

# The Complete Chloroplast and Mitochondrial DNA Sequence of *Ostreococcus tauri*: Organelle Genomes of the Smallest Eukaryote Are Examples of Compaction

Steven Robbens,\* Evelyn Derelle,† Conchita Ferraz,‡ Jan Wuyts,\* Hervé Moreau,† and Yves Van de Peer\*

\*Department of Plant Systems Biology, Flanders Institute for Biotechnology (VIB), Ghent University, Ghent, Belgium; †UMR 7628 CNRS, Université Paris VI, Laboratoire Arago, Banyuls sur Mer, France; and ‡Institut de Génétique Humaine, UPR CNRS 1142, Montpellier, France

The complete nucleotide sequence of the mt (mitochondrial) and cp (chloroplast) genomes of the unicellular green alga *Ostreococcus tauri* has been determined. The mt genome assembles as a circle of 44,237 bp and contains 65 genes. With an overall average length of only 42 bp for the intergenic regions, this is the most gene-dense mt genome of all Chlorophyta. Furthermore, it is characterized by a unique segmental duplication, encompassing 22 genes and covering 44% of the genome. Such a duplication has not been observed before in green algae, although it is also present in the mt genomes of higher plants. The quadripartite cp genome forms a circle of 71,666 bp, containing 86 genes divided over a larger and a smaller single-copy region, separated by 2 inverted repeat sequences. Based on genome size and number of genes, the *Ostreococcus* cp genome is the smallest known among the green algae. Phylogenetic analyses based on a concatenated alignment of cp, mt, and nuclear genes confirm the position of *O. tauri* within the Prasinophyceae, an early branch of the Chlorophyta.

## Introduction

The so-called green lineage (Viridiplantae) is divided into 2 major divisions, namely, Streptophyta and Chlorophyta. Streptophyta contain all known land plants and their immediate ancestors, a group of algae known as “charophyte green algae” (e.g., *Chaetosphaeridium globosum*), whereas Chlorophyta contain the other green algae (e.g., *Chlamydomonas reinhardtii*) that form a monophyletic assemblage and are a sister group to the Streptophyta (Graham and Wilcox 2000). So far, only 25 complete mt (mitochondrial) genomes have been sequenced for representatives of the green lineage, 17 from Streptophyta and 8 from Chlorophyta. Regarding plastid genomes, 68 genome sequences are available in public databases, of which 60 are from Streptophyta and 8 from Chlorophyta.

The mt genomes of chlorophytes are usually small (25–90 kb), whereas in general a bigger genome size is observed for the streptophytes (from 68 kb for *Chara vulgaris* to around 400 kb for higher plants). The great majority of these genomes are circular, except for some species of Chlamydomonales that have a linear genome (Vahrenholz et al. 1993). The increase of the genome size observed within Streptophyta does not necessarily reflect an increase in coding capacity. Indeed, the transfer of mt genes to the nucleus over evolutionary time (Brennicke et al. 1993), the enlargement and incorporation of new sequences within the mt intergenic spacers, the loss of genes, the increase of intron size, and the resulting decrease of the coding density are all characteristic for the mt genomes of higher land plants. In angiosperms, the most striking feature is the presence of a multipartite genome structure, which results in high-frequency recombination via repeated sequences in the genome (Fauron et al. 1995), altering the genome copy

number, which can result in different phenotypes (Kanazawa et al. 1994; Janska et al. 1998).

All cp (chloroplast) genomes that have been described for land plants have a very conserved genome size, usually around 150 kb covering about 70–80 genes. In contrast, the cp genomes of green algae, although having a rather similar genome size between 150 and 200 kb, show a tremendous variation in gene content, due to massive gene loss, genome erosion, and gene transfer to the nucleus (Grzebyk and Schofield 2003). All cp genomes described so far are circular. Previous studies have shown that, although in green algae (e.g., *C. reinhardtii*) more genes have been transferred to the nucleus compared with land plants (e.g., tobacco), the rate of gene flow has subsequently slowed down dramatically and the transfer of DNA from cp to the nucleus is now very rare (Lister et al. 2003). However, until very recently (Derelle et al. 2006; this study), there was no chlorophyte that had both its nuclear, cp, and mt genome published, and it therefore remained difficult to quantify precisely the extent of gene transfer from the organelles to the nucleus.

*Ostreococcus tauri* is a unicellular green alga that was discovered in the Mediterranean Thau lagoon (France) in 1994. With a size less than 1  $\mu\text{m}$ , comparable to that of a bacterium, it is the smallest eukaryotic organism currently described (Courties et al. 1994). Its cellular organization is rather simple with a relatively large nucleus with only 1 nuclear pore, a single chloroplast 1 mitochondrion, 1 Golgi body, and a highly reduced cytoplasm compartment (Chrétiennot-Dinet et al. 1995). A membrane surrounds the cells, but no cell wall can be observed. Apart from this simple cellular structure, the *O. tauri* nuclear genome is small (12.56 Mb) and is fragmented into 20 chromosomes (Derelle et al. 2006). Phylogenetically, *O. tauri* belongs to the Prasinophyceae, an early branch of the Chlorophyta (Courties et al. 1998). The presence of only 1 chloroplast and 1 mitochondrion and its basal position in the green lineage makes this alga interesting for studying the structure and evolution of both genomes, whereas comparison with other members of the green lineage sheds light on the evolution of organelle genomes.

Key words: chloroplast genome, mitochondrial genome, Chlorophyta, *Ostreococcus tauri*.

E-mail: yves.vandeppeer@psb.ugent.be.

Mol. Biol. Evol. 24(4):956–968. 2007

doi:10.1093/molbev/msm012

Advance Access publication January 23, 2007

## Materials and Methods

### Sequencing

For the sequencing of the nuclear genome, cellular DNA was used for the preparation of the shotgun libraries (Derelle et al. 2006). Consequently, mt and cp sequences were also obtained and identified by their high similarity with genes of other green algae or green plants. Purified DNA was broken by sonication, and after filling ends, DNA fragments ranging from 1 kb to 5 kb were separated in an agarose gel. Blunt-end fragments were inserted into pBluescript II KS (Stratagene, The Netherlands), digested with *EcoRV*, and dephosphorylated. Plasmid DNA from recombinant *Escherichia coli* strains was extracted according to the TempliPhi method (Amersham, GE Healthcare, France), and inserts were sequenced on both strands using universal forward and reverse M13 primers and the ET DYEnamic terminator kit (Amersham). Sequences were obtained with MegaBace 1000 automated sequencers (Amersham). Data were analyzed and contigs were assembled using Phred-Phrap (Ewing et al. 1998) and Consed software packages (<http://bozeman.mbt.washington.edu/consed/consed.html>). Gaps were filled through primer-directed sequencing using custom made primers.

### Gene Prediction and Annotation

All genes were annotated based on their similarity with cp and mt genes that were available in public databases and if necessary manually corrected using Artemis (Rutherford et al. 2000). Homologous relationships between publicly available genes and the *O. tauri* genes were identified through Blast (Altschul et al. 1990). Also small and large ribosomal subunit RNA genes were identified by Blast. Alignment and secondary structure annotation was done using the DCSE alignment editor (De Rijk and De Wachter 1993). The secondary structure drawings were made using RnaViz (De Rijk et al. 2003). tRNA genes were identified by tRNAscan-SE (Lowe and Eddy 1997) using the option "search for organellar tRNAs (-O)". The 5S rRNA gene of the cp genome was identified using the CMSEARCH program from the INFERNAL package (Eddy 2002) with the 5S rRNA covariance model (RF00001) from the RFAM database (Griffiths-Jones et al. 2005).

### Sequence Analyses

Pairwise comparison of gene permutations by inversions between different mt and cp genomes was obtained using the GRIMM web server (Tesler 2002). The data sets used contained, respectively, 54 conserved mt and 82 conserved cp genes. As this tool cannot deal with duplicated genes, genes located in the inverted repeats (IRs) were counted only once.

Duplicated sequences within both genomes were identified using DOTTER (Sonnhammer and Durbin 1995). For both genomes (but including only one of the IR sequences), short repeated sequences were identified with REPUTER 3.1 (Kurtz et al. 2001), using the -p (palindromic), -f (forward), -l (minimum length), and -allmax parameters; and MUMMER 3.0 (Kurtz et al. 2004), using the -l (minimum length) and -b (forward and reverse complement matches)

options. PIPMAKER (Schwartz et al. 2000) was used to visualize the location of the repeated sequences.

