rate of PsA-1 (peripheral fragment) can be reconciled with the structure of the PSI (fig. 6A and B). Indeed, the presence of a supercomplex of PSI and antenna proteins in C. velia is obvious from the 2D electrophoresis (fig. 3). Apart from the structure of PSI itself, the process of the PSI assembly seems to be highly conserved through evolution from cyanobacteria to higher plant chloroplasts including assistance of the same auxiliary factors such as Ycf3 and Ycf4 (Boudreau et al. 1997; Ruf et al. 1997).

Although many individual steps in the PSI biogenesis remain unknown, this process had to be remodeled in C. velia to assemble three proteins instead of two into the functional PSI core complex. Particularly intriguing is the binding of chlorophyll cofactors into the split PsA. Both in vivo and in vitro studies suggest that chlorophyll has to be inserted into core subunits of PSI co-translationally, probably as a prerequisite for correct protein folding (Kim et al. 1994; Eichacker et al. 1996). The putative interface between PsA-A and PsA-B is rich in chlorophyll molecules (fig. 6B) and, according to available crystal structures, these chlorophylls are coordinated by both parts of split PsA. It is not clear how C. velia inserts these chlorophylls into PSI and how stability and correct folding of the nascent PsA-1/2 is achieved. Perhaps, both PsA fragments are assembled together co-translationally, and this is synchronized with or even assisted by chlorophyll loading (which could also help to coordinate the orientation of the two fragments). In any case, the PSI core biogenesis in C. velia could involve additional synchronizing assembly step(s) and the recruitment of new, nuclear-encoded auxiliary factors. Interestingly, the second protein of the PSI antenna core, PsB, mirrors some of the PsA structure at functionally analogous positions: although not split, PsB also contains a variable loop between the fourth and fifth trans-membrane helices and a highly divergent N-terminus. Whether this plays a compensatory role directly related to the splitting of PsA is not clear, but it is an interesting possibility that would require direct biochemical testing.

Similar to the PsA protein, any split of the AtpB protein would seem improbable due to its critical function and conserved structure. Even though the proposed AtpB-1 part of the β-subunit appears to be relatively far from the catalytic site (fig. 6C), it is known that tilting of the β-subunits including the top β-sheet crown is critical for the catalytic cycle. The movement of the upper part of the β-subunit is subtle, but if it is restricted by inhibitor tentoxin, then the cyclic interconversion of nucleotide binding sites is blocked (Groth 2002). Tentoxin binds close to the place where the AtpB protein is split (Groth 2002). Correct reassembly of AtpB-1 and AtpB-2 proteins is therefore likely to be highly constrained to avoid any restrictions of AtpB-2 movement. On the other hand, it is also possible that the fragmented AtpB-2 is more flexible and perhaps less sensitive to some natural inhibitors. The ATP synthase was also noteworthy in that it migrated at the top of the native protein gel. This has never been observed in cyanobacteria (Herranen et al. 2004) or higher plants (Järvi et al. 2011), and indicates that ATP synthase might be integrated into a very stable megadalton supercomplex in the plastid of C. velia.
Highly Divergent Characteristics of Gene and Chromosomal Structure

A number of highly modified ORFs are found in the *C. velia* plastid genome and distributed without any obvious pattern. Long extensions or insertions in plastid proteins are uncommon in most plastids including those in other alveolates. Some have been identified in green algae (e.g., de Cambiaire et al. 2006), but never in the extent or number reported here. Many of the *C. velia* predicted proteins that contain extensions function in the small subunit of the ribosome. If the extensions are really present in the final functional proteins, this has implications for plastid ribosome structure. Alternatively, transcripts of the rps ORFs and other unusual ORFs could be modified so that only the homologous region is translated. This process may be related to transcript oligouridylation, but as yet we have no evidence to support this. OligoU tailing is another modification of general interest (Wang and Morse 2006) and its significance and distribution in *C. velia* plastid transcripts remain unknown. OligoU tails in some *C. velia* transcripts are found in mature mRNAs that are derived from polycistronic primary transcripts (fig. 1). This suggests a possible role in transcript processing, perhaps similar to that observed in dinoflagellate plastids (Nelson et al. 2007; Dang and Green 2009; Barbrook et al. 2012). If confirmed, the ubiquitous presence of polycistronic transcription would imply that oligoU tailing is widespread in the *C. velia* plastid and tightly intertwined with expression of many core plastid genes.

