The Exceptionally Large Chloroplast Genome of the Green Alga Floydiella terrestris Illuminates the Evolutionary History of the Chlorophyceae

Jean-Simon Brouard, Christian Otis, Claude Lemieux, and Monique Turmel*

Département de biochimie et de microbiologie, Université Laval, Québec, QC, Canada

*Corresponding author: E-mail: monique.turmel@bcm.ulaval.ca.

Accepted: 29 March 2010

The fully annotated sequence of the Floydiella chloroplast genome has been deposited in GenBank under the accession number GU196268.

Abstract

The Chlorophyceae, an advanced class of chlorophyte green algae, comprises five lineages that form two major clades (Chlamydomonadales + Sphaeropleales and Oedogoniales + Chaetopeltidales + Chaetophorales). The four complete chloroplast DNA (cpDNA) sequences currently available for chlorophyceans uncovered an extraordinarily fluid genome architecture as well as many structural features distinguishing this group from other green algae. We report here the 521,168-bp cpDNA sequence from a member of the Chaetopeltidales (Floydiella terrestris), the sole chlorophycean lineage not previously sampled for chloroplast genome analysis. This genome, which contains 97 conserved genes and 26 introns (19 group I and 7 group II introns), is the largest chloroplast genome ever sequenced. Intergenic regions account for 77.8% of the genome size and are populated by short repeats. Numerous genomic features are shared with the cpDNA of the chaetophoralean Stigeoclonium helveticum, notably the absence of a large inverted repeat and the presence of unique gene clusters and trans-spliced group II introns. Although only one of the Floydiella group I introns encodes a homing endonuclease gene, our finding of five free-standing reading frames having similarity with such genes suggests that chloroplast group I introns endowed with mobility were once more abundant in the Floydiella lineage. Parsimony analysis of structural genomic features and phylogenetic analysis of chloroplast sequence data unambiguously resolved the Oedogoniales as sister to the Chaetopeltidales and Chaetophorales. An evolutionary scenario of the molecular events that shaped the chloroplast genome in the Chlorophyceae is presented.

Key words: Oedogoniales, Chaetophorales, Chaetopeltidales, plastid genome evolution, phylogenomics, repeated sequences.

Introduction

The monophyletic class Chlorophyceae (sensu Mattox and Stewart) is part of the Chlorophyta, a major division of green algae that also includes the Prasinophyceae, the Ulvophyceae, and the Trebouxiophyceae (Lewis and McCourt 2004). In the Chlorophyta, the deep branching position of the Prasinophyceae is undisputed (Steinkötter et al. 1994; Nakayama et al. 1998; Fawley et al. 2000; Guillou et al. 2004), and although the branching order of the three other classes remains uncertain, increasing evidence suggests that the Trebouxiophyceae are sister to a clade uniting the Chlorophyceae and Ulvophyceae (Pombert et al. 2004, 2005). The members of the Chlorophyceae display diverse cell organizations (unicells, coccoids, colonies, simple flattened thalli, unbranched, and branched filaments) and among the chlorophytes exhibit the greatest variability at the level of the flagellar apparatus (Lewis and McCourt 2004). The flagellar basal bodies of most chlorophyceans are displaced in a clockwise (CW, 1–7 o’clock) direction or are directly opposed (DO, 12–6 o’clock), thus contrasting with the counterclockwise arrangement observed in the Ulvophyceae and Trebouxiophyceae (O’Kelly and Floyd 1984). Not only is the configuration of the flagellar apparatus the major feature unifying chlorophycean green algae but also is congruent with the subdivision of the Chlorophyceae into five orders (Chlamydomonadales, Sphaeropleales,
Evolution of the Chlorophycean Chloroplast Genome

Chaetophorales, Chaetopeltidiales, and Oedogoniales. In the Chlamydomonadales (also designated as CW clade), biflagellates display a CW orientation of basal bodies, whereas quadriflagellates harbor various flagellar apparatus ultrastructures (Nozaki et al. 2003). The Sphaeropleales order (DO clade) comprises vegetatively nonmotile unicellular or colonial taxa that produce zoospores with two flagella arranged in a DO configuration (Lewis and McCourt 2004). Quadriflagellates with the perfect DO configuration of flagellar bodies characterize the Chaetopeltidiales (O’Kelly et al. 1994), whereas quadriflagellates from the Chaetophorales display a polymorphic arrangement (DO + CW) where one pair of basal bodies has the DO configuration and the other is slightly displaced in a CW orientation (Manton 1964; Melkonian 1975; Floyd et al. 1980; Bakker and Lokhorst 1984; Watanabe and Floyd 1989). The members of the Oedogoniales are the only chlorophycean green algae that do not possess basal bodies in a DO or in a CW configuration: their unusual flagellar apparatus is characterized by a stephanokont arrangement of flagella (i.e., an anterior ring of flagella) (Pickett-Heaps 1975).

Phylogenies inferred from the nuclear-encoded 18S rRNA gene have been unable to unravel the interrelationships of the major chlorophycean lineages (Boothen et al. 1998; Buchheim et al. 2001; Nozaki et al. 2003; Shoup and Lewis 2003; Müller et al. 2004; Alverghina et al. 2006). Because phylogenomic studies based on comparative analysis of chloroplast genomes have been successful in resolving separate issues concerning relationships of algae or land plants (Martin et al. 1998; Qiu et al. 2006; Jansen et al. 2007; Lemieux et al. 2007; Rogers et al. 2007; Turmel et al. 2008; Turmel, Gagnon, et al. 2009), we have adopted this strategy to decipher the branching order of the chlorophycean lineages. Also, an impetus for sequencing chloroplast genomes from representatives of the five major chlorophycean lineages was our desire to gain insights into the molecular events that shaped the extremely plastic and ancestral lineages was our desire to gain insights into the molecular events that shaped the extremely plastic and ancestral lineages. However, the branching order of the lineages within the OCC clade could not be identified in an unambiguous manner: the protein and gene trees inferred from the data set of 44 proteins/gens from 20 green plants differed in topologies, whereas the trees inferred from the data set of 57 genes/proteins from the six chlorophyceans showed with strong support that the Oedogoniales diverged before the Chaetopeltidiales and the Chaetophorales. The latter topology was supported by the presence of uniquely shared trans-spliced introns in the Stigeoclonium and Floydiella rbcl genes.

Recently, phylogenetic analyses of multiple proteins/gens (44 or 57 depending on taxon sampling) derived from the abovementioned chlorophycean chloroplast genomes plus those from the partly sequenced C. moewusii and Floydiella chloroplast genomes provided strong support for the split of the Chlorophyceae into two major clades: the Chlamydomonadales + Sphaeropleales clade (CS clade) and the Oedogoniales + Chaetophorales + Chaetopeltidiales (OCC clade) (Turmel et al. 2008). Molecular signatures, namely trans-spliced group II introns in the psaC and petD genes and insertions/deletions in separate genes, were congruent with this dichotomy. However, the branching order of the lineages within the OCC clade could not be identified in an unambiguous manner: the protein and gene trees inferred from the data set of 44 proteins/gens from 20 green plants differed in topologies, whereas the trees inferred from the data set of 57 genes/proteins from the six chlorophyceans showed with strong support that the Oedogoniales diverged before the Chaetopeltidiales and the Chaetophorales. The latter topology was supported by the presence of uniquely shared trans-spliced introns in the Stigeoclonium and Floydiella rbcl genes.

In this study, we describe the complete chloroplast genome sequence of Floydiella and present unambiguous
evidence that the Oedogoniales diverged before the Chaeto-
opeltidales and the Chaetophorales. Exceeding 500 kb, the newly
analyzed genome is the largest chloroplast genome ever completely sequenced. Mapping of structural cpDNA
features on the inferred chlorophycean phylogeny enabled us not only to identify additional structural features supporting
the Chaetophorales + Chaetopectidales clade but also to better understand the evolutionary pathway followed by the chloroplast genome within the OCC clade.

Materials and Methods

Strains and Culture Conditions

_Floydiella terrestris_ was obtained from the Culture Collection of Algae at the University of Texas at Austin (UTEX
1709) and grown in C medium (Andersen et al. 2005) under
12 h light/dark cycles.

Cloning and Sequencing of the _Floydiella_ Chloroplast Genome

Most of the _Floydiella_ sequence was derived from plasmid
clones; the construction of the random plasmid clone library
has been previously described (Turmel et al. 2008). Briefly,
an A + T rich fraction containing cpDNA was isolated by
CsCl-bisbenzimide isopycnic centrifugation of total cellular
dNA (Turmel et al. 1999). This DNA fraction was sheared by
nebulization to produce 1,500 to 2,000-bp fragments that were
cloned into pSMART-HCKan (Lucigen Corporation). DNA
templates were prepared from selected clones with
the Qiagen Miniprep kit (Qiagen Inc) and sequenced
by the QIAprep 96 Miniprep kit (Qiagen Inc) and sequenced
of the National Center for Biotechnology and Information
(NCBI) server. Protein-coding genes and ORFs were localized
using tRNAscan-SE 1.23 (Lowe and Eddy 1997). Boundaries of introns were located by modeling intron
features on the inferred chlorophycean phylogeny using FRAMEALIGN
1997). Positions of transfer RNA (tRNA) genes
were determined using _tRNAscan-SE_ (Lowe and Eddy
1997). Genes and ORFs were identified by Blast similarity searches
using the TOPO TA cloning kit (Invitrogen) before sequencing.