### Phylogenetic Analysis

Homologous genes of *O. tauri* cp and mt genes were searched for in the public databases (GenBank/EMBL/DBJ) (Benson et al. 2002; Stoesser et al. 2002; Tateno et al. 2002) using BlastP (Altschul et al. 1997). Protein sequences were aligned with ClustalW (Thompson et al. 1994). Two different data sets were built:

1. Forty-seven cp protein sequences (*atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *clpP*, *petB*, *petG*, *psaA*, *psaB*, *psaC*, *psaJ*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbN*, *psbT*, *psbZ*, *rbcl*, *rpl14*, *rpl16*, *rpl2*, *rpl20*, *rpl36*, *rpoA*, *rpoB*, *rpoC1*, *rpoC2*, *rps11*, *rps12*, *rps14*, *rps18*, *rps19*, *rps3*, *rps4*, *rps7*, *rps8*, *ycf3*, and *ycf4*) from 14 different organisms (*Chlorella vulgaris* [Wakasugi et al. 1997] [AB001684], *Nephroselmis olivacea* [Turmel et al. 1999a] [AF137379], *Pseudoclonium akinetum* [Pombert et al. 2005] [AY835431], *Stigeoclonium helveticum* [Bélanger et al. 2006] [DQ630521], *Scenedesmus obliquus* [de Cambiaire et al. 2006] [DQ396875], *Oltmannsiellopsis viridis* [Pombert et al. 2006a] [DQ291132], *C. reinhardtii* [Maul et al. 2002] [BK000554], *Mesostigma viride* [Lemieux et al. 2000] [AF166114], *C. globosum* [Turmel et al. 2002a] [AF494278], *Marchantia polymorpha* [Ohyama et al. 1986] [M68929], *Nicotiana tabacum* [Shinozaki et al. 1986] [Z00044], *Pinus thunbergii* [Wakasugi et al. 1994] [D17510], *Cyanophora paradoxa* [Stirewalt et al. 1995] [U30821], and *O. tauri*) were independently aligned and concatenated into a data set of 9,553 amino acids.
2. A nuclear gene (small subunit [SSU] rRNA), 1 mt gene (*nad5*), and 2 cp genes (*rbcl* and *atpB*), encompassing 44 organisms, were combined into a data set of 5,053 nucleotides (based on Karol et al. 2001)

PHYML 2.4.4 (Guindon and Gascuel 2003) was used to compute maximum likelihood trees, using the cpREV45 model for cp sequences and the Hasegawa, Kishino and Yano (1985) model for the combined nucleic acid data set. Pairwise distance trees were obtained using TREE-CON (Van de Peer and De Wachter 1994), based on Poisson (Zuckerkindl and Pauling 1965; Dickerson 1971) and Kimura (1983) corrected distances for the protein alignment and Jukes and Cantor (1969) corrections for nucleic acid sequences. PHYLIP (the Phylogeny Inference Package; Felsenstein 1989) was used for 1) computing pairwise distance trees using the Dayhoff PAM matrix (1979) for protein alignment and Jukes and Cantor (1969) for nucleic acid sequences and 2) obtaining maximum parsimony trees for both data sets. For each method, bootstrap analyses with 500 replicates were performed to test the significance of the nodes. Finally, MrBayes (500,000 generations and 4 chains) was used for Bayesian inference of phylogenetic trees (Huelsenbeck et al. 2001), using a JTT +  $\gamma$  substitution model (Jones et al. 1992).

After manual improvement of the alignments using BIOEDIT (Hall 1999), only unambiguously aligned

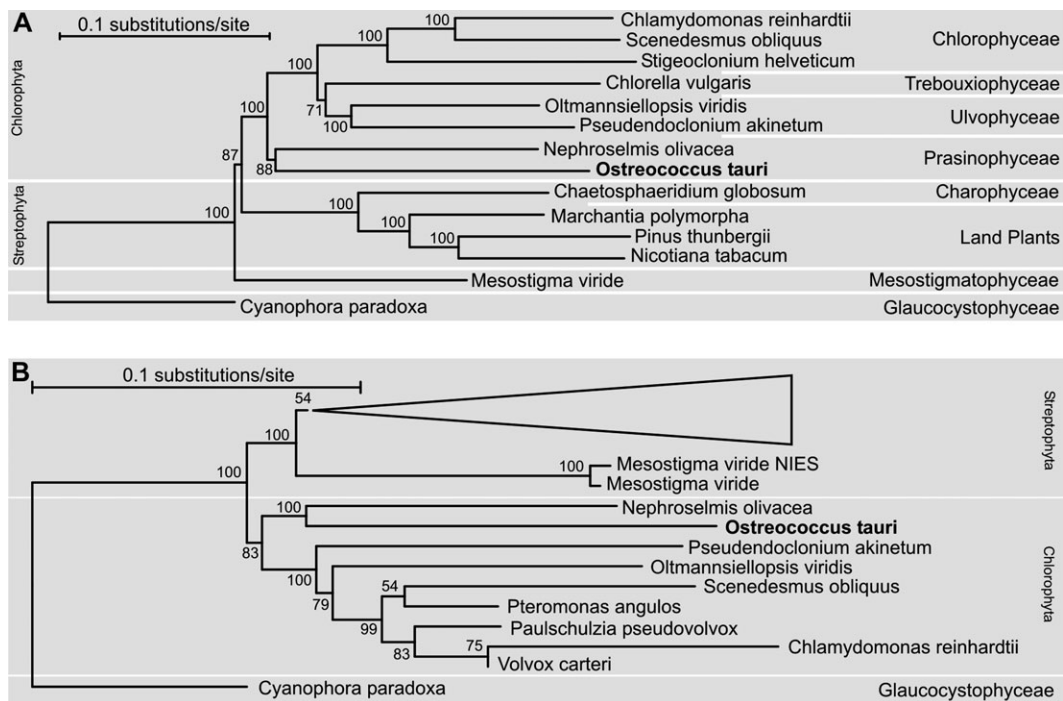


FIG. 1.—Phylogenetic position of *Ostreococcus tauri*. (A) Tree was made based on 47 concatenated sequences of cp genes, using the pairwise distance method. As each method gave the same topology, the lowest bootstrap values are shown at the branches of the nodes. (B) Tree inferred by pairwise distance methods of a combined data set of a mt, nuclear, and 2 cp genes, based on the Karol et al. (2001) data set. All members of the Streptophyta except *Mesostigma viride* are replaced by a triangle. The complete tree can be found in supplementary fig. S1 (Supplementary Material online). Lowest bootstrap values, using different methods, are shown at the nodes. For both trees, the branch lengths are drawn to scale.

positions were taken into account for tree construction. TREEVIEW was used to visualize the trees (Page 1996).

## Results and Discussion

### Phylogenetic Analyses

Previous phylogenetic analyses based on the 18S rDNA sequence of different Chlorophyta suggested that *O. tauri* belongs to the Prasinophyceae, an early diverging group within the green plant lineage (Courties et al. 1998). Now, with the availability of the cp, mt, and even nuclear (Derelle et al. 2006) genomes of *O. tauri*, a more extensive phylogenetic analysis could be performed. To this end, we prepared 2 different data sets, that is, 1 consisting of concatenated cp genes and 1 consisting of a mix of concatenated cp, mt, and nuclear genes (see Materials and Methods). Using these different data sets and different methods of phylogenetic tree construction, *O. tauri* always clustered with other members of the Chlorophyta, clearly confirming its Chlorophytean heritage (fig. 1; see supplementary fig. S1, Supplementary Material online). Furthermore, the different classes within the Chlorophyta, namely, Chlorophyceae, Trebouxiophyceae, Ulvophyceae, and Prasinophyceae formed monophyletic groups, well supported by bootstrap analyses. *O. tauri* and *N. olivacea* were always grouped together within an early diverging group referred to as the Prasinophyceae, thereby confirming the previous analyses done by Courties (1998).

In our phylogenetic analyses, we have also included the unicellular freshwater alga, *M. viride*, whose phylogenetic position is still being discussed (previously referred to

as the “enigma of *Mesostigma*” [McCourt et al. 2004]). After being classified as an primitive chlorophyte (Mattox and Stewart 1984; Grzebyk and Schofield 2003; Nozaki et al. 2003), a charophyte (Melkonian 1989; Bhattacharya et al. 1998; Karol et al. 2001; Martin et al. 2002), or as a species branching off prior to the divergence of the Streptophyta and Chlorophyta (Lemieux et al. 2000; Turmel et al. 2002b), Petersen et al. (2006) provided unequivocal support for its Streptophytean affiliation, based on the presence of a land plant–specific *gapB* gene and the absence of this gene in the different orders of chlorophyte green algae. However, our tree based on concatenated cp genes (fig. 1a) clearly support *M. viride* branching off before the divergence of Chlorophyta and Streptophyta. On the other hand, trees based on a combination of concatenated plastid, mt, and nuclear genes did group *M. viride* with the other streptophytes (fig. 1b, see supplementary fig. S1, Supplementary Material online). In addition, we recently showed (Robbins et al. 2007) that 2 *Ostreococcus* strains (*O. tauri* and *Ostreococcus lucimarinus*) also contain the *gapB* gene (DQ649078 and DQ649079), making this gene no longer land plant specific as postulated by Petersen et al. (2006). As a matter of fact, all this adds some more mystery to the phylogenetic position of *M. viride*.

### Structure and Gene Content of the mt Genome

The *O. tauri* mt genome assembles as a circle of 44,237 bp (fig. 2), with an overall GC content of 38%. This size is similar to the mt genome of another early branching chlorophyte *N. olivacea* (45,223 bp) (Turmel et al. 1999b).