No linear-mapping plastid genome has been documented to date, although there is a longstanding debate about whether circular-mapping plastid genomes might represent physically linear molecules (Bendich 2004). At a sufficient depth of coverage, the two ends of the *C. velia* plastid genome can be (barely) overlapped in sequence, but multiple lines of evidence suggest the genome is actually linear. The PFGE data support existence of linear monomers and rare...
dimers (which could arise by TIR-mediated recombination between two monomers), but is incongruent with a significant presence of linear concatemers. This leaves no direct comparisons possible with other plastid DNAs, including those in related apicomplexans, which comprise various mixtures of circular molecules, linear concatemers, and high-molecular-weight material (Lilly et al. 2001; Williamson et al. 2002; Bendich 2004; Day and Madesis 2007). Presence of TIRs in the C. velia plastid genome is a common characteristic of linear genomes (e.g., in apicomplexan mitochondria; Kairo et al. 1994; Hikosaka et al. 2010). The physical structure of the chromosome ends remains unknown, but given the sequence one can speculate based on structures observed in other linear genomes. These include single-stranded loops (possibly followed by a nick) or single-stranded overhangs in the 15 bp regions (reviewed in Nosek et al. 2004), or diverse associations with end-specific proteins (e.g., Rekosh et al. 1977; Tomáška et al. 1997). Gradually decreasing depth of coverage near the chromosome end may also suggest that the whole terminal 600 bp region, not just its very end, is protected, possibly by a t-loop (Tomáška et al. 2002). No repeated sequence can be found close to the ends, however, so determining which of these structures, if any, most closely resembles the physical state of the C. velia plastid chromosome will require direct testing. Altogether these data provide evidence for the first linear-mapping plastid genome in C. velia, and suggests that a similar topology could exist in other plastid genomes, particularly those that do not presently map as a circle (Gabrielsen et al. 2011).

The change in the plastid chromosome topology has important implications on the process of DNA replication and overcoming the end-replication problem. In the apicomplexan and other plastids, several types of replication origins have been identified and linked to the D-loop, rolling circle, and recombination-mediated replication strategies. The most common type of replication origin is associated with bidirectional D-loop replication and located inside the duplicated rRNA inverted repeat (Kolodner and Tewari 1975; Williamson et al. 2002; Krishnan and Rao 2009). In contrast, the C. velia plastid genome lacks both the duplicated rRNA unit and associated replication origin, and is probably replicated bidirectionally from the predicted putative replication origin toward the chromosome termini, which is a unique situation among all plastid genomes. The estimated plastid gene copy number (9× that of a nuclear marker) is also unusually low compared with most photosynthetic plastids (~50–100 copies; Day and Madesis 2007), and is more similar to the gene copy number in the non-photosynthetic plastids of apicomplexans (~25 copies; Matsuzaki et al. 2001) or plastid minicircles during exponential growth in dinoflagellates (2–4 copies; Koumandou and Howe 2007).

Searching for Forces Driving Structural Complexity

The C. velia plastid genome is unique in a number of ways. Here, we have examined several unusual characteristics of its organization and expression, all of which raises the question, why is this genome home to so many oddities? One property that might explain several of these characteristics concurrently is a high level of intragenomic duplication. The C. velia plastid genome has been extensively reshuffled and in order to understand the underlying forces we searched for palindromes, short repeats, and long repeats. Palindromes and short repeats were comparatively rare; however, the searches identified at least 46 pairwise matches between longer genomic sequences (>50 bp) at 65–100% similarity (fig. 7). Most of the duplicates are divergent and low copy number (2–3), and apart from the TIRs they include many intergenic regions, duplicated atpH and clpC genes, and duplicated fragments of otherwise complete genes (psaA, psaB, psbB, rpl3, and tufA; fig. 7). The region containing most repeated sequence is found around the TIR boundary (fig. 8). Multiple parts of this region, including several gene fragments, three atpH paralogs, and six trnK paralogs comprising both complete and fragmentary variants are scattered in several places in the genome (fig. 8). Long repeats are rare in most plastids, suggesting that duplication is an important force in shaping the C. velia plastid genome. Duplication could explain occurrence of paralogs, small gene fragments, extensive gene re-shuffling, and re-structuring of operon units through movement of promoters and intergenic elements. Similar processes could account for the addition of extensions and insertions to genes. The splitting of psaA and atpB could likewise be seen as a process involving duplication or partial duplication. For example, the Psaa split could have proceeded through an intermediate, in which the