Analyses of Coding Sequences and Gene Order

Genes and ORFs were identified by Blast similarity searches
(Altschul et al. 1990) against the nonredundant database of the National Center for Biotechnology Information
(NCBI) server. Protein-coding genes and ORFs were localized
precisely using ORFFINDER at NCBI, various programs of
the Genetics Computer Group package. Conserved gene pairs or gene clusters exhibiting identical
gene polarities in selected green algal cpDNAs were identified using a custom-built program.

Analyses of Repeated Sequences

To estimate the proportion of repeated sequences in the
_Floydiella_ chloroplast genome, repeated sequences were re-
trieved using REPIND of the REPuter 2.74 program (Kurtz
et al. 2001) with the options -f (forward) -p (palindromic) -l
(least repeat length = 30 bp) -allmax and then masked on
the genome sequence using REPEATMASKER (http://www.
repeatmasker.org/) running under the WU-Blast 2.0 search
engine (http://blast.wustl.edu/).

Repeated sequences were classified and counted using various programs of the VMATCH large-scale sequence
analysis software (http://www.vmatch.de/). After construct-
ing an index of repeated sequences using MKVTREE with the
-dna -pl -allout and -v options, direct repeats > 30 bp were
identified using VMATCH (-d and -l options) and then as-
signed to distinct families with MATCHCLUSTER by allowing
10% sequence dissimilarity (-erate option set to 10). For each family, sequences were retrieved with VMATCHSELECT
and a consensus was generated from a MUSCLE 3.7 (Edgar
2004) alignment. Repeat families were then sorted accord-
ing to their score values; the score of each family was ob-
tained by multiplying the size of the prototype sequence by the copy number determined using FUZZNUC in EMBOSS.
Like REPuter, VMATCH identifies all overlapping repeated sequences and thus overestimates the total number of re-
peated elements in a genome. To prevent the detection of overlapping repeats, prototypes of the various families
were submitted to a second round of counting using a custom-built program that finds and masks repeats sequen-
tially on the genome sequence, starting with the prototype having the highest score value.

Phylogenetic Analyses of Sequence Data

An amino acid data set and the corresponding nucleotide
data set with first and second codon positions were derived
from the completely sequenced chloroplast genomes of 12
chlorophytes. Species names and accession numbers are as
follows: _Chlamydomonas reinhardtii_, NC_005353 (Maul
et al. 2002); _Parachlorella kessleri_, NC_012978 (Turmel,
Otis, and Lemieux 2009); _Chlorella vulgaris_, NC_001865
(Wakasugi et al. 1997); _F. terrestris_, GU196268 (this study);
_Lepota teresita_, NC_009681 (de Cambiaire et al. 2007);
_O. cardium_, NC_011031 (Brouard et al. 2008); _Oltman-
ssiopsis viridis_, NC_008099 (Pombert et al. 2006); _Oocystis
solitaria_, FJ968739 (Turmel, Otis, and Lemieux 2009); _Ped-
nomonas minor_, FJ968740 (Turmel, Otis, and Lemieux
2009); _Pseudendoclonium akinetum_, NC_008114 (Pombert
et al. 2005); _S. obliquus_, NC_008101 (de Cambiaire et al.
2006), and *S. helvetica*, NC_008372 (Bélanger et al. 2006). In addition, protein-coding genes from the partially sequenced *C. moewusii* chloroplast genome were incorporated in the data sets; the accession numbers of these gene sequences are reported in Turmel et al. (2008).

To limit the proportion of missing data, we selected for analysis the protein-coding genes that are shared by at least eight taxa. Sixty-nine genes met this criterion: *atpA*, *B*, *E*, *F*, *H*, *I*, *ccsA*, *cemA*, *chib*, *L*, *N*, *clpP*, *ftsH*, *infA*, *petA*, *B*, *D*, *G*, *L*, *psaA*, *B*, *C*, *J*, *M*, *psbA*, *B*, *C*, *D*, *E*, *F*, *H*, *I*, *J*, *K*, *L*, *M*, *N*, *T*, *Z*, *rcbL*, *rpl2*, *5, 12, 14, 16, 20, 23, 32, 36, rpoA, *B*, *C1*, *C2*, *rps2*, *3, 4, 7, 8, 9, 11, 12, 14, 18, 19, tufA, *ycf1*, *3, 4, 12*. The amino and nucleotide data sets were prepared as described by Turmel, Gagnon, et al. (2009), except that ambiguously aligned regions were removed using the -b2 option (minimal number of sequences for a flank position) of GBLOCKS set to 7. All phylogenetic inferences were carried out using the maximum likelihood (ML) method as implemented in Treefinder (version of October 2008) (Jobb et al. 2004). Treefinder was also used to identify the best models fitting the data under the Akaike information criterion. The amino acid data set was analyzed using the LG + F + I (gamma distribution of rates across sites with eight categories) model of sequence evolution. Trees were inferred from the nucleotide data set using the general time reversible + I (eight categories) model. Confidence of branch points was estimated by 100 bootstrap replications.

### Reconstruction of Ancestral Character States

Using MacClade 4.08 (Maddison D and Maddison W 2000), we prepared a data set of genomic characters for the *Chlamydomonas, Chlorella, Floydiella, Oedogonium, Oltmannsiellopsis, Pseudendocladium, Scenedesmus*, and *Stigeoclonium* chloroplast genomes by coding the presence/absence of an IR, genes, ancestral gene pairs, derived/inherited genes, fragmented genes, duplicated genes, *trans*-spliced group II introns, and inteins. Most genomic features were coded as Dollo characters; the presence/absence of *trans*-spliced group II introns were coded as irreversible characters, and the features related to the *rps4* and *rpoB* gene structures were treated as ordered 3-states characters. In the case of *rps4*, state 0 represents the ancestral structure of the gene, state 1 the expanded form, and state 2 the structure lacking both the last 40 codons of the gene and the preceding insertion of more than 2,500 codons. In the case of *rpoB*, state 0 denotes the expanded form of the gene, state 1 the gene fragment into two adjacent ORFs (*rpoBα* and *rpoBβ*), and state 2 the form of the gene consisting of these two unlinked ORFs. MacClade was used to map the gains and losses of all characters on tree topologies and to calculate GC skew analysis did not disclose any putative origin and terminus of replication that are consistent with a bidirectional mode of replication. Aside from the presence of conserved genes, we identified 89 ORFs greater than 100 codons in the *Floydiella* chloroplast genome. Blast searches

### Data Deposition

The fully annotated sequence of the *Floydiella* chloroplast genome has been deposited in GenBank under the accession number GU196268.

### Results and Discussion

#### The Exceptionally Large Size of the *Floydiella* Chloroplast Genome Is Largely Explained by the Expansion of Intergenic Spacers

At 521,168 bp, the circular-mapping chloroplast genome of *Floydiella* (fig. 1) is the largest cpDNA ever sequenced, being more than 2.3-fold larger than its counterparts in the OCC and CS clades (table 1) but exceeding to a lesser extent the minimal size estimated (420,650 bp) for the partially decoded cpDNA of *Volvox*, a member of the Chlamydomonadales (Smith and Lee 2009). The *Volvox* genome sequence, which consists of 34 contigs, could not be deciphered in its entirety because the high abundance of repeats aborted the sequencing reactions and hampered sequence assembly. With an A + T content of 65.5%, the *Floydiella* genome falls within the range of base composition observed for the four completely sequenced chlorophycean cpDNAs (table 1) but deviates significantly from *Volvox* cpDNA (57% A + T). Multiple mutational events can promote chloroplast genome expansion, including growth of the IR (Turmel et al. 1999; Chumley et al. 2006; Brouard et al. 2008), duplication of genes (Lee et al. 2007; Cai et al. 2008; Haberle et al. 2008), proliferation of introns and repeated elements (Maul et al. 2002; Pombert et al. 2005; Bélanger et al. 2006; Chumley et al. 2006; Cai et al. 2008; Haberle et al. 2008; Smith and Lee 2009), acquisition of foreign sequences through lateral DNA transfer (Brouard et al. 2008; Turmel, Gagnon, et al. 2009), and accumulation of noncoding and coding sequences through strand slippage during DNA replication (Sears et al. 1995; Turmel et al. 2005). Like its *Stigeoclonium* homolog, the *Floydiella* genome lacks an IR (table 1). In the case of *Volvox* cpDNA, it is currently unknown whether an IR is present.