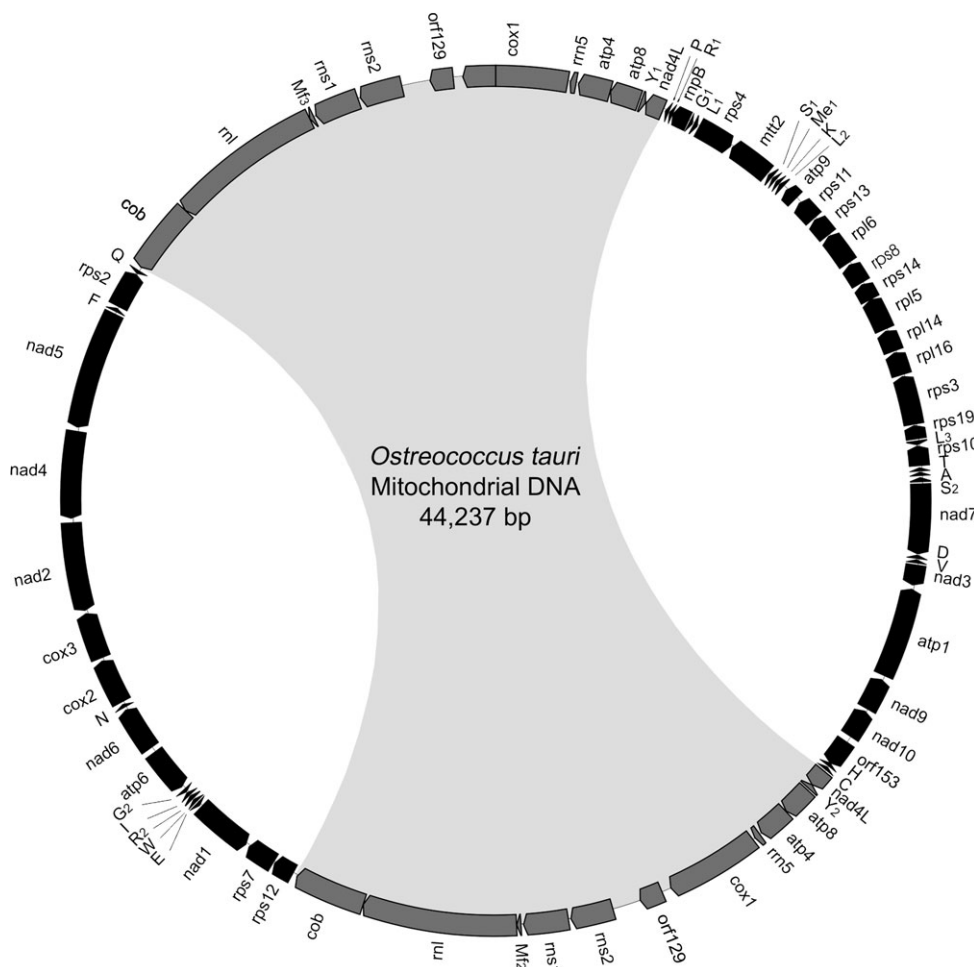


FIG. 2.—*Ostreococcus tauri* mt genome. Genes located in the unique duplicated region are colored in gray; single-copy genes are in black. The length of the boxes is proportional to their amino acid length. tRNA genes are represented by the 1-letter amino acid code, and the unique ORFs are indicated by orf followed by their amino acid length.

However, in contrast to the *N. olivacea* genome, the *O. tauri* mt sequence contains a duplicated region, containing 22 genes and covering 44% (19,542 bp) of the genome (see further). Sixty-five genes (unique open reading frames [ORFs] were not taken into account, and duplicated genes were counted only once) are encoded on both strands, encompassing 93% of the genome, which makes the mt genome of *O. tauri* the most gene dense among the

Chlorophyta. For comparison, both *M. viride* (Turmel et al. 2002b) and *N. olivacea* also have sixty-five genes, but only covering 87% and 81% of their genome, respectively (table 1). Among the 65 genes, 36 are protein-encoding genes, 26 are transfer RNAs, and 3 are rRNAs (see supplementary table S1, Supplementary Material online). Two predicted proteins (*orf129* and *orf153*) coding for 129 and 153 amino acids, respectively, did not show any

**Table 1**  
General Features of mt Genomes of Different Green Organisms

Feature	Ot	No	Cr	Mv	Cg	Mp	At
Size (bp)	44,237	45,223	15,758	42,424	56,574	186,609	366,924
GC (%)	38.0	32.8	45.2	32.2	34.4	42.2	44.8
Gene number <sup>a</sup>	65	65	13	65	67	69	50
Gene density (%)	93.0	80.6	83.1	86.6	76.3	65.0	45.5
Intron number	—	4 (I)	—	4 (I), 3 (II)	9 (I), 2 (II)	7 (I), 25 (II)	23 (II)
Repeated sequence (bp) <sup>b</sup>	9,771	—	500	—	—	—	11,372
Map	Circular	Circular	Linear	Circular	Circular	Circular	Circular

NOTE.—Ot: *Ostreococcus tauri*, No: *Nephroselmis olivacea*, Cr: *Chlamydomonas reinhardtii*, Mv: *Mesostigma viride*, Cg: *Chaetosphaeridium globosum*, Mp: *Marchantia polymorpha*, At: *Arabidopsis thaliana*. “—”, Not present.

<sup>a</sup> Unique ORFs were not taken into account and duplicated genes were counted only once.

<sup>b</sup> Minimum 500 bp long and 100% identical.

clear similarity to other known genes. The compactness of the *O. tauri* mt genome is further illustrated by the shortness of the intergenic regions, ranging from 1 to 475 bp, with an average of 42 bp. Only 5 intergenic regions exceed 100 bp, and these are all located in the duplicated region. In addition, there are 3 cases of overlapping genes (*trnR1-rnpB*, *rps14-rpl5*, and *orf153-trnH*). Lastly, in contrast to other members of the green lineage, neither group I nor group II type introns are present in any of the genes.

All 26 tRNAs fold into the conventional cloverleaf secondary structure and are able to decode all codons. The small subunit rRNA (SSU rRNA, *rns* in fig. 2) gene is fragmented into 2 parts, but retains its ability to fold into the normal secondary structure model (see supplementary fig. S2, Supplementary Material online). The fragmentation site is located near the hairpin loop of helix 29 (indicated by gray area) of the secondary structure model (Wuyts et al. 2004), and the location of the 2 fragments has been rearranged in the genome such that both fragments are located on the forward strand but their order is reversed. However, this fragmentation site does not correspond to one of the several fragmentation sites that have been previously identified in the small subunit rRNA genes of chlorophyte mt genomes (Nedelcu et al. 2000) (other known fragmentation sites are indicated by gray areas on supplementary fig. S2, Supplementary Material online). The SSU rRNA, LSU rRNA (large subunit rRNA, *rnl* in fig. 2), and 5S rRNA gene (*rrn5* in fig. 2) are all located in the duplicated region. Like other known members of the green lineage, the *O. tauri* mitochondrion uses the standard genetic code, and all the 61 codons are used. As in *M. viride*, there is a strong bias in favor of codons that end in A or U. The 3 types of stop codons are present with UAA being the one most often used (81.4%) (see supplementary table S2, Supplementary Material online).

The most striking feature of the *O. tauri* mt genome is the presence of a large duplicated segment (19,542 bp; shaded box in fig. 2). This duplication is also observed in the partially sequenced mt genome of another *Ostreococcus* strain (*O. lucimarinus*; Palenik B, personal communication), thereby excluding erroneous genome assembly. The presence of such a duplicated sequence has not been observed in any other member of the Chlorophyta, except for *C. reinhardtii*, wherein its mt genome, which is linear instead of circular, terminal IRs of approximately 500 bp have been described (Vahrenholz et al. 1993; table 1). No duplication is present in the mt genome of the charophyte *Chara vulgaris* (Turmel et al. 2003). The only large repeated sequences previously reported are present in higher land plants (e.g., *Arabidopsis thaliana*: 366,924 bp, containing repeat sequences of 6.5 and 4.5 kb and *Beta vulgaris*: 368,799 bp, containing a repeat sequence of 6.2 kb) (Unselde et al. 1997; Kubo et al. 2000). These repeated regions in the mt genome of angiosperms gave rise to a multipartite genome structure (Fauron et al. 1995) and lead to high-frequency intramolecular recombination. Indeed, a master circle, containing the complete genetic information, can lead to different subgenomic circles by homologous recombination via a repeated sequence motif (e.g., the tobacco mt genome can provide 6 different subgenomic circles by homologous recombination between the different repeated sequences) (Knoop 2004; Sugiyama et al.

2005). The presence of this multipartite genome structure enables them to change their gene and genome copy number, resulting in an altered plant phenotype (Kanazawa et al. 1994; Janska et al. 1998).

The unique repeated segment in the mt genome of *O. tauri* contains 5 protein-coding genes (*cob*, *cox1*, *atp4*, *atp8*, and *nad4L*), 2 tRNAs (*trnMe* and *trnY*), the SSU rRNA (*rns*), LSU rRNA (*rnl*), and 5S rRNA (*rrn5*) genes and the *orf129*. The duplicated nucleotide sequences are 100% identical over a length of 9,771 bp, covering 44% of the genome (see supplementary fig. S3, Supplementary Material online). Repeats that are a 100% identical in mt genomes are not exceptional. For instance, *Brassica napus* has 2 repeats with 1 mismatch over 2,427 bp (Handa 2003), *B. vulgaris* has 2 repeats (Kubo et al. 2000), and *A. thaliana* has 3 repeats that are 100% identical (Unselde et al. 1997).

Short dispersed repeats (SDR) are also thought to play an important role in mt genome rearrangements, thereby altering the gene content and genome size. This is not only true for members of the Chlorophyta, but also in land plants, yeasts, and even animals, where they serve as hot spots for recombination (Pombert et al. 2004). Short dispersed repeats have been described in all known members of the Chlorophyta, although their abundance is highly variable. All Chlorophyta members hold SDRs of at least 15 bp in their genome. This number is reduced to 52 repeats in *N. olivacea* (Pombert et al. 2006b) and to only 11 in *O. tauri*. The largest repeats found in *O. tauri* and *N. olivacea* are rather short, being 34 bp and 42 bp, respectively. The GC content of the SDRs present in *O. tauri* does not differ much from the overall GC content of the mt genome (36% for the SDRs vs. 38% for the complete genome). In general, more derived lineages show an increase of the number of SDRs: *O. viridis* contains 1,206 (Pombert et al. 2006b), *Scenedesmus obliquus* 4,086 (Nedelcu et al. 2000), and *P. akinetum* 8,002 (Pombert et al. 2004) SDRs of at least 15 bp long. It seems that after the split of the Prasinophyceae, an increase of SDRs took place (with the exception of the *Chlamydomonadales*), and it is tempting to correlate this increase with the gene rearrangements that took place within the other members of the Chlorophyta (see further).