![Fig. 7. Large repeats in the Chromera velia plastid genome. The linear genome is drawn as an incomplete circle with genes in black, intergenics in light gray, and two TIRs (white lines) at chromosome ends (15 bp regions at the termini are not visible here). Links connect homologous regions of 50 bp or more identified by BLAST (dark gray) and additional matches identified by Pipmaker (medium gray). Homology between the TIRs is not shown for clarity. Names of genes most affected by duplication events are indicated.](http://mbe.oxfordjournals.org/)

Downloaded from http://mbe.oxfordjournals.org/ at The University of British Colombia Library on February 17, 2014
plastids have maintained a conserved genomic architecture known to be actively prevented, which explains why most repair and replication. Unwanted genome reorganization have been extensive in the plastid genome, although sequence shuffling has involved homologous TIRs and 15 bp repeats now located at the termini. Interestingly, although sequence shuffling has been active in the dinoflagellate plastids and possibly even led to a massive fragmentation of their genome to small mini-circles (Zhang et al. 1999; Zhang et al. 2001), so it might be tempting to speculate that the core of many of these forces could have significantly contributed to the genome remodeling. For example, all alveolate plastids (C. velia, apicomplexans, Vitrella, and dinoflagellates) have a reduced gene complement, somewhat modified gene order, and comparatively fast rate of protein evolution (Janouškovec et al. 2010). Oligouridylylation of plastid transcripts is also found in dinoflagellates and may significantly predate C. velia, although it seems to be missing in apicomplexans (Janouškovec et al. 2010; J.J. unpublished data). The AtpB insertion present in dinoflagellates and Vitrella at a homologous position to the split in C. velia could be viewed as an ancestral acquisition that may have individualized the two parts of the protein preparing ground for the split. Similarly, recombination has been very active in the dinoflagellate plastids and possibly even led to a massive fragmentation of their genome to small mini-circles (Zhang et al. 1999; Zhang et al. 2001), so it might be tempting to speculate that the core cause of many of these conditions traces back to an ancient ancestor. Distinguishing between ancient and independent gains is nevertheless difficult. The plastid genomes of apicomplexans and Vitrella are generally not that unusual in structure as C. velia and dinoflagellates, so it is more accurate to propose that all alveolate plastids were ancestrally somewhat divergent, but evolved in different directions.

The fact that in a single organism two conservative photosynthesis proteins are split is astonishing given the high efficiency of C. velia photosynthesis (Quigg et al. 2012). The mechanism behind a particular split can be relatively simple, but it is more difficult to see how each of the PsA and AtpB fragments became integrated into a functional multiprotein complex. For example, both PsA fragments had to coevolve with mechanisms allowing co-translational insertion of chlorophylls, and simultaneously acquire/remodel interactions with numerous cofactors that mediate their assembly into a fully functional PSI complex. Many other structures in the C. velia plastid appear similarly complicated. Both the ribosome and RNA polymerase complex comprise a number of highly modified components including proteins with hundreds of amino acid long extensions and insertions. Explaining the appearance of highly modified structures

canonical and split PsAAs coexisted. In this scenario, each of the two complexes could have been regulated independently and potentially even have different functions much alike the cyanobacterial monomeric and trimeric PSI (Majeed et al. 2012).

Duplicated sequences in C. velia plastid are low copy number, long, and divergent from each other. They are dissimilar from short dispersed repeats found in some algal plastids (e.g., in Chlamydomonas; Maul et al. 2002) and longer repeats found in angiosperm plastids (e.g. Guisinger et al. 2011), all of which are conserved in sequence and comparatively high copy number. It remains unclear how duplication is mediated in C. velia. Few duplicates are in tandem so the duplication process is not likely to be related to DNA replication and is more likely related to recombination. A functional DNA recombination machinery has been documented in several land plant and algal plastids (Palmer 1983; Boynton et al. 1988; Haberle et al. 2008), where it primarily acts in DNA repair and replication. Unwanted genome reorganization is known to be actively prevented, which explains why most plastids have maintained a conserved genomic architecture (reviewed in Maréchal and Brisson 2010). It is possible that mechanisms preventing unwanted recombination or, alternatively, those allowing for recombination-mediated repair were loosened or lost in the C. velia plastid, which led to propagation of duplicates and rearrangements. The apparent linearization of the chromosome could also stem from a recombination-mediated rearrangement involving homologous TIRs and 15 bp repeats now located at the termini. Interestingly, although sequence shuffling has been extensive in the C. velia plastid genome, it has not been completely random. Most genes have been reorganized in large clusters with a strongly pronounced strand polarity (supplementary fig. S8B, Supplementary Material online). More importantly, functionally related genes often cluster together and are most likely transcribed polycistronically (fig. 4), so even with high levels of rearrangement a functionally reorganized genomic architecture can result from selection.