The *Floydiella* genome encodes 97 conserved genes, that is, the same number identified in the *Stigeoclonium* chloroplast (table 1); however, its gene content differs by the presence of *infA* and the absence of *trnS*(gca) (table 2). Relative to the *Oedogonium* genome, it lacks only the *trnR*(ucg) and *trnR*(ccu) genes (table 2). Contiguous genes in the *Floydiella* genome show a pronounced propensity to be clustered on the same strand; however, in contrast to the *Stigeoclonium* genome (Bélanger et al. 2006), genes are unequally distributed between the two DNA strands (76:21) and a cumulative GC skew analysis did not disclose any putative origin and terminus of replication that are consistent with a bidirectional mode of replication. Aside from the presence of conserved genes, we identified 89 ORFs greater than 100 codons in the *Floydiella* chloroplast genome. Blast searches
indicated that the vast majority of these ORFs have no significant similarities with known genes. The free-standing orf120, orf150, and orf183 are related to HNH homing endonucleases, whereas the orf102, which is located 2,560-bp downstream of clpP, represents the duplicated 3’ coding region of this gene (97% identity at the protein level).

As observed for other chlorophyte cpDNAs (Pombert et al. 2005, 2006; Bélanger et al. 2006; de Cambiaire et al. 2006, 2007), numerous genes in Floydiella cpDNA (cemA, clpP, ftsH, rpoB, rpoC1, rpoC2, rps3, rps4, and ycf1) have enlarged coding regions relative to their counterparts in the streptophyte green alga Mesostigma viride. Alignments of the deduced amino acid sequences of these Floydiella genes with their homologs in the Chlorophyta and the Streptophyta revealed that their expansion is caused by sequence insertions at one or more sites within internal regions. These insertions are generally part of variable regions showing heterogeneity in size and sequence, and with the exception of the RPB2 intein encoded by rpoB, their nature remains largely unknown. There is no correlation between the sites of gene expansion and the presence of repeated sequences in the Floydiella genome; the repeats in expanded genes were found to represent less than 1% of the total amount of repeated sequences. This observation mirrors the situation in the Stigeoclonium chloroplast genome where repeats are largely excluded from a comparable set of expanded genes (Bélanger et al. 2006). Other chlorophyte genomes, including the compact and repeat-poor genomes of Scenedesmus and Oedogonium, share expanded genes with Floydiella cpDNA, reinforcing the idea that expansion of coding sequences occurred independently of repeat proliferation in the Chlorophyceae. Chloroplast coding regions

**Fig. 1.**—Gene map of the Floydiella chloroplast genome. Genes are colored according to their function. Coding sequences on the outside of the map are transcribed in a CW direction. Introns are represented by open boxes; the single intron ORF (in rr) is denoted by a narrow, blue box. The rpoB gene consists of two separate ORFs (rpoBa and rpoBb) that are not associated with sequences typical of group I or group II introns; the rpoBb fragment contains the Fte RPB2 intein. The three ORFs display sequence similarity with group I intron-encoded HNH homing endonucleases. tRNA genes are indicated by the one-letter amino acid code followed by the anticodon in parentheses (Me, elongator methionine; Mf, initiator methionine).
Table 1
General Features of Floydiella and Other Chlorophycean cpDNAs

| Feature                  | OCC clade | CS clade |
|-------------------------|-----------|----------|
|                         | Oedogoniales | Chaetopeltidales | Chaetophorales | Chlamydomonadales | Sphaeropleales |
| Size (bp)               | 196,547 | 521,168 | 223,902 | 203,827 | 161,452 |
| IR                      | 35,492  | —        | —       | —       | —       |
| SC1                     | 80,363  | —        | —       | —       | —       |
| SC2                     | 45,200  | —        | —       | —       | —       |
| A + T (%)               | 70.5    | 65.5     | 71.1    | 65.5    | 73.1    |
| Sidedness index         | 0.74    | 0.91     | 0.95    | 0.87    | 0.88    |
| Conserved genes (no.)   | 99      | 97       | 97      | 94      | 96      |
| Intron                  |         |          |         |         |         |
| Fraction of genome (%)  | 17.9    | 4.3      | 7.9     | 6.8     | 8.6     |
| Group I (no.)           | 17      | 19       | 16      | 5       | 7       |
| Group II (no.)          | 4       | 7        | 5       | 2       | 2       |
| Intergenic sequencesa   |         |          |         |         |         |
| Fraction of genome (%)  | 22.6    | 77.8     | 46.7    | 49.2    | 34.3    |
| Average size (bp)       | 370     | 3,824    | 1,026   | 937     | 517     |
| Short repeated sequencesf|         |          |         |         |         |
| Fraction of genome (%)  | 1.3     | 49.9     | 17.8    | 15.8    | 3.0     |

a Because Floydiella and Stigeoclonium cpDNAs lack an IR, only the total size of this genome is given.

f Nonoverlapping repeated elements ≥ 30 bp were identified as described in the Materials and Methods.

Table 2
Differences between the Repertoires of Conserved Genes in Floydiella and Other Chlorophycean cpDNAs

| Genea     | OCC Clade | CS Clade |
|-----------|-----------|----------|
|           | Oedogonium | Floydiella | Stigeoclonium | Chlamydomonadales | Sphaeropleales |
| infA       | +         | +         | —         | —       | +         |
| petA       | —         | —         | —         | —       | +         |
| psaM       | +         | +         | +         | —       | —         |
| rpl12      | —         | —         | —         | —       | +         |
| rps2       | +         | +         | +         | —       | —         |
| tufA       | +         | +         | +         | —       | —         |
| trnRccu    | +         | +         | +         | —       | —         |
| trnRucg     | +         | +         | —         | —       | —         |
| trnS(gag)  | —         | —         | +         | —       | —         |

a Only the genes that are missing in one or more genomes are indicated. Plus and minus signs denote the presence and absence of genes, respectively. A total of 93 genes are shared by all compared-cpDNAs: atpA, B, E, F, H, I, ccsA, cemA, chla, L, N, clpP, ftsh, petB, D, G, L, psaA, B, C, J, psbA, B, C, D, E, F, H, I, J, K, L, M, N, T, Z, rbcL, rps2, S, 14, 16, 20, 23, 36, rpaA, B, C, D, E, F, H, I, J, K, L, M, N, T, Z, rbcL, rpl2, rps2, trnRccu, trnRucg, trnS(gag).

b Among all completely sequenced chlorophyte cpDNAs, the Oedogonium genome is unique in encoding trnRucg. In a BlastN search against the NCBI database, this chloroplast gene revealed a best hit with the mitochondrial trnRucg gene of the fern Asplenium nidus (E value = 9 × 10⁻¹⁸) followed by hits with numerous bacterial trnRucg and trnRucg genes (E values ranging from 5 × 10⁻⁶ to 6 × 10⁻¹⁵), suggesting that the Oedogonium trnRucg was acquired through horizontal transfer from a mitochondrial or bacterial donor. Interestingly, a mitochondrial origin has previously been reported for two other genes (int and dpoB) unique to the Oedogonium chloroplast (Brouard et al. 2008).
five additional introns compared with the Stigeoclonium and Oedogonium genomes. The core sequences of the Floydiella group I introns are not notably different in size relative to their chlorophycean homologs. However, all four transspliced group II introns are much larger than their homologs in Oedogonium and Stigeoclonium. In case of the psaC intron, considerable expansion was noted for the loop of domain VI (1,501 nt compared with 34 and 296 nt in Oedogonium and Stigeoclonium, respectively).

The exceptionally large size of the Floydiella chloroplast genome is mostly explained by bloated intergenic regions. Among the fully sequenced cpDNAs, this genome is the most loosely packed with genes (table 1). Representing 78.1% of the genome sequence, intergenic regions vary from 68 to 29,364-bp in size, with an average size of 3,824 bp, that is, 3.7-fold larger than observed for the Stigeoclonium genome. The estimated proportion of intergenic DNA in the Volvox genome (Smith and Lee 2009) is only 1.4% lower relative to Floydiella cpDNA. Interestingly, there is accumulating evidence that chloroplast genomes lacking an IR (e.g., those of Chlorella and Leptosira) are more loosely packed with genes relative to their closest relatives having an IR (Parachlorella) and tend to be richer in short dispersed repeats (de Cambiare et al. 2007; Turmel, Otis, and Lemieux 2009). The Floydiella and Stigeoclonium chloroplast genomes conform to these trends.

A Myriad of Short Repeats Populate the Intergenic Regions of the Floydiella Chloroplast Genome

As reported for the Volvox lineage, proliferation of short repeats is mainly responsible for the overall genome expansion in the Floydiella lineage. Repeats larger than 30 bp account for half of the Floydiella genome, an almost 3-fold higher proportion compared with the Stigeoclonium and Chlamydomonas cpDNAs (table 1), both of which are recognized for their high level of repetitive DNA. In Volvox cpDNA, palindromic repeats were found to represent 64% of the partial sequence analyzed and 84% of the identified intergenic regions (vs. 63% for the repeats in the Floydiella intergenic regions) (Smith and Lee 2009). In contrast, short dispersed repeats are scarce in the more compact Oedogonium and Scenedesmus cpDNAs (table 1), representing 1.3% and 3% of these chlorophycean genomes, respectively (table 1). As in other repeat-rich cpDNAs, the great majority of the repeats (>99%) in the Floydiella genome reside in intergenic spacers, the remaining ones being present in expanded genes, psbD, psbI, rps18, and some introns (Ft. psaB.1, Ft. psaC.1, Ft. rbcL.1, Ft. rbcL.2, Ft. rnl.3).