### Comparison with Other mt Genomes

Comparison of the *O. tauri* mt genome with 9 other species of the Viridiplantae lineage (Cr: *C. reinhardtii*, No: *N. olivacea*, Ov: *O. viridis*, Pa: *P. akinetum*, So: *S. obliquus*, At: *A. thaliana*, Mp: *M. polymorpha*, Cg: *C. globosum*, and Mv: *M. viride*) unveiled only 9 genes (not including tRNAs), which are common to all these species (table 2). However, when removing *C. reinhardtii* (Michaelis et al. 1990) and *S. obliquus*, 2 members of the Chlorophyceae, from this comparison, this number increases to 25 shared genes. When further removing the 2 ulvophyte green algae (*O. viridis* and *P. akinetum*), the number of conserved genes increases to 30, thus, representing the gene content conservation between the 2 prasinophytes and the land plants. However, when only considering the protein-coding genes of *O. tauri*, *N. olivacea*, and *M. viride*, 36 genes are shared, which represents 95% of the *O. tauri* and 92% of the *M. viride* protein-coding gene content. Apparently, the

**Table 2**  
**Comparison of Gene Content in Green Algal and Land Plant mt Genomes**

	Ov	Pa	So	Cr	No	Ot	Mv	Cg	At	Mp
NADH dehydrogenase										
<i>nad1</i>	*	*	*	*	*	*	*	*	*	*
<i>nad2</i>	*	*	*	*	*	*	*	*	*	*
<i>nad3</i>	*	*	*	*	*	*	*	*	*	*
<i>nad4</i>	*	*	*	*	*	*	*	*	*	*
<i>nad4L</i>	*	*	*	*	*	*	*	*	*	*
<i>nad5</i>	*	*	*	*	*	*	*	*	*	*
<i>nad6</i>	*	*	*	*	*	*	*	*	*	*
<i>nad7</i>	*	*			*	*	*	*	*	*+
<i>nad9</i>	*				*	*	*	*	*	*
<i>nad10</i>					*	*				
Cob-complex										
<i>cob</i>	*	*	*	*	*	*	*	*	*	*
Cytochrome c oxidase										
<i>cox1</i>	*	*	*	*	*	*	*	*	*	*
<i>cox2</i>	*	*	*	*	*	*	*	*	*	*
<i>cox3</i>	*	*	*	*	*	*	*	*	*	*
ATP synthase										
<i>atp1</i>	*	*			*	*	*	*	*	*
<i>atp4</i>	*	*			*	*	*	*	*	*
<i>atp6</i>	*	*	*	*	*	*	*	*	*	*
<i>atp8</i>	*	*			*	*	*	*	*	*
<i>atp9</i>	*	*	*	*	*	*	*	*	*	*
Conserved protein										
<i>mtt2</i>	*	*			*	*	*	*	*	*
LSU ribosomal proteins										
<i>rpl2</i>								*	*	*
<i>rpl5</i>		*			*	*	*	*	*	*
<i>rpl6</i>					*	*	*	*	*	*
<i>rpl14</i>					*	*	*	*	*	*
<i>rpl16</i>	*	*			*	*	*	*	*	*
SSU ribosomal proteins										
<i>rps1</i>							*	*		*
<i>rps2</i>	*	*			*	*	*	*	*	*
<i>rps3</i>	*	*			*	*	*	*	*	*
<i>rps4</i>		*			*	*	*	*	*	*
<i>rps7</i>					*	*	*	*	*	*
<i>rps8</i>					*	*				*
<i>rps10</i>		*			*	*	*	*	*	*
<i>rps11</i>	*	*			*	*	*	*	*	*
<i>rps12</i>	*	*			*	*	*	*	*	*
<i>rps13</i>	*	*			*	*	*	*	*	*
<i>rps14</i>	*	*			*	*	*	*	*+	*
<i>rps19</i>	*	*			*	*	*	*	*+	*
Ribosomal RNAs										
<i>rns</i>	*	*	*	*	*	*	*	*	*	*
<i>rnl</i>	*	*	*	*	*	*	*	*	*	*
<i>rrn5</i>	*				*	*	*	*	*	*
Rnase P RNA										
<i>rnpB</i>					*	*				
Transfer RNAs	24	25	27	3	26	26	26	28	17	27

Cr: *Chlamydomonas*, No: *Nephroselmis*, Ot: *Ostreococcus*, Mv: *Mesostigma*, Ov: *Oltmannsiellopsis*, Pa: *Pseudoclonium*, So: *Scenedesmus*, Cg: *Chaetosphaeridium*, Mp: *Marchantia*, At: *Arabidopsis*.

\*+ Pseudogene.

\* present

Bold gene names indicate conserved genes

gene content conservation between these genomes, which are assumed to represent a more ancestral state, is still very high. One of the 7 protein-coding genes that are absent in the *O. tauri* mt genome, namely *rpl2*, could be uncovered in the

nuclear genome (see supplementary table S3, Supplementary Material online).

Disregarding the unique ORFs and tRNAs (*trnG*[gcc] and *trnL*[gag] seem to have been lost in *N. olivacea* compared with *O. tauri*, whereas *trnR*[ucg] is lost in *O. tauri* compared with *N. olivacea*), the gene repertoires of *O. tauri* and *N. olivacea* are identical (table 2). Furthermore, there is a high degree of synteny between these 2 algae, with 5 gene clusters of at least 5 genes and 1 of 2 genes, which are almost identical in both mt genomes (genes denoted in black in fig. 3). However, when one considers gene polarities, synteny is limited to only 2 gene clusters (12 genes extending from *rps11* to *rps10* and 5 genes extending from *atp6* to *cox3*). The major difference between both mt genomes is the duplication in *O. tauri* and the presence of 4 group I introns in *N. olivacea* (3 within the *rnl* and 1 in the *cob* gene) (Turmel et al. 1999b).

A certain degree of synteny can still be detected when adding *C. globosum* (charophyte) (Turmel et al. 2002a) and *Marchantia polymorpha* (streptophyte) (Oda et al. 1992) to the 2 previous species (genes in black and gray in fig 3), except for cluster 3 where no clear synteny could be detected among the 4 organisms and; for cluster 2 (fig. 3), where the genes of *M. polymorpha* are divided into 2 parts ([*atp6*, *nad6*, and *trnN*] and [*cox2* and *cox3*]). In contrast, synteny conservation still exists in cluster 5 where 9 genes are present in the 4 organisms, all oriented in the same direction, indicating that although the genome size increases from green algae to higher land plants, the gene organization of some clusters are extremely well conserved in evolution.

Additionally, we estimated the number of gene inversions needed to transform the gene organization of one genome into another, thereby providing quantitative measurement of their evolutionary distances. Fifty-four conserved genes (duplicated genes were used only once) of 3 Chlorophyta (*O. tauri*, *N. olivacea*, and *P. akinetum*) and *M. viride* were used, showing that a minimum of 29 inversions are needed to transform the gene organization of *O. tauri* into that of *N. olivacea*. When comparing *O. tauri* with the other mt genomes, almost twice as many inversions are needed (50 for both *P. akinetum* and *M. viride*), again indicating the close relationship between the 2 Prasinophyceae.

### Structure and Gene Content of the cp Genome

With a circular cp genome of 71,666 bp long (fig. 4), *O. tauri* contains the smallest cp genome known so far within the Viridiplantae (except for the parasite *Helicosporidium* sp. [de Koning and Keeling 2006]). Cp genome size in green algae ranges from 118,360 bp in *M. viride* (Lemieux et al. 2000) to 203,395 bp in *C. reinhardtii* (Maul et al. 2002; table 3). The GC content (39.9%) of the *O. tauri* cp genome is close to that of *N. olivacea* (42.1%) (Turmel et al. 1999a) and *O. viridis* (40.5%) (Pombert et al. 2006a), but higher than that of other chlorophytes, such as *Chlorella vulgaris* (31.6%) (Wakasugi et al. 1997), *C. reinhardtii* (34.6%), and *M. viride* (30.1%). Like all known members of the Chlorophyta, except *C. vulgaris*, the cp genome of *O. tauri* has a quadripartite structure containing 2 large IRs of 6,825 bp (covering 9.5% of the genome) separating a large

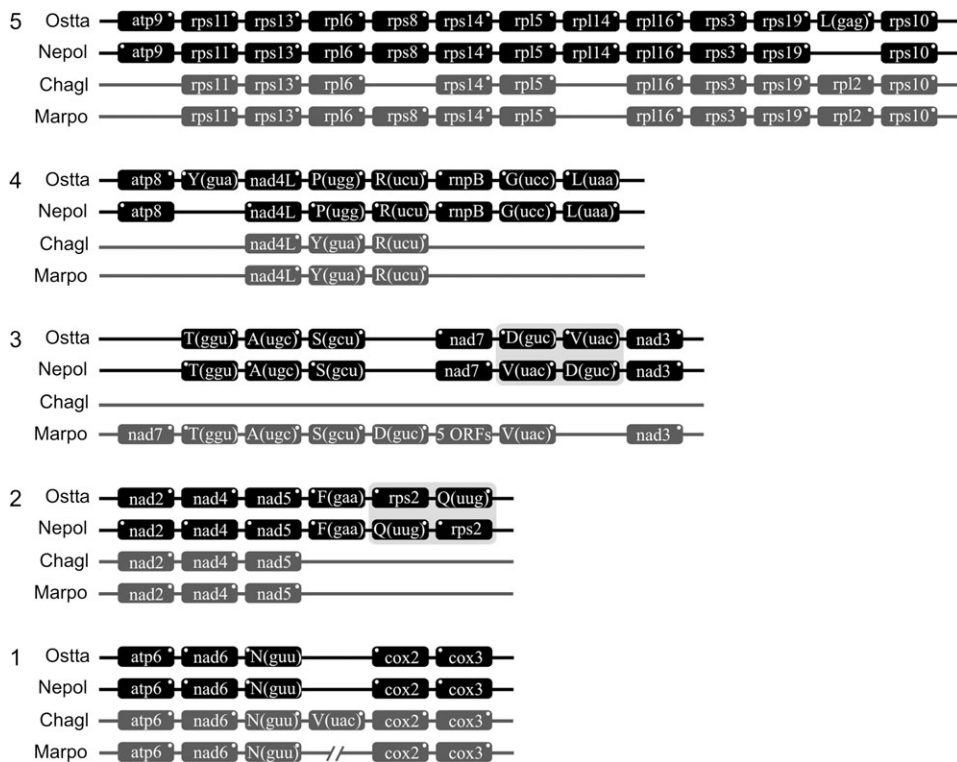


FIG. 3.—Overview of gene clusters conserved between *O. tauri* (Ostta), *Nephroselmis olivacea* (Nepol), *Chaetosphaeridium globosum* (Chagl), and *Marchantia polymorpha* (Marpo). Black boxes represent the genes of *O. tauri* and *N. olivacea*, and gray boxes represent the genes of both Streptophyta. White dots in the upper-left or upper-right corner indicate the orientation of the genes. Gaps represent the loss of genes. tRNA genes are represented by the one-letter amino acid code followed by the anticodon in parentheses. Light gray background boxes indicate inversions of gene order.

single-copy (LSC) region (35,684 bp, covering 49.8%) and a small single-copy (SSC) region (22,332 bp, covering 31.2%) (fig. 4 and supplementary fig S4, Supplementary Material online). Despite the difference in size, both the LSC and the SSC contain 41 genes, whereas the IR sequences contain, next to *psbA*, the rRNA operon (*rrs*, *trnI*[*gau*], *trnA*[*ugc*], *rrl*, and *rrf*).