Intragenomic duplication may be the cause of most peculiarities in the C. velia plastid genome, but not all of them. Features such as transcript oligoU tailing and noncanonical genetic code (Moore et al. 2008; Janouškovec et al. 2010) may have different causes including increased mutation rate, small population size, and a short replication cycle. Moreover, early evolutionary acquisition and long-term combined effect of some these forces could have significantly contributed to the genome remodeling. For example, all alveolate plastids (C. velia, apicomplexans, Vitrella, and dinoflagellates) have a reduced gene complement, somewhat modified gene order, and comparatively fast rate of protein evolution (Janouškovec et al. 2010). Oligouridylylation of plastid transcripts is also found in dinoflagellates and may significantly predate C. velia, although it seems to be missing in apicomplexans (Janouškovec et al. 2010; J.J. unpublished data). The AtpB insertion present in dinoflagellates and Vitrella at a homologous position to the split in C. velia could be viewed as an ancestral acquisition that may have individualized the two parts of the protein preparing ground for the split. Similarly, recombination has been very active in the dinoflagellate plastids and possibly even led to a massive fragmentation of their genome to small mini-circles (Zhang et al. 1999; Zhang et al. 2001), so it might be tempting to speculate that the core cause of many of these conditions traces back to an ancient ancestor. Distinguishing between ancient and independent gains is nevertheless difficult. The plastid genomes of apicomplexans and Vitrella are generally not that unusual in structure as C. velia and dinoflagellates, so it is more accurate to propose that all alveolate plastids were ancestrally somewhat divergent, but evolved in different directions.

The fact that in a single organism two conservative photosynthesis proteins are split is astonishing given the high efficiency of C. velia photosynthesis (Quigg et al. 2012). The mechanism behind a particular split can be relatively simple, but it is more difficult to see how each of the PsA and AtpB fragments became integrated into a functional multiprotein complex. For example, both PsA fragments had to coevolve with mechanisms allowing co-translational insertion of chlorophylls, and simultaneously acquire/remodel interactions with numerous cofactors that mediate their assembly into a fully functional PSI complex. Many other structures in the C. velia plastid appear similarly complicated. Both the ribosome and RNA polymerase complex comprise a number of highly modified components including proteins with hundreds of amino acid long extensions and insertions. Explaining the appearance of highly modified structures
is more difficult than their components individually, because many parts of these systems are mutually constrained. Likewise, molecular processes such as oligoU tailing are difficult to justify in adaptive terms. Considering the complexity of similar unnecessary processes in other organelles, such as RNA editing or intron splicing (Gray et al. 2010), it is attractive to hypothesize that complicated molecular machineries have evolved in the C. velia plastid that serve no general advantage.