The repeats in the Floydiella chloroplast are extremely diversified in sequence and consist mostly of dispersed repeats. We classified the repeats larger than 30 bp into 196 nonredundant families and for each family identified the sequence and number of copies of the prototype in the genome. As indicated in table 3, the most abundant repeats (i.e., those present in more than 15 copies) represent 26 nonredundant families (designated A through Z) and span 34,246 bp. Most of these repeats are less than 34-bp long and are characterized by mononucleotide repeats. Degenerated versions of these repeats as well as composite repeats formed of two or more repeat units can also be found in the Floydiella genome. The longest composite repeats are 518-bp long and are present at two distant loci. The Floydiella repeats differ from those previously reported in Stigeoclonium, Volvox, Pseudendoclonium, and Oltmansiellopsis cpDNAs by the higher heterogeneity of their sequence and their lesser propensity to adopt secondary structures. Most of the repeats in the latter chlorophyte cpDNAs occur as perfect palindromes or stem-loop structures with loops of a few bases (Pombert et al. 2005, 2006; Belanger et al. 2006; Smith and Lee 2009).

The origin of the dispersed repeats in the Floydiella genome and the process by which these sequences proliferated remain unknown. The palindromic repeats found in the Volvox and Pseudendoclonium chloroplasts have been suggested to descend from a selfish DNA element carried by a mobile intron involved in interorganellar lateral DNA transfers (Pombert et al. 2005; Smith and Lee 2009). Invoking the presence of a putative group-II intron-encoded reverse transcriptase (RT) and a putative group-I intron-encoded endonuclease, Smith and Lee (2009) hypothesized that the palindromic repeats could have been disseminated throughout the Volvox chloroplast genome via a retrotransposition mechanism of mobility. This mechanism, which involves a target DNA-primed reverse transcription step mediated by a RT encoded by a non-long terminal repeat retrotransposable element, was originally proposed to explain the proliferation of a mitochondrial ultra-short element in the mitochondrial genome of the filamentous fungus Podospora anserina (Koll et al. 1996). However, our observation that a RT gene is lacking in the repeat-rich chloroplast genomes of Floydiella and Stigeoclonium but is present in the repeat-poor genomes of Oedogonium and Scenedesmus provides no evidence supporting the hypothesis of RT-mediated proliferation of dispersed repeats.

An Unusually Small Fraction of Mobile Group I Introns in the Floydiella Chloroplast

Nineteen group I introns interrupt six genes in the Floydiella genome. The rRNA operon alone contains 11 introns (eight in rrl and three in rrs), whereas the remaining genes contain one (psaB and trnL(UAA)), two (psbC), or four (psbA) introns (for their predicted positions, sizes, and assigned subgroups, see table 4). In figure 2, the predicted insertion sites of the group I introns are compared with those found in other chlorophycean cpDNAs. Irregular intron distributions are observed at all the 41 insertion sites, except site 2593 in rnl, thus confirming that group I introns are not phylogenetically
informative in the Chlorophyceae (Brouard et al. 2008; Turmel et al. 2008) and that these genetic elements must arise and die relatively frequently. Most of the Floydiella group I introns have positional and structural homologs in other chlorophycean and chlorophyte cpDNAs. To our knowledge, only four map to genomic sites not previously documented for introns. There exists no evidence suggesting that these introns, found in the rrs, rrl, psbA, and psbC genes and belonging to three distinct subgroups, result from group I intron proliferation within the lineage leading to Floydiella because none bears striking similarity with other introns in the Floydiella chloroplast. Although the IAI introns inserted at site 1769 in psaB and at site 276 in psbA appear to be widespread within the Chlorophyceae, these insertion sites have not been found in other completely sequenced green algal cpDNAs; therefore, they might have evolved just before the emergence of the Chlorophyceae.

It is intriguing that there is just one Floydiella group I intron (the rrl intron at site 1065) encoding a potential homing endonuclease when we consider that eight or more mobile group I introns are found in the chloroplasts of the two other representatives of the OCC clade (fig. 2). Also surprising is our finding that the LAGLIDADG endonuclease specified by this unique mobile intron displays sequence similarity with proteins encoded by introns in the mitochondrial rnl, cox1, and atp6 genes of the fungi Smittium culisetae (E value = 3 x 10⁻¹²), Giberella zeae (E value = 5 x 10⁻¹¹), and Neurospora crassa (E value = 9 x 10⁻⁹), respectively. No similarity was observed with the LAGLIDADG endonuclease encoded by the Stigeoclonium chloroplast site-1725 rrl intron, a protein that is closely related to those encoded by group I introns inserted at the same site in the chloroplast genomes of Chlamydomonas species. Furthermore, consistent with the sequence divergence observed between the proteins encoded by the Floydiella and Stigeoclonium site-1065 introns, the primary sequences and putative secondary structures of these introns display substantial dissimilarity. These observations suggest that the single mobile intron in the Floydiella chloroplast is of recent origin and was acquired through lateral transfer of a mobile intron from a mitochondrial genome donor. In this context, it is interesting to mention that a case of horizontal transfer of mobile elements originating from the mitochondria of an unknown donor has also been reported for the Oedogonium chloroplast (Brouard et al. 2008). Coding sequences not carried out by introns (i.e., genes encoding members of the tyrosine recombinase family and type B DNA-directed DNA polymerases) were involved in this horizontal gene transfer.

Another interesting result is our finding that the free-standing orf120, orf150, and orf183 feature similarities with the HNH endonuclease encoded by the Stigeoclonium psbD genes.

### Table 3

| Designation | Prototype Sequence | Size (bp) | Copy Number |
|-------------|--------------------|----------|-------------|
| A           | ACCCGAGCAGAGCTGGCAGAAAGCCTTTT | 30       | 141         |
| B           | CGGGGCGCAGAAADAGAAKAAAGGCTGAAC | 30       | 112         |
| C           | MAMKAGYCTTTTTAAAAGGAGGGG   | 25       | 94          |
| D           | AAAKAGGCGCTTTTTAAGGTTGCA   | 28       | 91          |
| E           | TTTTTTTCTTTTTTACCAAAAGG   | 33       | 62          |
| F           | GCTTTGCAGGCTTGCTTTTTAAAAGGAGGGG | 33   | 60          |
| G           | CCTYTTAAGAKTCTTTTAAAGGCC   | 30       | 55          |
| H           | TAAAACCCCTCAAGAAAGGCTCAATTGCTTC | 33   | 53          |
| I           | CCCCCTCCTCTCTCTTTTTGAAGGAAA | 31       | 44          |
| J           | TTTTTCTTYTCTATCTATTTTMYCTT | 31       | 44          |
| K           | AAAAATGGCCGCCCTCTCTTTAAAGAACGGG   | 32       | 36          |
| L           | GYKTTTTCTTTTTAAAAGGGCCTTTTTAA | 31       | 35          |
| M           | AAATTTTTGCTTCAGTGGGCTTTTACAC | 30       | 33          |
| N           | AGAGGCCTTTTTAAAAGGAGCGGCTC | 30       | 29          |
| O           | CCTGAAACCCAAAAATTAAAAGAGCTGTC | 31       | 29          |
| P           | GGCCCTCCTCCTTTTTTTAAGGTTTCGCTT | 28       | 28          |
| Q           | AAAAGGACGGCTTTCTTCTTTTTAAAGGG   | 30       | 26          |
| R           | AAAGGTTGCAACCCCGAACCCGCTCAAAAA | 30   | 24          |
| S           | GGGCTTTTTAAGAAGGGCTTTTTTTTTT | 30       | 23          |
| T           | GGGCCCTCCTAATTTTTGCTGCTAAAC | 28       | 23          |
| U           | AACCAGACCTAAATTATTGTTGCTGGG   | 31       | 19          |
| V           | GAAAAACCCGAGGAGCTGTCGGGAGGAGGG | 32       | 18          |
| W           | GGTTGCAGCTCTCTCTCTTTTTAAAGRAA | 29       | 17          |
| X           | TTTTCTTTTTAAAAGGCTGTTGGGCTGAC | 30       | 16          |
| Y           | ACTGCCCGCTCTCTTTTTACAGAAAAAA  | 28       | 16          |
| Z           | RDRAGGGCCCTGCTTTTTAAAGAATCT | 27       | 15          |

* Families of nonoverlapping repeats sharing ≥ 90% sequence identities were identified as described in the Materials and Methods.
intron and that the \textit{orf150} is contiguous to \textit{psbD}. The \textit{orf150} displays the complete HNH motif, whereas the two others have retained only the coding region corresponding to the C-terminal portion of the endonuclease. In addition, two free-standing ORFs located between \textit{atpA} and \textit{atpI} (\textit{orf265} and \textit{orf412}) show weak similarities with intron-encoded endonuclease genes found in the \textit{Pseudenuclclia} chloroplast genome. These observations suggest that the five free-standing ORFs are remnants of endonuclease genes that were originally present in group I introns, thereby raising the possibility that colonization of intergenic regions by mobile group I introns could have contributed to their expansion.