Besides its ultrasmall cp genome, the gene content is reduced to a minimum: 86 genes (unique ORFs were not taken into account, and duplicated genes were counted only once) were identified, including 25 tRNAs and the rRNA gene cluster (*rrf*, *rrl*, and *rrs*). Two predicted proteins (*orf537* and *orf1260*) coding for 537 and 1,260 amino acids, respectively, show little similarity with known genes: *ycf1* and *ycf2*. These genes will be indicated as *orf537/ycf1* and *orf1260/ycf2*. (see supplementary table S4, Supplementary Material online) This gene repertoire is the smallest known to date among the green algae: *C. reinhardtii* has a slightly higher number of genes (94 genes, not including the duplicated genes and unique ORFs), but this number is also much lower than the number of genes present in other Chlorophyta (e.g., *N. olivacea* contains 127 genes and *M. viride* contains 135 genes) (table 3). Twenty-five tRNAs could be detected, a number that is low compared with that of other members of the green lineage (e.g., *N. olivacea*: 32, *M. viride*: 37, and *A. thaliana*: 37 [Sato et al. 1999]), but enables the *O. tauri* cp genome to decode all 61 codons. Two of the 3 types of stop codons are present with UAA being the one most often used (95%) and with

UGA being absent (see supplementary table S5, Supplementary Material online). The average length of intergenic regions is 116 bp, varying from 1 to 476 bp. There are 3 cases of overlapping genes (*psbC*–*psbD*, *rpoC1*–*rpoC2*, and *rps3*–*rpl16*), resulting in an average coding density (including conserved genes, unique ORFs, and introns) of 84.7%. One group II intron, present in *atpB*, could be detected.

#### Comparison with Other cp Genomes

The gene repertoire of the cp genomes of 7 Chlorophyta (Cr: *C. reinhardtii*, Cv: *C. vulgaris*, No: *N. olivacea*, Ov: *O. viridis*, Pa: *P. akinetum*, So: *S. obliquus*, and Ot: *O. tauri*), 2 Streptophyta (At: *A. thaliana* and Nt: *N. tabacum*), and *M. viride* (Mv) were compared and the results shown in table 4. Fifty-three core genes are shared between both Chlorophyta and Streptophyta (bold gene names), whereas 4 additional core genes (*ycf12*, *tufA*, *rpl5*, and *rps9*) are present when only considering the Chlorophyta lineage. The 53 core cp genes are involved either in photosynthesis, energy metabolism, or some housekeeping functions. Gene loss and gene transfer to the nucleus is a common feature of cp genomes (Stegemann et al. (2003)), and (Grzebyk and Schofield 2003) reported the loss of 7 genes (*rpl21*, *rpl22*, *rpl33*, *rps15*, *rps16*, *odpB*, and *ndhJ*) at the base of the Chlorophyta lineage. These genes were also not detected in the *O. tauri* cp genome, but 5 of them are present in the nuclear genome (see supplementary table S3, Supplementary Material online).

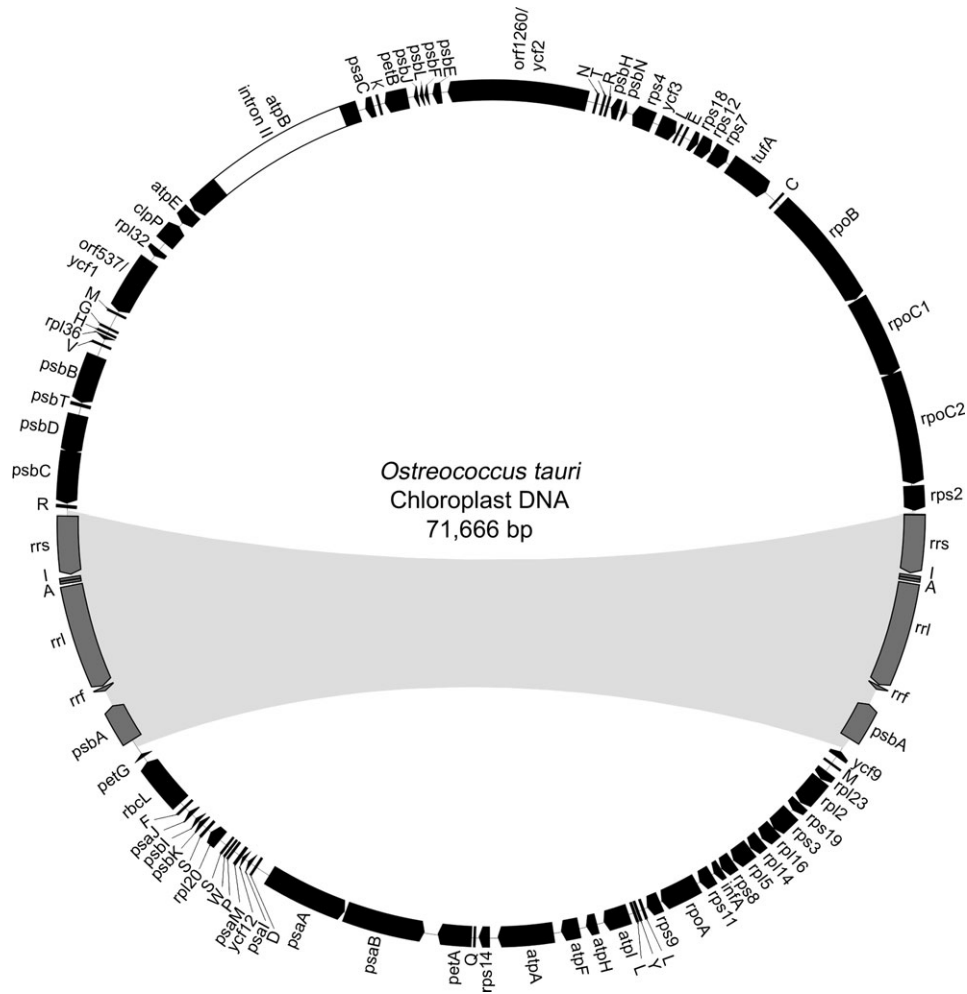


FIG. 4.—*Ostreococcus tauri* plastid genome. Genes located in the inverted repeat sequences are colored in gray; genes in the single copy regions are black. The single intron, located in *atpB*, is shown as a white box. The length of the boxes is proportional to their amino acid length. tRNA genes are represented by the 1-letter amino acid code and the unique ORFs are indicated by orf followed by their amino acid length.

In *O. tauri*, 34 genes are lost in the cp genome compared with other Chlorophyta: 1) the 10 homologs of the *ndh* genes, subunits of the NADH:ubiquinone oxidoreductase. None of these genes were present in the nuclear genome; 2) the genes *chlB*, *chlI*, *chlL*, and *chlN* involved in the chlorophyll synthesis in dark. In almost all known green algal cp genomes, these 4 genes are present, but not in *O. tauri* where only *chlI* was found in the nuclear

genome (on chromosome 2). The absence of *chlB*, *chlL*, and *chlN* in the cp or nuclear genome of *O. tauri* confirms the inability of this organism to produce chlorophyll in dark (Derelle et al. 2006); 3) both the *petL* and *petD* genes are absent in *O. tauri*, whereas they are present in all other studied organisms where they encode a small subunit of the cytochrome b6f complex. The *petL* has not been transferred to the nucleus, whereas *petD* could be located on chromosome 7.

**Table 3**  
General Features of cp Genomes of Different Green Organisms

Feature	Ot	No	Cr	Mv	Cg	Mp	At
Size (bp)	71,666	200,799	203,827	118,360	131,183	121,024	154,478
GC (%)	39.9	42.1	34.5	30.1	29.6	28.2	36.3
Gene number <sup>a</sup>	86	128	94	136	125	120	87
Gene density (%)	84.7	68.7	50.1	?	?	?	?
Intron number	1 (II)	—	5 (I), 2 (II)	—	1 (I), 17 (II)	1 (I), 19 (II)	1 (I), 20 (II)
IR (bp)	6,825	46,137	22,211	6,057	12,430	10,058	26,264
SSC (bp)	22,332	16,399	78,088	22,619	17,640	19,813	17,780
LSC (bp)	35,684	92,126	81,307	83,627	88,683	81,095	84,170

NOTE.—Ot: *O. tauri*, No: *N. olivacea*, Cr: *C. reinhardtii*, Mv: *M. viride*, Cg: *C. globosum*, Mp: *M. polymorpha*, At: *A. thaliana*. “—”, Not present. “?”, Not known.

<sup>a</sup> Unique ORFs were not taken into account and duplicated genes were counted only once.