**Materials and Methods**

**DNA Extraction, Sequencing, Annotation, and Fragmented Gene Analysis**

Pelleted cells of *C. velia* were ground in liquid nitrogen using mortar and pestle and the resulting slurry was incubated in CTAB buffer (2% w/v cetyltrimethylammonium bromide; 1.42 M NaCl; 20 mM EDTA; 100 mM Tris HCl, pH 8.0; 2% w/v polyvinylpyrrolidone 0.5% β-mercaptoethanol; 1 mg/ml RNase A) at 65 °C for 20 min. After two extractions with phenol/chloroform and one with chloroform only, DNA was precipitated with isopropanol and washed with ethanol. Pellet was dried at room temperature and resuspended in TE buffer. Extracted DNA was separated by CsCl-Hoechst gradient ultracentrifugation and AT-enriched fractions were tested for the presence of plastid DNA using the *psba* probe. Plastid-enriched DNA fractions were sequenced using Illumina 54 bp paired-end reads technology (a total of 9,591,179 read pairs were obtained) and deposited in GenBank Sequence Read Archive (Bioproject PRJNA193178). De novo sequence assembly using MIRA3 extended the previous plastid contig (NC_014340.1) at both ends to the final length of 120,426 nucleotides at 572× average depth of coverage. Two errors at homopolymeric regions were corrected, which led to the merging of *rps7* with *orf142* and *rpl4* with *orf128*. The gene for 5S ribosomal RNA was identified upstream of 23S rRNA, and its fold verified using Mfold (http://mfold.rna.albany.edu/?q=mfold, last accessed May 25, 2013). Proteins were re-annotated by comparison with plastid homologs from an NTG start codon with the exception of 10 proteins that were more consistent with an AT[A,C,T] start codon. Six of the proteins had no alternative NTG start (*atpA*, *rps31*, *rps7*, *rps12*, *rps14*, *ycf3*). Unknown ORFs and genes with long N-terminal extensions were annotated from the first NTG start codon. The extended plastid genome sequence and new annotations were deposited in an updated *C. velia* plastid genome entry (NC_014340.2). PsA trans-membrane domains were predicted using TMHMM 2.0 (http://www.cbs.dtu.dk/services/TMHMM/, last accessed May 25, 2013) and by comparison to PsA of *Synechococcus elongatus*. The corrected genetic distances for PsA-1 and PsA-2 were calculated in Tree-Puzzle 5.2 (Schmidt et al. 2002) using a data set of 40 representative plastid PsA homologs and the following parameters: Likelihood mapping, Slow (exact) parameter estimates, WAG model with four gamma categories, estimated alpha parameter, and estimated aa frequencies. PsA-1/2 phylogenies were computed using a broader data set in RAxML using -m PROTGAMMALGF -f a -# 100. PsA-1/2 and AtpB data sets were aligned using -localpair option in MAFFT.

**RNA Extraction, RACE, and Northern Blot Hybridization**

Total RNA was isolated from ~0.2 g of *C. velia* cells ruptured by repeated freezing and thawing, followed by grinding with a pestle in liquid nitrogen, by the addition of 5 ml TRI reagent, followed by manufacturer’s instructions (Sigma). Total RNA was ligated with 5′ adapter (5′ RACE) or self-ligated with T4 RNA ligase (circular RACE), and RT-PCR was carried out with either random or specific primers, or with polyT primers (3′ RACE). Sequences of circularized transcripts were deposited under GenBank Nucleotide accessions KC734564–KC734569. For Northern blot analysis, ~10 μg per lane of total RNA was separated in 1% formaldehyde agarose gel in 1× MOPS buffer, blotted, and UV cross-linked as described elsewhere (Vondrušková et al. 2005). PCR-amplified DNA fragments of selected genes (for primers see supplementary table S4, Supplementary Material online) were labeled by random priming with [α-32P]-dATP and used as probes for hybridization at 55 °C. After washing the results were visualized on Typhoon Imaging System (GE Healthcare).

**Preparations of Cell Membranes and Two Dimensional Electrophoresis**

Cells of *C. velia* (optical density at 750 nm = ~0.5) were washed and resuspended in 1 ml of the working buffer containing 25 mM MES/NaOH, pH 6.5, 5 mM CaCl2, 10 mM MgCl2, and 20% glycerol. The resuspended cells were mixed with 0.5 ml of glass beads (0.1 mm diameter) in 2 ml eppendorf tube and broken using Mini-BeadBuster (BioSpec; eight shaking cycles, 10 s each with 2-min breaks for cooling the suspension on ice). Membranes were separated from the cell extract by centrifugation (40,000 × g, 20 min). The isolated membranes were resuspended in the working buffer and solubilized by gentle shaking with 1% dodecyl-β-maltoside at 10 °C for 1 h. Insoluble parts were removed by centrifugation (65,000 × g, 20 min). Analysis of native membrane complexes was performed using Clear-Native electrophoresis as described in Wittig and Schägger (2008). Individual proteins in membrane complexes were resolved in the second dimension by SDS-PAGE in a 12–20% linear gradient polyacrylamide gel containing 7 M urea (Sobotka et al. 2008).