### Table 4

| Designation | Predicted Insertion Site\(^a\) | Subgroup\(^b\) | Size (bp) |
|-------------|-------------------------------|---------------|-----------|
| Group I introns |                             |               |           |
| \textit{Ft.psaB.1} | 1769                         | 1A1           | 851       |
| \textit{Ft.psaB.2} | 276                          | 1A1           | 372       |
| \textit{Ft.psaB.3} | 333                          | 1B            | 331       |
| \textit{Ft.psaB.4} | 414                          | 1A1           | 412       |
| \textit{Ft.psaB.5} | 790                          | 1B            | 425       |
| \textit{Ft.psbC.1} | 579                          | 1A2           | 695       |
| \textit{Ft.psbC.2} | 1089                         | 1A1           | 820       |
| \textit{Ft.rrs.1}  | 508                          | 1A3           | 257       |
| \textit{Ft.rrs.2}  | 531                          | 1A3           | 339       |
| \textit{Ft.rss.1}  | 692                          | 1AI           | 256       |
| \textit{Ft.rrf.1}  | 958                          | 1AI           | 321       |
| \textit{Ft.rrf.2}  | 1065                         | 1AI           | 1275\(^c\) |
| \textit{Ft.rrf.3}  | 1766                         | 1AI           | 449       |
| \textit{Ft.rrf.4}  | 1931                         | 1B            | 406       |
| \textit{Ft.rrf.5}  | 2449                         | 1A1           | 375       |
| \textit{Ft.rrf.6}  | 2590                         | 1A1           | 381       |
| \textit{Ft.rrf.7}  | 2511                         | 1A3           | 402       |
| \textit{Ft.rrf.8}  | 2596                         | 1A3           | 431       |
| \textit{Ft.trnl(uaa.1) | 35                          | 1C3           | 997       |
| Group II introns |                               |               |           |
| \textit{cis-spliced} |                             |               |           |
| \textit{Ft.psaB.1} | 80                           | IIA           | 876       |
| \textit{Ft.rbcL.3} | 285                          | IIA           | 1672      |
| \textit{Ft.rbcL.4} | 1225                         | IIB           | 940       |
| \textit{trans-spliced} |                             |               |           |
| \textit{Ft.psaC.1} | 25                           | IIB (I)       | 2532      |
| \textit{Ft.petD.1} | 4                            | IIB (I)       | 1313      |
| \textit{Ft.rbcL.1} | 67                           | IIB (II)      | 2672      |
| \textit{Ft.rbcL.2} | 120                          | IIA (I)       | 1598      |

\(^{a}\) Insertion sites of introns in genes coding for rRNAs and proteins are given relative to the corresponding genes in \textit{Mesostigma} cpDNA, whereas those in \textit{rrs} and \textit{mf} are given relative to \textit{Escherichia coli} 16S and 23S rRNAs, respectively. For each insertion site, the position corresponding to the nucleotide immediately preceding the intron is reported.

\(^{b}\) Group I introns were classified according to Michel and Westhof (1990), whereas classification of group II introns was according to Michel et al. (1989). For each trans-spliced intron, the domain containing the site of discontinuity is indicated in parentheses. An homing endonuclease of 431 amino acids with two copies of the LAGJEDAG motif is encoded in loop L9 of the \textit{Ft.rl.2} intron.

### The \textit{Floydiella} Trans-Spliced Group II Introns Have Structural Homologs in Other Members of the OCC Lineage

Our recent phylogenomic analyses were somewhat ambiguous regarding the branching order of the Oedogoniales, Chaetophorales, and Chaetopeltidales; nevertheless, based on genomic features, in particular the presence/absence of \textit{trans}-spliced group II introns at common sites, we favored the hypothesis that the Oedogoniales diverged before the Chaetophorales and the Chaetopeltidales (Turmel et al. 2008). \textit{Trans}-spliced group II introns are the products of rare recombination events leading to the fragmentation of \textit{cis}-spliced introns, and their reversion to \textit{cis}-spliced introns is thought to be very unlikely (Malek et al. 1996; Malek and Knoop 1998). Each fragmentation event occurs within a \textit{cis}-spliced group II intron sequence, so that the regions 5’ and 3’ of the breakpoint become part of independent transcription units, often located far apart on the genome (Michel et al. 1989). The separate intron pieces derived from these transcription units can assemble in trans at the RNA level to reconstitute a complete and fully spliceable intron structure.

Our analysis of the complete set of group II introns present in \textit{Floydiella} is consistent with the view that \textit{trans}-spliced group II introns are reliable phylogenetic markers (fig. 2). Three \textit{cis}-spliced and four \textit{trans}-spliced group II introns, none of which is mobile, occur in the \textit{Floydiella} chloroplast (table 4). The \textit{rbcL} gene contains two \textit{trans}-spliced and two \textit{cis}-spliced introns, and the two remaining \textit{trans}-spliced introns are located in \textit{psaC} and \textit{petD}. The \textit{cis}-spliced \textit{psbA} intron is the only group II intron that was not reported earlier (Turmel et al. 2008). All three \textit{cis}-spliced introns lack homologs inserted at identical gene positions in previously investigated cpDNAs and are thus lineage specific. The \textit{trans}-spliced introns, however, have positional and structural homologs in one (\textit{rbcL} introns) or two other members (\textit{petD} and \textit{psaC} introns) of the OCC lineage (fig. 2). The \textit{Floydiella} \textit{trans}-spliced \textit{petD} and \textit{psaC} introns as well as the first \textit{trans}-spliced intron in \textit{rbcL} were modeled as group IIB introns. As mentioned above, the \textit{psaC} intron is peculiar in featuring an oversized loop in domain VI. Like its \textit{Stigeoclonium} homolog (Bélanger et al. 2006), the second \textit{trans}-spliced intron in \textit{rbcL} is missing domains IA and IB; this is an unusual characteristic for group IIA introns. Sites of discontinuities were mapped in domain I for the introns in \textit{petD}, \textit{psaC} and the first \textit{rbcL} intron and near domain II for the second \textit{rbcL} intron.

### Analysis of Chloroplast Gene Order Supports a Close Relationship between the Chaetophorales and Chaetopeltidales

Pairwise comparisons of overall gene order in the \textit{Floydiella}, \textit{Oedogonium}, and \textit{Stigeoclonium} chloroplast genomes suggest that the \textit{Floydiella} genome is more closely related to its...
Stigeoclonium counterpart. Fourteen gene clusters including a total of 39 genes are conserved between the latter genomes. By comparison, the Floydiella and Oedogonium genomes share 11 gene clusters encoding 34 genes, whereas the Stigeoclonium and Oedogonium cpDNAs share eight clusters comprising 26 genes.

The Floydiella chloroplast genome resembles its chlorophycean counterparts in lacking most of the ancestral gene clusters found in other chlorophyte cpDNAs. Like the Oedogonium and Stigeoclonium genomes, it has lost the triad psbH-psbN-psbT and the gene pairs rpl14-rpl5 and rpl2-rps19, all of which are present in the chloroplasts of representatives of the CS clade (fig. 3A). The chaetopeltidalean alga has retained the same set of ancestral gene clusters as Oedogonium plus the rps12-rps7 gene pair.

On the other hand, the gene clusters that were recently acquired by the chlorophycean lineage robustly support a specific affiliation between the Chaetopeltidales and Chaetophorales (fig. 3B). Floydiella specifically shares the triads atpA-atpI-petG, psaC(ex1)-psbN-psaC(ex2) and the gene pairs atpF-rpl16, chlN-psaM, and psbT-psbH with the two other representatives of the OCC clade. These clusters are shown as nine gene pairs in figure 3B. Nine additional derived gene pairs, forming seven clusters, are shared exclusively by Floydiella and Stigeoclonium. In contrast, only three derived gene pairs are common to Oedogonium and Stigeoclonium and a single pair unites Oedogonium and Floydiella.