**Table 4**  
**Comparison of Gene Content in Green Algal and Land Plant cp Genomes**

	Ov	Pa	So	Cr	Cv	No	Ot	Mv	At	Nt
<b>Photosystem I</b>										
<i>psaA</i>	*	*	*	*	*	*	*	*	*	*
<i>psaB</i>	*	*	*	*	*	*	*	*	*	*
<i>psaC</i>	*	*	*	*	*	*	*	*	*	*
<i>psaI</i>	*	*	*	*	*	*	*	*	*	*
<i>psaJ</i>	*	*	*	*	*	*	*	*	*	*
<i>psaM</i>	*	*		*	*	*	*	*	*	*
<b>Photosystem II</b>										
<i>psbA</i>	*	*	*	*	*	*	*	*	*	*
<i>psbB</i>	*	*	*	*	*	*	*	*	*	*
<i>psbC</i>	*	*	*	*	*	*	*	*	*	*
<i>psbD</i>	*	*	*	*	*	*	*	*	*	*
<i>psbE</i>	*	*	*	*	*	*	*	*	*	*
<i>psbF</i>	*	*	*	*	*	*	*	*	*	*
<i>psbG</i>										
<i>psbH</i>	*	*	*	*	*	*	*	*	*	*
<i>psbI</i>	*	*	*	*	*	*	*	*	*	*
<i>psbJ</i>	*	*	*	*	*	*	*	*	*	*
<i>psbK</i>	*	*	*	*	*	*	*	*	*	*
<i>psbL</i>	*	*	*	*	*	*	*	*	*	*
<i>psbM</i>	*	*	*	*	*	*	*	*	*	*
<i>psbN</i>	*	*	*	*	*	*	*	*	*	*
<i>psbT</i>	*	*	*	*	*	*	*	*	*	*
<i>psbZ (ycf9)</i>	*	*	*	*	*	*	*	*	*	*
<b>Cytochrome b6/F</b>										
<i>petA</i>	*	*	*	*	*	*	*	*	*	*
<i>petB</i>	*	*	*	*	*	*	*	*	*	*
<i>petD</i>	*	*	*	*	*	*	*	*	*	*
<i>petG</i>	*	*	*	*	*	*	*	*	*	*
<i>petL</i>	*	*	*	*	*	*	*	*	*	*
<i>petN</i>						*	*	*	*	*
<b>ATP synthase</b>										
<i>atpA</i>	*	*	*	*	*	*	*	*	*	*
<i>atpB</i>	*	*	*	*	*	*	*	*	*	*
<i>atpE</i>	*	*	*	*	*	*	*	*	*	*
<i>atpF</i>	*	*	*	*	*	*	*	*	*	*
<i>atpH</i>	*	*	*	*	*	*	*	*	*	*
<i>atpI</i>	*	*	*	*	*	*	*	*	*	*
<b>Chlorophyll biosynthesis</b>										
<i>chlB</i>	*		*	*	*	*		*		*
<i>chlI</i>	*	*		*	*	*		*		*
<i>chlL</i>	*		*	*	*	*		*		*
<i>chlN</i>	*		*	*	*	*		*		*
<b>Rubisco</b>										
<i>rbcL</i>	*	*	*	*	*	*	*	*	*	*
<b>RNA polymerase</b>										
<i>rpoA</i>	*	*	*	*	*	*	*	*	*	*
<i>rpoB</i>	*	*	*	*	*	*	*	*	*	*
<i>rpoC1a</i>				*						*
<i>rpoC1b</i>				*						*
<i>rpoC1</i>	*	*	*		*	*	*	*	*	*
<i>rpoC2</i>	*	*	*	*	*	*	*	*	*	*
<b>NADH oxidoreductase</b>										
<i>ndhA-I</i>						*		*	*	*
<i>ndhJ</i>								*	*	*
<i>ndhK</i>						*		*	*	*
<b>Ribosomal RNAs</b>										
<i>23S</i>	*	*	*	*	*	*	*	*	*	*
<i>16S</i>	*	*	*	*	*	*	*	*	*	*
<i>5S</i>	*	*	*	*	*	*	*	*	*	*
<i>tRNAs</i>	25	28	27	31	33	32	27	37	37	30
<i>rnpB</i>						*				

**Table 4**  
**Continued**

	Ov	Pa	So	Cr	Cv	No	Ot	Mv	At	Nt
<b>LSU ribosomal proteins</b>										
<i>rpl2</i>	*	*	*	*	*	*	*	*	*	*
<i>rpl5</i>	*	*	*	*	*	*	*	*	*	*
<i>rpl12</i>	*	*	*	*	*	*	*	*	*	*
<i>rpl14</i>	*	*	*	*	*	*	*	*	*	*
<i>rpl16</i>	*	*	*	*	*	*	*	*	*	*
<i>rpl19</i>	*	*		*	*	*	*	*	*	*
<i>rpl20</i>	*	*	*	*	*	*	*	*	*	*
<i>rpl22</i>								*	*	*
<i>rpl23</i>	*	*	*	*	*	*	*	*	*	*
<i>rpl32</i>	*	*			*	*	*	*	*	*
<i>rpl33</i>								*	*	*
<i>rpl36</i>	*	*	*	*	*	*	*	*	*	*
<b>SSU ribosomal proteins</b>										
<i>rps2</i>	*	*	*	*	*	*	*	*	*	*
<i>rps3</i>	*	*	*	*	*	*	*	*	*	*
<i>rps4</i>	*	*	*	*	*	*	*	*	*	*
<i>rps7</i>	*	*	*	*	*	*	*	*	*	*
<i>rps8</i>	*	*	*	*	*	*	*	*	*	*
<i>rps9</i>	*	*	*	*	*	*	*	*	*	*
<i>rps11</i>	*	*	*		*	*	*	*	*	*
<i>rps12</i>	*	*	*	*	*	*	*	*	*	*
<i>rps14</i>	*	*	*	*	*	*	*	*	*	*
<i>rps15</i>								*	*	*
<i>rps16</i>								*	*	*
<i>rps18</i>	*	*	*	*	*	*	*	*	*	*
<i>rps19</i>	*	*	*	*	*	*	*	*	*	*
<b>Translation factors</b>										
<i>infA</i>	*	*	*		*	*	*	*	*	*
<i>tufA</i>	*	*	*	*	*	*	*	*	*	*
<b>Division</b>										
<i>ftsI</i>							*		*	
<i>ftsW</i>							*		*	
<i>minD</i>	*	*				*	*	*	*	
<i>minE</i>					*					
<b>Miscellaneous proteins</b>										
<i>accD</i>	*	*			*	*			*	*
<i>cemA</i>	*	*	*	*	*	*		*	*	*
<i>cIpp</i>	*	*	*	*	*	*	*	*	*	*
<i>ccsA</i>	*	*	*	*	*	*	*	*	*	*
<i>cysA</i>					*	*		*	*	*
<i>cysT</i>					*	*		*	*	*
<i>I-Cvul</i>					*					
<b>Conserved proteins</b>										
<i>ycf1</i>	*	*	*		*	*	*	*	*	*
<i>ycf2</i>					*	*	*	*	*	*
<i>ycf3</i>	*	*	*	*	*	*	*	*	*	*
<i>ycf4</i>	*	*	*	*	*	*	*	*	*	*
<i>ycf5</i>					*				*	*
<i>ycf6/PetN</i>						*		*	*	*
<i>ycf9/PsbZ</i>	*	*	*	*	*	*	*	*	*	*
<i>ycf10/CemA</i>					*				*	*
<i>ycf12</i>	*	*	*	*	*	*	*	*	*	*

Cr: *Chlamydomonas*, Cv: *Chlorella*, No: *Nephroselmis*, Ot: *Ostreococcus*, Ov: *Oltmannsiellopsis*, So: *Scenedesmus*, Pa: *Pseudoclonium*, Mv: *Mesostigma*, At: *Arabidopsis*, Nt: *Nicotiana*. Shaded boxes indicate specific absence of genes in *Ostreococcus tauri*. Squared boxes indicate conserved genes among all organisms. \* present

However, it has been shown in *C. reinhardtii* that a free *petL* N-terminus is not required for the b6f complex function (Zito et al. 2002); 4) *psbM*, a part of the photosystem II reaction center, is absent in the cp genome, but is present in the nucleus (chromosome 12); and 5) at least 13

additional genes (*petN*, *minE*, *minD*, *ftsI*, *ftsW*, *ftsH*, *rpl12*, *rpl19*, *accD*, *cemA*, *ccsA*, *cysA*, and *cysT*) and 3 unknown conserved genes (*ycf4*, *ycf6*, and *ycf10*) have been lost in *O. tauri*. However, 5 of them are present in the nuclear genome (*ycf4*, *minD*, *rpl12*, *rpl19*, and *cemA/ycf10*) (see supplementary table S3, Supplementary Material online).

Despite these differences in gene content, 10 conserved blocks, ranging from 2 to 12 genes are shared between *O. tauri* and *N. olivacea*, 11 between *O. tauri* and *C. vulgaris*, and 12 between *O. tauri* and *M. viride*. When aligning the 4 genomes together, 9 conserved blocks of at least 2 genes can be unveiled. However, when adding the cp genome of *C. reinhardtii*, whose genome is structurally the most comparable to that of *O. tauri* (see below), almost no conserved blocks shared by all species, can be detected. Comparison of the cp genome of *O. tauri* with the one of *O. viridis*, a member of the Ulvophyceae, also showed shared gene clusters. So in general, without considering *C. reinhardtii*, 9 conserved blocks of at least 2 genes can be unveiled between different members of the Chlorophyta, representing 33 genes (for *O. tauri* 37% of its gene content), indicating the importance of maintaining certain gene clusters throughout evolution. However, if we compare the gene order of *O. tauri* cp genome with the 24 “ancestral” gene clusters present in *N. olivacea* and *M. viride* (de Cambiaire et al. 2006), only 7 of them are completely present in *O. tauri*, indicating the loss of its ancestral characteristics.

The number of gene inversions necessary to transform the gene organization of one genome into another has been estimated for 4 Chlorophyta (*O. tauri*, *N. olivacea*, *O. viridis*, and *C. vulgaris*) and for *M. viride*. An average of 50 inversions is needed to transform the gene organization of *O. tauri* into that of any other of these cp genomes.