**Protein Identification by LC-MS/MS Analysis**

Gel slices were placed in 200 μl of 40% acetonitrile, 200 mM ammonium bicarbonate, and incubated at 37 °C for 30 min. The solution was then discarded, the procedure was repeated one time, and the gel was finally dried in a vacuum centrifuge. Twenty microliters of 40 mM ammonium bicarbonate containing 0.4 μg trypsin (proteomics grade, Sigma) was added to the tube, incubated at 4 °C for 45 min, and then dried in a vacuum centrifuge. To digest proteins, 20 μl of 9%
acetonitrile in 40 mM ammonium bicarbonate was added to the gel and incubated at 37°C overnight. Peptides were purified using ZipTip C18 pipette tips (Millipore Corporation) according to manufacturer’s protocol. MS analysis was performed on NanoAcquity UPLC (Waters) on-line coupled to the ESI Q-Tof Premier mass spectrometer (Waters). One microliter of the sample was diluted in 3% acetonitril/0.1% formic acid and tryptic peptides were separated using the Symmetry C18 Trapping column (180 μm i.d. × 20 mm length, particle size 5 μm, reverse phase, Waters) with a flow rate of 15 μl/min for 1 min. It was followed by a reverse-phase UHPLC using the BEH300 C18 analytical column (75 μm i.d. × 150 mm length, particle size 1.7 μm, reverse phase, Waters). Linear gradient elution was from 97% solvent A (0.1% formic acid) to 40% solvent B (0.1% formic acid in acetonitrile) at a flow rate of 0.4 μl/min. Eluted peptides flowed directly into the ESI source. Raw data were acquired in data-independent MS^n examination mode (Waters). Precursor ion spectra were acquired with collision energy 5 V and fragment ion spectra with a collision energy 20–35 V ramp in alternating 1 s scans. For the second analysis, data-dependent analysis mode was used; peptide spectra were acquired with collision energy 5 V, and peptides with charge states of +2, +3, and +4 were selected for MS/MS analysis. Fragment spectra were collected with a collision energy 20–40 V ramp. In both modes, acquired spectra were submitted for database searching using the PLGS2.3 software (Waters) against the predicted proteins coded by the plastid genome of C. velia and against the available EST sequences (www.ncbi.nlm.nih.gov, last accessed May 25, 2013). Acetyl N-terminal, deamidation N and Q, carbamidomethyl C, and oxidation M were set as variable modifications. Identification of three consecutive y- or b-ions was required for a positive peptide match and a minimum of three peptide matches were required for a positive protein identification.

Transcriptome Analysis
Total RNA and polyA RNA fraction were sequenced using Illumina paired-end read technology. Read coverage depth was averaged across the two TIRs. Coverage values for all sites were then exported and plotted in a spread sheet editor. Incorrect read mapping to duplicated genes (atpH and clpC) and repeated regions was ruled out based on match length or sequence divergence (all matches were 96.6% similar or lower). Presence of RNA editing was assessed by comparing relative representation of high-quality nucleotides (Phred score 30 or higher) at each site in between transcriptomic and genomic reads mapped on the plastid contig.

DNA Read Mapping, Chromosome Ends Analysis, and Repeat Analysis
Illumina paired-end DNA reads were separately mapped on the linear and artificially circularized plastid genome sequence using Consed 22 and Bowtie 2.0 under default settings (667,715 read pairs mapped). Read coverage depth was averaged across the two TIRs and incorrect read mapping to repeats was ruled out as above. Coverage depth by forward-oriented (i.e., oriented outside of the chromosome) and reverse-oriented reads was calculated using Bedtools 2.17. No forward-oriented reads were found within the last 130 bp of the chromosome ends (all reads in this region were reverse-oriented). Only 23 forward-oriented reads were found within the last 250 bp of the chromosome ends, all of which had a reverse-oriented pair mapping close to the chromosome end except 6 reads whose pairs did not map to the plastid genome sequence or each other. In order to estimate the mean DNA fragment size, full-length read pairs (54 bp each) were mapped on the plastid contig using Bowtie 2.0, and sorted and analyzed using Picard Tools SortSam.jar and CollectInsertSizeMetrics.jar utilities. Discrepant bases in figure S8 were visualized using WebLogo 3 (http://weblogo.threeplusone.com/create.cgi, last accessed May 25, 2013) using “Base probabilities” as units and “No adjustment for base composition” setting. PCR on genomic DNA was done at the following conditions: annealing at 55°C, 30 s to 4 min elongation times, and 35 cycles. All six primers used in bridging the gap between the two contig ends were confirmed to give functional products with a different primer pair under the same conditions (supplementary fig. S6C and D, Supplementary Material online). Palindromes and repeats were searched using EMBOSS 6.3.1 repeat identification tools at Pasteur Mobyle website (http://mobyle.pasteur.fr/, last accessed May 25, 2013), Pipmaker (http://pipmaker.bx.psu.edu/pipmaker/, last accessed May 25, 2013), and pairwise BlastN searches (length >50 nucleotides, e-value <0.01, and sequence similarity 65–100%), all at default settings. Figure 7 was plotted using Circos.