**Phylogenies Inferred from Sequence Data and Genomic Features Are Congruent in Identifying the Oedogoniales as the First Branch of the OCC Lineage**

To examine the branching order of the three recognized lineages of the OCC clade, we analyzed an amino acid data set (14,101 sites) and a nucleotide data set (codons excluding third positions, 31,858 sites) derived from 69 protein-coding genes of 13 completely sequenced chlorophyte chloroplast genomes (fig. 4). Both the protein and gene trees placed the Oedogoniales before the divergence of

![Fig. 2.](https://academic.oup.com/gbe/article-abstract/doi/10.1093/gbe/evq014/570533) /by guest on 27 July 2018

denoted by colored numbers. In the last column are indicated the introns of Chlamydomonas species other than Chlamydomonas reinhardtii that have homologs in completely sequenced chlorophycean algal genomes. References for the latter introns are as follows: psaB (Turmel, Mercier, and Côté 1993), psbA (Turmel et al. 1989), psbB (Turmel, Mercier, and Côté 1993), rrs (Duracer et al. 1989; Turmel, Mercier, et al. 1995), and rrf (Turmel et al. 1991; Côté et al. 1993; Turmel, Gutell, et al. 1993; Turmel, Côté, et al. 1995). An asterisk denotes the absence of the ORF in some Chlamydomonas species. Intron insertion sites are designated as indicated in table 4. Oc, Oedogonium cardiaicum; Ft, Floydiella terrestris; Sh, Stigeoclonium helveticum; So, Scenedesmus obliquus; Cr, Chlamydomonas reinhardtii; C, Chlamydomonas species.
Although weaker bootstrap support was observed in the gene tree. This observation represents a significant improvement in resolution, considering that the ML tree reconstructed from nucleotide data in our previous analyses of data sets derived from 44 protein-coding genes (Turmel et al. 2008) differed from the corresponding protein tree in showing the Chaetopeltidales as the first branch of the OCC clade (T2 topology). Moreover, the sister relationship between the Chaetophorales and the Chaetopeltidales received stronger support (97%) in the protein tree reported here compared with the ML tree inferred earlier from 44 proteins (71%) (Turmel et al. 2008).

To gain independent evidence that the T1 topology reflects the true interrelationships between the Oedogoniales, Chaetophorales, and the Chaetopeltidales, structural genomic characters were mapped on the three possible topologies of the OCC clade, and the lengths of the resulting trees were compared (fig. 5). Only the parsimoniously informative characters that evolved in the OCC lineages were examined in this analysis. As expected, the most parsimonious tree (25 steps) was consistent with the T1 topology. The trees with the alternative T2 and T3 topologies comprised 11 and 12 extra steps, respectively, which are mainly attributable to convergent IR losses and acquisitions of trans-spliced rbcL introns and derived gene pairs. Our phylogenetic analyses based on sequence data and genomic characters are thus congruent in supporting the notion that the Oedogoniales diverged before the Chaetopeltidales and the Chaetophorales.

Several phylogenetic studies based on nuclear-encoded rRNA sequences also placed the Oedogoniales at a basal position but failed to resolve the relationships among the five major groups of the Chlorophyceae (Booton et al. 1998; Buchheim et al. 2001; Krienitz et al. 2003; Müller et al. 2004; Alberghina et al. 2006). Pickett-Heaps (1975) speculated that the Oedogoniales represent the earliest branch of an evolutionary lineage that gave rise to filamentous taxa currently included in the Chaetophorales. Although radically different from those observed in the Chaetophorales and the Chaetopeltidales (O’Kelly et al. 1994), the flagellar apparatus of the Oedogoniales can be viewed as a modification of the cruciate arrangement of basal bodies, which appeared with the proliferation of the flagella (Moestrup 1982; Van den Hoek et al. 1995). The fibrous ring of the flagellar apparatus of the Oedogoniales presumably arose from the repetition of the upper transversely striated fiber.

**Fig. 3.**—Conservation of ancestral and derived gene pairs in fully sequenced chlorophycean chloroplast genomes. (A) Conserved gene pairs dating back to a distant chlorophyte ancestor (3’ psaI-5’ rps12) or to the last common ancestor of all green plants (all other gene pairs). (B) Conserved gene pairs that emerged during the evolution of the Chlorophyceae. For each gene pair, adjoining termini of the genes are indicated. Dark boxes indicate the presence of gene pairs with the same polarities in two or more genomes, whereas light or open boxes indicate the absence of gene pairs. A light box indicates that the two genes associated with a gene pair are found in the genome but are unlinked. An open box indicates that one or both genes associated with a gene pair are absent from the genome. Gene pairs linked by brackets are contiguous on the genome. Six categories of derived gene pairs were distinguished according to their distribution: 1) those present in all three lineages of the OCC clade (OCC), 2) those supporting a sister relationship between the Chaetophorales and the Chaetopeltidales (T1), 3) those supporting a sister relationship between the Oedogoniales and Chaetophorales (T2), 4) the single gene pair supporting a sister relationship between the Oedogoniales and Chaetopeltidales (T3), 5) those present in both lineages of the CS clade (CS), and 6) the three remaining gene pairs found in some lineages of the OCC and CS clades.
interconnecting the basal bodies in other chlorophycean lineages (Pickett-Heaps 1975; Van den Hoek et al. 1995).

Dynamic Evolution of the Chloroplast Genome in the Chlorophyceae

Considering that the chloroplast genomes of the green algae representing the five recognized lineages of the Chlorophyceae vary considerably in architecture and bear little similarity with other chlorophyte chloroplast genomes, it is difficult to pinpoint the main factors responsible for the extraordinarily dynamic evolution of these genomes. Although the detailed suite of events that led to their widely differing architectures is poorly understood, the origins of some genomic characters can be traced. As shown in the evolutionary scenario presented in figure 6, major genomic changes were mapped at all internal nodes of the chlorophycean phylogeny, implying that all steps of lineage diversification were accompanied by important reorganization of the chloroplast genome. Some events such as the breakup of rpoB into two contiguous ORFs, the disruption of multiple ancestral operons and the expansion of clpP, rps3, and rps4 coincided with the appearance of the Chlorophyceae, whereas gains of numerous derived gene pairs and transspliced group II introns marked the emergence and subsequent divergence of the CS and OCC clades.

FIG. 5.—Scenarios of gains/losses of chloroplast genomic features predicted by the three possible branching orders of the OCC lineages (T1, T2, and T3). Gains of derived gene pairs, trans-spliced rbcL introns (rbcL_i67 and rbcL_i120), and the RPB2 intein are denoted by blue symbols, whereas losses of IR, derived gene pairs and the RPB2 intein are denoted by orange symbols. Characters supporting a clade are denoted by squares, whereas homoplastic characters are denoted by triangles.
The large DNA segment separating rpoBa and rpoBb in the common ancestor of all chlorophycean algae became the target of recombinational events in the common ancestor of the OCC algae, resulting in the localization of the two gene pieces at distant loci (fig. 6). The rpoB gene is likely functional in chlorophycean algae because gene disruption of this gene has revealed an essential function in C. reinhardtii (Fischer et al. 1996) and also because no chloroplast-targeted RNA polymerase gene was identified in the nuclear genome of C. reinhardtii (Merchant et al. 2007). Splitting of rpoB took place independently in the trebouxiophyte lineage leading to Leptosira (de Cambiaire et al. 2007) and...
therefore was not an unprecedented event during chlorophyte evolution. Genes encoding other subunits of the chloroplast RNA polymerase \((rpoC1\) and \(rpoC2\)) sustained fragmentation in the lineages leading to \(C.\) reinhardtii and \(C.\) moewusii (Turmel et al. 2008).

Chloroplast-encoded components of chloroplast ribosomes also continued to evolve under relaxed constraints in the CS lineages, as \(rps2\) was fragmented into two pieces and the 3’ end of \(rps4\) was trimmed of the last 40 codons. Because the latter region is highly conserved in bacterial and all other chloroplast \(rps4\) homologs, we examined the possibility that it could be distantly located from the 5’ end of \(rps4\) in chlamydomonadalean and sphaeroplealean cpDNAs; however, our searches were unsuccessful. Evidence that the \(rps4\) genes of these genomes must be functional comes from a proteomic analysis of the chloroplast ribosome from \(C.\) reinhardtii (Yamaguchi et al. 2003). The finding that the missing 3’ conserved sequence is immediately adjacent to the site of the prominent insertion sequence characterizing the \(rps4\) genes of the OCC lineages led us to envision that this insert was initially gained by the last common ancestor of all chlorophyceans and was subsequently lost along with the 3’ conserved coding region before the emergence of the CS clade. It is well documented that chloroplast ribosomes contain proteins with extensions relative to their bacterial homologs as well as unique proteins (Yamaguchi et al. 2002, 2003; Manuell et al. 2007). A recent cryo-electron microscopy study of the \(C.\) reinhardtii chloroplast ribosome identified chloroplast-specific domains in the small subunit as novel structural additions to a basic bacterial ribosome (Manuell et al. 2007). Among the additional domains visualized in this study is that corresponding to the major insertion responsible for the expansion of the chloroplast \(rps3\) gene in chlorophyceans. The observed chloroplast-unique ribosomal domains/proteins were located at optimal positions for interactions with mRNAs, prompting the hypothesis that they interact with chloroplast-specific translation factors and RNA elements to facilitate the regulation of translation. A large body of evidence has indicated that translation is the key regulated step in chloroplast gene expression (Zerges 2000). In contrast, bacterial gene expression is strongly influenced by the rate of transcription, and translation and transcription are often closely coupled.