Although some genes and gene clusters are well conserved among green algae, the overall structure of the cp genomes can show remarkable differences. First, both the LSC and the SSC region of *O. tauri* cp genome contain 41 genes, in contrast to the cp genomes of other green algae (*N. olivacea*, *M. viride*, *O. viridis*, and *P. akinetum*), where most of the genes are located in the LSC region (Pombert et al. 2006a). Second, the difference in length between the 2 SSCs is much smaller than in other Chlorophyta (e.g., in *N. olivacea*, the LSC region is 5.6 times larger than its SSC region) or even Streptophyta (e.g., in *A. thaliana*, the LSC region is 4.7 times larger than its SSC region) (table 3). In this respect, the cp genome of *O. tauri* is more similar to the cp genome of *C. reinhardtii* (Maul et al. 2002) for 2 reasons: 1) the SSCs have almost identical lengths and both contain an almost identical number of genes (81 and 78, respectively) and 2) the IRs, which in both cases cover almost 20% of the genome, contain exactly the same genes, orientated in the same direction.

The distribution of different genes over the LSC and SSC regions is highly conserved, not only in the entire streptophyte lineage (*M. viride* and land plant genomes share essentially the same gene partitioning), but also in the early diverging *N. olivacea*, indicating that the last common ancestor of all chlorophytes featured a gene partitioning very similar to that observed in land plants. In this respect, Pombert et al. (2006a) created an ancestral cp genome based on the genomes of *O. viridis* and *P. akinetum*

(both Chlorophyta, belonging to the Ulvophyceae) and compared that with the genome of *N. olivacea*, which is a prasinophyte and can be considered as ancestral to the 2 ulvophyte. They concluded that the LSC region of the ancestral genome of both Ulvophyceae contained only genes characteristic of the LSC region of *N. olivacea* and that the SSC region contained genes usually found in the SSC and LSC region of *N. olivacea*. However, in the *O. tauri* cp genome, the genes are scattered across the LSC and SSC region, and the previous assumption made by Pombert (2006a) holds no longer true for *O. tauri*. Because the Prasinophyceae are not a monophyletic group, it is not surprising that the *O. tauri* cp genome differs significantly from the *N. olivacea* cp genome and that changes in gene partitioning have occurred independently in *O. tauri* from those observed in ulvophycean and chlorophycean algae. With the availability of more cp genomes it will become clearer whether *O. tauri* is an exception to the rule and has undergone specific genome reshuffling or whether different species all have their own independent evolutionary history regarding their cp genome structure.

Also in the cp genome, we looked for the presence of SDRs. Sixty-four repeats larger than 15 bp are present, but none of the detected repeats exceed the length of 25 bp. Almost all these SDRs are located in the coding region of 5 protein-coding genes (*rpl23*, *psbD*, *psaB*, *psaA*, and *psbA*) and 5 tRNAs (see supplementary fig. S5, Supplementary Material online). The GC content of the SDRs is comparable to the overall GC content of the cp genome (38% for the SDRs vs. 39.9% for the cp genome). The number of SDRs in *N. olivacea* is similar, but substantially differs from *C. reinhardtii*, which cp genome is more similar to the *O. tauri* cp genome regarding its structure (see above). In the *O. tauri* cp genome, no direct link can be made between the major reshuffling that took place and the abundance of SDRs, whereas for *C. reinhardtii* the major rearrangements could be explained by the huge collection of SDRs present in its cp genome. Consequently, another mechanism is probably responsible for the large number of rearrangements present in the cp genome of *O. tauri*.

## Conclusion

*Ostreococcus tauri* is the smallest eukaryotic organism known to date, and recently, its small (12.56 Mb), but gene dense nuclear genome has been described (Derelle et al. 2006). Here, we present its mt and cp genome, which makes *O. tauri* one of the very few green lineage organisms for which the 3 genome sequences are available. The 2 *O. tauri* organellar genomes are small and display both common and special features compared with their closest relatives.

The main difference between the *O. tauri* and the other Chlorophyta mt genomes is the presence of a unique duplication, previously unobserved in the Chlorophytae. On the other hand, the mt genome of *O. tauri*, which is the most gene dense among all known green algae, closely resembles the one of *Nephroselmis olivacea*, another member of the Prasinophyceae. This is illustrated by a number of common characteristics: 1) the gene content is almost identical in both genomes; 2) there is a high degree of synteny between the 2 genomes, which is illustrated by the presence of

a number of conserved gene blocks and by a low number of gene inversions necessary to transform the *O. tauri* gene structure into the one of *N. olivacea*; and finally 3) Pombert (2006b) showed that there is an increase in the number of Short Dispersed Repeats (SDR) when moving in the tree from *N. olivacea* to the more derived lineages within the Chlorophyta. These analyses were confirmed by *O. tauri*, which contains even fewer SDRs than *N. olivacea*. All these data clearly show that the mt genome of *O. tauri* shares the “after ancestral pattern of evolution typified by the *N. olivacea* genome. This conclusion for *N. olivacea* representing an ancestral state (Turmel et al. 1999b) was based on its basal phylogenetic position in the chlorophyte lineage, on the presence of 3 genes (*nad10*, *rpl14*, and *rnpB*) that had not been identified at that time in any other mt genome (today, *rpl14* is also identified in *P. akinetum*), and on its ancestral organizational pattern. These arguments also hold for the *O. tauri* mt genome, and most likely, both the *O. tauri* and *N. olivacea* mt genome represent the most ancestral form known to date for the green lineage. Whether the unique duplication seen in *Ostreococcus* is restricted to this organism will hopefully become clear with the availability of more mt genomes of basal green algae (e.g., the one of *Micromonas pusilla*, another prasinophyte which is currently being sequenced; Worden A, personal communication).

The *O. tauri* cp genome is very compact, and both the genome size and the gene number are the smallest known among the green plants and green algae. Looking at the gene content, the *O. tauri* cp genome lost many genes compared with other prasinophyte green algae or to *M. viride*. This is well illustrated by the small number of ancestral gene clusters still present in the *O. tauri* cp genome where only 7 of the 24 *Mesostigma/Nephroselmis* gene clusters (de Cambiaire et al. 2006) could be uncovered. Finally, although gene partitioning among LSC and SSC regions is well conserved in all Streptophyta and early-diverging Chlorophyta, the genes in the *O. tauri* cp genome are randomly distributed between both regions. All these data strongly suggest that, in contrast to its mt genome, the *O. tauri* cp genome seems to have lost most of the ancestral features observed in the *M. viride* and *N. olivacea* genomes.

### Supplementary Material

The genome data have been submitted to the European Molecular Biology Laboratory, [www.embl.org](http://www.embl.org) (accession numbers CR954200 [mt genome] and CR954199 [cp genome]) or can be found at <http://bioinformatics.psb.u-gent.be/>. Supplementary tables S1–S5 and figures S1–S5 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

### Acknowledgments

We would like to thank Saker Grootjans for his help in the phylogenetic analyses, Jeroen Raes for discussions, Yvan Saeys for help with the figures, and Igor Grigoriev, Brian Palenik, and the Joint Genome Institute for prior access to the *O. lucimarinus* data. S.R. is indebted to the Institute for the Promotion of Innovation by Science and Technology in Flanders for a predoctoral fellowship. This work was sup-

ported by the Génopole Languedoc-Roussillon and the French research ministry, and was conducted within the framework of the “Marine Genomics Europe” European Network of Excellence (GOCE-CT-2004-505403).

### Literature Cited

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.
- Bélanger AS, Brouard JS, Charlebois P, Otis C, Lemieux C, Turmel M. 2006. Distinctive architecture of the chloroplast genome in the chlorophyte green alga *Stigeoclonium helveticum*. *Mol Genet Genomics.* 276:464–477.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. 2002. GenBank. *Nucleic Acids Res.* 30:17–20.
- Bhattacharya D, Weber K, An SS, Berning-Koch W. 1998. Actin phylogeny identifies *Mesostigma viride* as a flagellate ancestor of the land plants. *J Mol Evol.* 47:544–550.
- Brennicke A, Grohmann L, Hiesel R, Knoop V, Schuster W. 1993. The mitochondrial genome on its way to the nucleus: different stages of gene transfer in higher plants. *FEBS Lett.* 325:140–145.
- Chrétiennot-Dinet MJ, Courties C, Vaquer A, Neveux J, Claustre H, Lautier J, Machado MC. 1995. A new marine picoeukaryote: *Ostreococcus tauri* gen. et sp. Nov. (Chlorophyta, Prasinophyceae). *Phycologia.* 4:285–292.
- Courties C, Perasso R, Chrétiennot-Dinet MJ, Gouy M, Guillou L, Troussellier M. 1998. Phylogenetic analysis and genome size of *Ostreococcus tauri* (Chlorophyta, Prasinophyceae). *J Phycol.* 34:844–849.
- Courties C, Vaquer A, Troussellier M, Lautier J, Chrétiennot-Dinet M-J, Neveux J, Machado MC, Claustre H. 1994. Smallest eukaryotic organism. *Nature.* 370:255.
- Dayhoff MO. 1979. Atlas of protein sequence and structure. Vol. 5 (suppl 3). Silver Spring (MD): National Biochemical Foundation.
- de Cambiaire JC, Otis C, Lemieux C, Turmel M. 2006. The complete chloroplast genome sequence of the chlorophyte 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 Koning AP, Keeling PJ. 2006. The complete plastid genome sequence of the parasitic green alga *Helicosporidium* sp. is highly reduced and structured. *BMC Biol.* 21:12.
- De Rijk P, De Wachter R. 1993. DCSE, an interactive tool for sequence alignment and secondary structure research. *Comput Appl Biosci.* 9:735–740.
- De Rijk P, Wuyts J, De Wachter R. 2003. RnaViz 2: an improved representation of RNA secondary structure. *Bioinformatics.* 19:299–300.
- Derelle E, Ferraz C, Rombauts S, et al. (26 co-authors). 2006. From the cover: genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc Natl Acad Sci USA.* 103:11647–11652.
- Dickerson RE. 1971. The structures of cytochrome c and the rates of molecular evolution. *J Mol Evol.* 1:26–45.
- Eddy SR. 2002. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics.* 3:18.
- Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8:175–185.