PFGE and Southern Blot Hybridization
Chromera velia cells (10^9–10^10) were slowly pelleted, embedded in low-melting agarose plugs, and treated with 2% N-laurylsarcosine and 2 mg/ml proteinase K for 28 h at 56°C. Thick cell walls in C. velia prevented efficient penetration of cell membranes and DNA release leaving between 10% and 30% of all cells undigested. The plugs were inserted into 1% agarose gel and DNA was resolved on CHEF-DR III PFGE (Bio-Rad) in 0.5× TBE at 14°C and using two different settings: 1) U = 6 V/cm with 0.5–25 s pulses and 120° angle for 20 h and 2) U = 6 V/cm with 0.1–2 s pulses and 120° angle for 14 h. After treatment with 0.25 M HCl for 20 min, the gels were denatured, neutralized, and blotted to nylon membrane and UV cross-linked following standard protocols. DNA probes were labeled as described above. Southern blot analysis with psbA, tufA, petA, and atpA probes (see supplementary table S4, Supplementary Material online, for primers) was performed in NaPi solution (0.5 M Na_2HPO_4, pH 7.2, 1 mM EDTA, 7% SDS, 1% BSA) at 65°C overnight, and the membranes were washed 20 min in 2× SSC, 0.1% SDS at room temperature and another 20 min in 0.2× SSC, 0.1% SDS at 65°C, and visualized on Typhoon Imaging System (GE Healthcare).
Determining the Replication Origin and Plastid Gene Copy Number

The cumulative GC skew plot was drawn using the utility at http://gcat.davidson.edu/DGPB/gc_skew/gc_skew.html (last accessed May 25, 2013). Coverage depth by Illumina genomic reads was determined as above. Relative abundance of selected plastid and nuclear genes in *C. velia* were estimated as follows. A 519-bp-long fragment of nuclear topoisomerase II gene (a typical single-copy nuclear gene), and 337-bp-long and 327-bp-long fragments of chloroplast *tufA* and *atpH*2 genes, respectively, were cloned in tandem into a single plasmid, using unique restriction sites in the primers (supplementary table S4, Supplementary Material online). The resulting construct was labeled pS5+13. Separate serial dilutions of total DNA of *C. velia* digested with DraI and Sph1103I and the EcoRV-linearized pS5+13 were spotted on a Biotine membrane (Pall), cross-linked, and blotted as described above. Each gene fragment (Topo II, *tufA*, *atpH*) was labeled by [α-32P]ATP, hybridized at 65°C to one of three identical blots, and visualized as above. The *atpH*2 probe was tested to hybridize efficiently to all three *atpH*2 paralogs (two of which are identical to the probe and the third is 93% identical). For this, genomic DNA digested with DraI and Sph1103I was resolved in a 0.75% agarose gel, blotted, and hybridized as above, and signals from *atpH*1 and *atpH*2-specific fragments were compared.

Supplementary Material

Supplementary figures S1–S9 and tables S1–S4 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by a grant from the Canadian Institutes of Health Research to P.J.K. (MOP-42517); by the Czech Science Foundation projects P506/12/1522 and P501/12/G055 to M.O.; by the Praemium Academiae award to J.L.; by Award IC/2010/09 by the King Abdullah University of Science and Technology (KAUST) to A.P., M.O., and J.L.; and by the project Algatche (CZ.1.05/2.10/03.0110) to R.S., J.K., and O.P. P.J.K. and J.L. are Fellows of the Canadian Institute for Advanced Research. P.J.K. was supported by a Fellowship from the John Simon Guggenheim Foundation.

References


Plastid Genome of *Chromera velia* · doi:10.1093/molbev/mst144

MBE