The establishment of group II introns in chlorophycean chloroplasts was crucial in modeling the genomic landscape, but the origin of these introns remain elusive. Group II introns are rare among the chlorophyte chloroplast genomes examined so far, and \(trans\)-spliced group II introns have been found only in the Chlorophyceae (Lemieux et al. 2007; Turmel, Gagnon, et al. 2009; Turmel, Otis, and Lemieux 2009). Because \(trans\)-spliced group II introns were undoubtedly derived from \(cis\)-spliced versions of cognate introns (Malek et al. 1996; Malek and Knoop 1998), it is intriguing that no putative \(cis\)-spliced intron precursors were uncovered in the chlorophycean cpDNAs examined to date. The absence of such precursors undoubtedly reflects the extreme scrambling in gene order sustained by the chloroplast genome during the evolution of chlorophyceans. The extraordinarily fluid architecture of the chlorophycean genome is thought to result predominantly from intramolecular and intermolecular recombination between homologous and nonhomologous regions, with the presence of numerous dispersed repeats enhancing opportunities for recombinational exchanges. Obviously, complete cpDNA sequences from close relatives of chlorophycean green algae are needed to better understand the dynamics of chloroplast genome evolution in the Chlorophyceae.

Our data do not suggest that there is a positive correlation between the extent of gene rearrangements and the rate of sequence evolution observed for chloroplast genomes in the Chlorophyta. Instead, although the five main chlorophycean lineages show extreme rearrangements and also differ considerably from other chlorophyte lineages in terms of chloroplast gene order, the phylogenies inferred from multiple chloroplast genes do not reveal any radical length differences for the branches of chlorophycean lineages as compared with other chlorophyte lineages (fig. 4). In contrast, a positive correlation between changes in gene order, gene/intron loss, and lineage-specific rate acceleration has been identified in a recent study of chloroplast genomes from a broad sampling of photosynthetic angiosperms (Guisinger et al. 2008). For the family Geraniaceae, which features extreme changes in gene content and order relative to the typically conserved chloroplast genomes of most angiosperms (the IR-containing genome of Pelargonium \(x\) hortorum contains dispersed repeats and is the largest and most rearranged land plant genome completely sequenced so far), accelerated rates of sequence evolution were observed for the ribosomal protein and RNA polymerase genes (Guisinger et al. 2008). To explain their observations, Guisinger et al. (2008) proposed a model of aberrant DNA repair coupled with altered gene expression. According to this model, improper repair arising from mutations in organellar-targeted \(rec\) genes would lead not only to genome rearrangements and increased substitution rates but also to extreme size variation. Moreover, possible transcriptional control of chloroplast genes by the nucleus following loss of \(rpoA\) function (\(rpoA\)-like ORFs are found in Pelargonium cpDNA) would result in altered gene expression and nucleotide substitutions. It is remarkable that RNA polymerase and ribosomal protein genes are affected in both chlorophycean and Geraniaceae chloroplast genomes; this may be a common feature of highly rearranged genomes.

Contrasting with their uniformity in gene content, the 3-fold size variation displayed by chlorophycean chloroplast genomes raises questions about the regulation of genome size. The positive correlation observed between genome size and the proportion of noncoding and repeated DNA in
chlorophycean chloroplasts are in concordance with the selfish-DNA hypothesis. According to this hypothesis, accumulation of noncoding DNA is caused by the proliferation of selfish elements, which in turn is limited by the harmful effects of these elements on host fitness (Doolittle and Sapienza 1980; Orgel and Crick 1980; Gregory 2001; Lynch 2007). Still, little is known about how genome size is regulated even though various models have been proposed (Petrov 2002; Oliver et al. 2007; Pettersson et al. 2009). The relative rates of small insertions and deletions and the degree to which these mutations are favored or not by natural selection appear to be the main forces driving genome size evolution (Petrov et al. 2000; Lynch 2007).

Is the 521 kb Floydiella cpDNA near the high end of size variation for the chloroplast genome? A broader sampling of chlorophycean green algae will be necessary to answer this question. We have shown here that the increased intergenic regions account largely for the expansion of the Floydiella genome and that these regions consist primarily of dispersed repeats, but also to a minor extent, of remnants of homing endonuclease genes derived from degenerated mobile group I introns. The absence of homing endonuclease genes in almost all Floydiella group I introns is particularly intriguing, as this observation contrasts sharply with the higher proportion of mobile group I introns in other chlorophycean genomes. It is tempting to speculate that mobile introns were once present in the common ancestor of the Chaetopeltidales and that the homing endonuclease genes conferring their mobility were extinguished because of their role in amplifying noncoding DNA through intron transpositions and constraints to eliminate excessive noncoding DNA in the Floydiella lineage. Of course, the paucity of mobile introns and presence of remnants of endonuclease genes may be simply coincidental and unrelated to the pressure to reduce the size of a burdened genome. The chloroplast genomes of closely related chlorophycean green algae will need to be analyzed to gain deeper insight into the forces driving the evolution of genome size in the Chlorophyceae.

Acknowledgments

This study was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (to M.T. and C.L.).

Literature Cited

Alberghina JS, Vigna MS, Confalonieri VA. 2006. Phylogenetic position of the Oedogoniales within the green algae (Chlorophyta) and the evolution of the absolute orientation of the flagellar apparatus. PI Syst Evol. 261:151–163.

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

Andersen RA, Berges JA, Harrison PJ, Watanabe MM. 2005. Appendix A—recipes for freshwater and seawater media. In: Andersen RA, editor. In: Algal culturing techniques. Burlington (VT): Elsevier Academic Press. p. 429–538.

Bakker ME, Lokhorst GM. 1984. Ultrastructure of Drapanoalga glomerata (Chaetophorales, Chlorophyceae). I. The flagellar apparatus of the zoospore. Nord J Bot. 4:261–273.

Bélanger A-S, et al. 2006. Distinctive architecture of the chloroplast genome in the chlorophycean green alga Stigeoclonium helveticum. Mol Genet Genomics. 276:464–477.

Booton GC, Floyd GL, Fuerst PA. 1999. Origins and affinities of the filamentous green algal orders Chaetophorales and Oedogoniales based on 18S rRNA gene sequences. J Phycol. 34:312–318.

Boudreau E, Otis C, Turmel M. 1994. Conserved gene clusters in the highly rearranged chloroplast genomes of Chlamydomonas moewusii and Chlamydomonas reinhardtii. Plant Mol Biol. 24:585–602.

Boudreau E, Turmel M. 1995. Gene rearrangements in Chlamydomonas chloroplast DNAs are accounted for by inversions and by the expansion/contraction of the inverted repeat. Plant Mol Biol. 27:351–364.

Boudreau E, Turmel M. 1996. Extensive gene rearrangements in the chloroplast DNAs of Chlamydomonas species featuring multiple dispersed repeats. Mol Biol Evol. 13:233–243.

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

Buchheim MA, Michalopoulos EA, Buchheim JA. 2001. Phylogeny of the Chlorophyceae with special reference to the Sphaeropleales: a study of 18S and 26S rDNA data. J Phycol. 37:819–835.

Cai Z, et al. 2008. Extensive reorganization of the plastid genome of Trifolium subterraneum (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. J Mol Evol. 67:696–704.

Chumley T, et al. 2006. The complete chloroplast genome sequence of Pelargonium x hortorum: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol Biol Evol. 23:2175.

Côté V, Mercier JP, Lemieux C, Turmel M. 1993. The single group-I intron in the chloroplast rrnL gene of Chlamydomonas humicola encodes a site-specific DNA endonuclease (i-Chul). Gene. 129:69–76.

Cui L, et al. 2006. Adaptive evolution of chloroplast genome structure inferred using a parametric bootstrap approach. BMC Evol Biol. 6:13.

de Cambiaria J-C, Otis C, Lemieux C, Turmel M. 2006. The complete chloroplast genome sequence of the chlorophycean green alga Scenedesmus obliquus reveals a compact gene organization and a biased distribution of genes on the two DNA strands. BMC Evol Biol. 6:37.

de Cambiaria J-C, Otis C, Lemieux C, Turmel M. 2007. The chloroplast genome sequence of the green alga Leptosira terrestris: multiple losses of the inverted repeat and extensive genome rearrangements within the Trebouxiophyceae. BMC Genomics. 8:212.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature. 284:601–603.

Duroucher V, Gauthier A, Bellermare G, Lemieux C. 1989. An optional group-I intron between the chloroplast small subunit rRNA genes of Chlamydomonas moewusii and C. eugametos. Curr Genet. 15:277–282.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
Fawley MW, Yun Y, Qin M. 2000. Phylogenetic analyses of 18S rDNA sequences reveal a new coccoid lineage of the Prasinophyceae (Chlorophyta). J Phycol. 36:387–393.

Fischer N, Stampacchia O, Redding K, Rochaix JD. 1996. Selectable marker recycling in the chloroplast. Mol Gen Genet. 251:373–380.