- Fauron C, Casper M, Gao Y, Moore B. 1995. The maize mitochondrial genome: dynamic, yet functional. *Trends Genet.* 11:228–235.
- Felsenstein J. 1989. PHYLIP (phylogeny inference package). Version 3.2. *Cladistics.* 5:164–166.
- Graham LE, Wilcox LW. 2000. Green algae I—introduction and prasinophyceans. In: Graham LE, Wilcox LW, editors. *Algae*. Upper Saddle River, Prentice Hall. p. 397–419.
- Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. 2005. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* 33:D121–D124.
- Grzebyk D, Schofield O. 2003. The mesozoic radiation of eukaryotic algae: the portable plastid hypothesis. *J Phycol.* 39:259–267.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.
- Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser.* 41:95–98.
- Handa H. 2003. The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus L.*): comparative analysis of the mitochondrial genomes of rapeseed and *Arabidopsis thaliana*. *Nucleic Acids Res.* 31:5907–5916.
- Hasegawa M, Kishino H, Yano T. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22:160–174.
- Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics.* 17:754–755.
- Janska H, Sarria R, Woloszynska M, Arrieta-Montiel M, Mackenzie SA. 1998. Stoichiometric shifts in the common bean mitochondrial genome leading to male sterility and spontaneous reversion to fertility. *Plant Cell.* 10:1163–1180.
- Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* 8:275–282.
- Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. *New York: Mammalian Press.* p. 21–123.
- Kanazawa A, Tsutsumi N, Hirai A. 1994. Reversible changes in the composition of the population of mtDNAs during dedifferentiation and regeneration in tobacco. *Genetics.* 138:865–870.
- Karol KG, McCourt RM, Cimino MT, Delwiche CF. 2001. The closest living relatives of land plants. *Science.* 294:2351–2353.
- Kimura M. 1983. *The neutral theory of molecular evolution.* Cambridge: Cambridge University Press.
- Knoop V. 2004. The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. *Curr Genet.* 46:123–139.
- Kubo T, Nishizawa S, Sugawara A, Ichihara N, Estiati A, Mikami T. 2000. The complete nucleotide sequence of the mitochondrial genome of sugar beet (*Beta vulgaris L.*) reveals a novel gene for tRNA(Cys)(GCA). *Nucleic Acids Res.* 28:2571–2576.
- Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29:4633–4642.
- Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* 5:R12.
- Lemieux C, Otis C, Turmel M. 2000. Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. *Nature.* 403:649–652.
- Lister DL, Bateman JM, Purton S, Howe CJ. 2003. DNA transfer from chloroplast to nucleus is much rarer in *Chlamydomonas* than in tobacco. *Gene.* 316:33–38.
- Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25:955–964.
- Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci USA.* 99:12246–12251.
- Mattox KR, Stewart KD. 1984. Classification of the green algae: a concept based on comparative cytology. In: Irvine DEG, John DM, editors. *The systematics of green algae.* London: Academic Press. p. 29–72.
- Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH, Stern DB. 2002. The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. *Plant Cell.* 14:2659–2679.
- McCourt RM, Delwiche CF, Karol KG. 2004. Charophyte algae and land plant origins. *Trends Ecol Evol.* 19:661–666.
- Melkonian M. 1989. Flagellar apparatus ultrastructure in *Mesostigma viride* (Prasinophyceae). *Plant Syst Evol.* 164:93–122.
- Michaelis G, Vahrenholz C, Pratje E. 1990. Mitochondrial DNA of *Chlamydomonas reinhardtii*: the gene for apocytochrome b and the complete functional map of the 15.8 kb DNA. *Mol Gen Genet.* 223:211–216.
- Nedelcu AM, Lee RW, Lemieux C, Gray MW, Burger G. 2000. The complete mitochondrial DNA sequence of *Scenedesmus obliquus* reflects an intermediate stage in the evolution of the green algal mitochondrial genome. *Genome Res.* 10:819–831.
- Nozaki H, Misumi O, Kuroiwa T. 2003. Phylogeny of the quadriflagellate Volvocales (Chlorophyceae) based on chloroplast multigene sequences. *Mol Phylogenet Evol.* 29:58–66.
- Oda K, Kohchi T, Ohyama K. 1992. Mitochondrial DNA of *Marchantia polymorpha* as a single circular form with no incorporation of foreign DNA. *Biosci Biotechnol Biochem.* 56:132–135.
- Ohyama K, Fukuzawa H, Kohchi T, et al. (13 co-authors). 1986. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature.* 322:572–574.
- Page RD. 1996. TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci.* 12:357–358.
- Petersen J, Teich R, Becker B, Cerff R, Brinkmann H. 2006. The GapA/B gene duplication marks the origin of Streptophyta (charophytes and land plants). *Mol Biol Evol.* 23:1109–1118.
- Pombert JF, Lemieux C, Turmel M. 2006a. The complete chloroplast DNA sequence of the green alga *Oltmannsiellopsis viridis* reveals a distinctive quadripartite architecture in the chloroplast genome of early diverging ulvophytes. *BMC Biol.* 4:3.
- Pombert JF, Beauchamp P, Otis C, Lemieux C, Turmel M. 2006b. The complete mitochondrial DNA sequence of the green alga *Oltmannsiellopsis viridis*: evolutionary trends of the mitochondrial genome in the Ulvophyceae. *Curr Genet.* 50:137–147.
- 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 *Pseudendoclonium akinetum* (Ulvophyceae) reveals unusual structural features and new insights into the branching order of chlorophyte lineages. *Mol Biol Evol.* 22:1903–1918.

- Robbens S, Petersen J, Brinkmann H, Rouzé P, Van de Peer Y. Forthcoming 2007. Unique regulation of the calvin cycle in the ultrasmall green algae *Osterococcus*. *J. Mol. Evol.*
- Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000. Artemis: sequence visualization and annotation. *Bioinformatics*. 16:944–945.
- Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S. 1999. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Res.* 6:283–290.
- Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W. 2000. PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.* 10:577–586.
- Shinozaki K, Ohme M, Tanaka M, et al. (23 co-authors). 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J.* 5:2043–2049.
- Sonnhammer EL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*. 167:GC1–G10.
- Stegemann S, Hartmann S, Ruf S, Bock R. 2003. High-frequency gene transfer from the chloroplast genome to the nucleus. *Proc Natl Acad Sci USA.* 100:8828–8833.
- Stirewalt VL, Michalowski CB, Löffelhardt W, Bohnert HJ, Bryant DB. 1995. Nucleotide sequence of the cyanelle genome from *Cyanophora paradoxa*. *Plant Mol Biol Rep.* 13:327–332.
- Stoesser G, Baker W, van den Broek A, et al. (16 co-authors). 2002. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* 30:21–26.
- Sugiyama Y, Watase Y, Nagase M, Makita N, Yagura S, Hirai A, Sugiura M. 2005. The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants. *Mol Genet Genomics.* 272:603–615.
- Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, Gojobori T. 2002. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res.* 30:27–30.
- Tesler G. 2002. GRIMM: genome rearrangements web server. *Bioinformatics.* 18:492–493.
- Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–4680.
- Turmel M, Otis C, Lemieux C. 1999a. The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes. *Proc Natl Acad Sci USA.* 96:10248–10253.
- Turmel M, Lemieux C, Burger G, Lang BF, Otis C, Plante I, Gray MW. 1999b. The complete mitochondrial DNA sequences of *Nephroselmis olivacea* and *Pedinomonas minor*. Two radically different evolutionary patterns within green algae. *Plant Cell.* 11:1717–1730.
- Turmel M, Otis C, Lemieux C. 2002a. The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. *Proc Natl Acad Sci USA.* 99:11275–11280.
- Turmel M, Otis C, Lemieux C. 2002b. The complete mitochondrial DNA sequence of *Mesostigma viride* identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants. *Mol Biol Evol.* 19:24–38.
- Turmel M, Otis C, Lemieux C. 2003. The mitochondrial genome of *Chara vulgaris*: insights into the mitochondrial DNA architecture of the last common ancestor of green algae and land plants. *Plant Cell.* 15:1888–1903.
- Unsel M, Marienfeld JR, Brandt P, Brennicke A. 1997. The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. *Nat Genet.* 15:57–61.
- Vahrenholz C, Riemen G, Pratje E, Dujon B, Michaelis G. 1993. Mitochondrial DNA of *Chlamydomonas reinhardtii*: the structure of the ends of the linear 15.8-kb genome suggests mechanisms for DNA replication. *Curr Genet.* 24:241–247.
- Van de Peer Y, De Wachter R. 1994. TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput Appl Biosci.* 10:569–570.
- Wakasugi T, Nagai T, Kapoor M, et al. (15 co-authors). 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 USA.* 94:5967–5972.
- Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, Sugiura M. 1994. Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. *Proc Natl Acad Sci USA.* 91:9794–9798.
- Wuyts J, Perriere G, Van De Peer Y. 2004. The European ribosomal RNA database. *Nucleic Acids Res.* 32:D101–D103.
- Zito F, Vinh J, Popot JL, Finazzi G. 2002. Chimeric fusions of subunit IV and PetL in the b6f complex of *Chlamydomonas reinhardtii*: structural implications and consequences on state transitions. *J Biol Chem.* 277:12446–12455.
- Zuckerklund E, Pauling L. 1965. Molecules as documents of evolutionary history. *J Theor Biol.* 8:357–366.

Peter Lockhart, Associate Editor

Accepted January 18, 2007