Floyd GL, Hoops HJ, Swanson JA. 1980. Fine structure of the zoosporangium of Ulothrix belkai with emphasis on the flagellar apparatus. Protoplasma. 104:17–32.

Gregory TR. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev Camb Philos Soc. 76:65–101.

Guilou L, et al. 2004. Diversity of picoplanktonic prasinophytes assessed by direct nuclear SSU rDNA sequencing of environmental samples and novel isolates retrieved from oceanic and coastal marine ecosystems. Protist. 155:193–214.

Guisinger MM, Kuehl JV, Boore JL, Jansen RK. 2008. Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. Proc Natl Acad Sci U S A. 105:18424–18429.

Haberle RC, Fourcade HM, Boore JL, Jansen RK. 2008. Extensive rearrangements in the chloroplast genome of Trachelium caeruleum are associated with repeats and tRNA genes. J Mol Evol. 66:350–361.

Jansen RK, et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci U S A. 104:19369–19374.

Jobb G, von Haeseler A, Strimmer K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol. 4:18.

Koll F, Boulay J, Belcour L, d’Aubenton-Carafa Y. 1996. Contribution of ultra-short invasive elements to the evolution of the mitochondrial genome in the genus Podospora. Nucleic Acids Res. 24:1734–1741.

Kriemitz L, Hegewald E, Hepperle D, Wolf A. 2003. The systematics of coccoid green algae: 18S rRNA gene sequence data versus morphology. Botologia. 58:437–446.

Kurtz S, et al. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 29:4633–4642.

Lee HL, Jansen RK, Chumley TW, Kim KJ. 2007. Gene relocations within the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. Mol Phylogenet Evol. 427–528.

Lee HL, Jansen RK, Chumley TW, Kim KJ. 2007. Gene relocations within the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. Mol Phylogenet Evol. 29:58–66.

O’Kelly CJ, Floyd GL. 1984. Flagellar apparatus absolute orientations and the phylogeny of the green algae. Biosystems. 16:227–251.

O’Kelly CJ, Watanabe S, Floyd GL. 1994. Ultrastructure and phylogenetic relationships of Chaetopeltidales ord nov (Chlorophyta, Chlorophyceae). J Phycol. 30:118–128.

Oliver MJ, Petrov D, Ackerly D, Falkowski P, Schofield OM. 2007. The mode and tempo of genome size evolution in eukaryotes. Genome Res. 17:594–610.

Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. Nature. 284:604–607.

Pettersson ME, Kurland CG, Berg OG. 2009. Deletion rate evolution and the phylogeny of the green algae. Biosystems. 16:227–251.

Petrov DA. 2002. Mutational equilibrium model of genome size evolution. Theor Popul Biol. 61:531–544.

Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. 2000. Evidence for DNA loss as a determinant of genome size. Science. 287:1060–1062.

Petersson ME, Kurland CG, Berg OG. 2009. Deletion rate evolution and its effect on genome size and coding density. Mol Biol Evol. 26:1421–1430.

Pickett-Heaps J. 1975. Green algae: structure, reproduction and evolution in selected genera. Sunderland (MA): Sinauer Associates, Inc.

Pombert J-F, Lemieux C, Turmel M. 2006. The complete chloroplast DNA sequence of the green alga Oltmannsia spiculata reveals a distinctive quadripartite architecture in the chloroplast genome of early diverging ulvophytes. BMC Biol. 4:3.

Pombert JF, Otis C, Lemieux C, Turmel M. 2004. The complete mitochondrial DNA sequence of the green alga Pseudendoclonium akinetum (Ulvophyceae) highlights distinctive evolutionary trends in the Chlorophyta and suggests a sister-group relationship between the Ulvophyceae and Chlorophyceae. Mol Biol Evol. 21:922–935.
Pombert JF, Otis C, Lemieux C, Turmel M. 2005. The chloroplast genome sequence of the green alga *Pseudodanionium akinetum* (Ulvophyceae) reveals unusual structural features and new insights into the branching order of chlorophyte lineages. Mol Biol Evol. 22:1903–1918.

Qiu YL, et al. 2006. The deepest divergences in land plants inferred from phylogenomic evidence. Proc Natl Acad Sci U S A. 103:15511–15516.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16:276–277.

Rogers MB, Gilson PR, Su V, McFadden GI, Keeling PJ. 2007. The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. Mol Biol Evol. 24:54–62.

Sears BB, Chiu W-L, Wolfson R. 1995. Replication slippage as a molecular mechanism for evolutionary variation in chloroplast DNA due to deletions and insertions. In: Tsenewaki K, editor. Plant genome and plastome: their structure and evolution. Tokyo (Japan): Kodansha Scientific Ltd. p. 139–146.

Shoup S, Lewis LA. 2003. Polyphyletic origin of parallel basal bodies in swimming cells of Chlorophycean green algae (Chlorophyta). J Phycol. 39:789–796.

Smith DR, Lee RW. 2009. The mitochondrial and plastid genomes of *Prasinophyceae*. Mol Biol Evol. 26:2317–2331.

Steinkötter J, Bhattacharya D, Semmelroth I, Bibeau C, Melkonian M. 1994. Prasinophytes form independent lineages within the Chlorophyta: evidence from ribosomal RNA sequence comparisons. J Phycol. 30:340–345.

Turmel M, Boulanger J, Lemieux C. 1989. Two group I introns with long internal open reading frames in the chloroplast *psbA* gene of *Chlamydomonas moewusii*. Nucleic Acids Res. 17:3875–3887.

Turmel M, Boulanger J, Schnare MN, Gray MW, Lemieux C. 1991. Six group I introns between different subcellular compartments (chloroplast and mitochondrion). Mol Biol Evol. 12:533–545.

Turmel M, Gagnon MC, O’Kelly CJ, Otis C, Lemieux C. 2009. The chloroplast genomes of the green algae *Pyramimonas*, *Monomastix*, and *Pyrococculus* shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. Mol Biol Evol. 26:631–648.

Turmel M, Gutell RR, Mercier J-P, Otis C, Lemieux C. 1993. Analysis of the chloroplast large subunit ribosomal RNA gene from 17 *Chlamydomonas* taxa. Three internal transcribed spacers and 12 group I intron insertion sites. J Mol Biol. 232:446–460.

Turmel M, Mercier J-P, Côté M-J. 1993. Group I introns interrupt the chloroplast *psaB* and *psbC* and the mitochondrial *rml* gene in *Chlamydomonas*. Nucleic Acids Res. 21:5242–5250.

Turmel M, Mercier J-P, Côté V, Otis C, Lemieux C. 1995. The site-specific DNA endonuclease encoded by a group I intron in the *Chlamydomonas palidostigmatica* chloroplast small subunit rRNA gene introduces a single-strand break at low concentrations of Mg2+. Nucleic Acids Res. 23:2519–2525.

Turmel M, Otis C, Lemieux C. 1999. The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes. Proc Natl Acad Sci U S A. 96:10248–10253.

Turmel M, Otis C, Lemieux C. 2005. The complete chloroplast DNA sequences of the chlorophycean green algae *Staurastrum* and *Zygnema* reveal that the chloroplast genome underwent extensive changes during the evolution of the Zygnematales. BMC Biol. 3:22.

Turmel M, Otis C, Lemieux C. 2009. The chloroplast genomes of the green alga *Pedinomonas minor*, *Parachlorella kessleri*, and *Oocystis solitaria* reveal a shared ancestry between the Pedinomonadales and Chlororellales. Mol Biol Evol. 26:2317–2331.

Turmel M, Pombert JF, Charlebois P, Otis C, Lemieux C. 2007. The green algal ancestry of land plants as revealed by the chloroplast genome. Int J Plant Sci. 168:679–689.

Van den Hoek C, Mann DG, Jahns HM. 1995. Algae: an introduction to phycology. Cambridge: Cambridge University Press.

Wakasugi T, et al. 1997. Complete nucleotide sequence of the chloroplast genome from the green alga *Chlorella vulgaris*: the existence of genes possibly involved in chloroplast division. Proc Natl Acad Sci U S A. 94:5967–5972.

Watanabe S, Floyd GL. 1989. Ultrastructure of the quadriflagellate zoospores of the filamentous green algae *Chaetophora incrassata* and *Pseudoschizomeris caudata* (Chaetophorales, Chlorophyceae) with emphasis on the flagellar apparatus. Bot Mag Tokyo. 102:533–546.

Yamaguchi K, et al. 2003. Proteomic characterization of the *Chlamydomonas reinhardtii* chloroplast ribosome. Identification of proteins unique to the 70 S ribosome. J Biol Chem. 278:33774–33785.

Yamaguchi K, et al. 2002. Proteomic characterization of the small subunit of *Chlamydomonas reinhardtii* chloroplast ribosome: identification of a novel S1 domain-containing protein and unusually large orthologs of bacterial S2, S3, and S5. Plant Cell. 14:2957–2974.

Zerges W. 2000. Translation in chloroplasts. Biochimie. 82:583–601.

**Associate editor:** William Martin