Mol Genet Genomics (2007) 277:427–439
DOI 10.1007/s00438-006-0199-4

ORIGINAL PAPER

# Chloroplast genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*: comparison with other plastid genomes of the red lineage

**Marie-Pierre Oudot-Le Secq · Jane Grimwood · Harris Shapiro · E. Virginia Armbrust · Chris Bowler · Beverley R. Green**

Received: 1 September 2006 / Accepted: 30 November 2006 / Published online: 25 January 2007


**Abstract** The chloroplast genomes of the pennate diatom *Phaeodactylum tricornutum* and the centric diatom *Thalassiosira pseudonana* have been completely sequenced and are compared with those of other secondary plastids of the red lineage: the centric diatom *Odontella sinensis*, the haptophyte *Emiliania huxleyi*, and the cryptophyte *Guillardia theta*. All five chromist genomes are compact, with small intergenic regions and no introns. The three diatom genomes are similar in gene content with 127–130 protein-coding genes, and genes for 27 tRNAs, three ribosomal RNAs and two small RNAs (tmRNA and signal recognition particle RNA). All three genomes have open-reading frames corresponding to ORFs148, 355 and 380 of *O. sinensis*, which have been assigned the names *ycf88*, *ycf89* and *ycf90*. Gene order is not strictly conserved, but there are a number of conserved gene clusters showing remnants of red algal origin. The *acpP, tsf* and *psb28* genes appear to be on the way from the plastid to the host nucleus, indicating that endosymbiotic gene transfer is a continuing process.

**Keywords** Chloroplast · Plastid · Genome · Secondary endosymbiosis · Diatom

Communicated by R. Herrmann.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00438-006-0199-4) contains supplementary material, which is available to authorized users.

M.-P. Oudot-Le Secq · B. R. Green (✉)
Department of Botany, University of British Columbia,
#3529-6270 University Blvd., Vancouver, BC,
Canada, V6T 1Z4
e-mail: brgreen@interchange.ubc.ca

J. Grimwood
Stanford/JGI, Stanford Human Genome Center,
975 California Avenue, Palo Alto, CA 94304, USA

H. Shapiro
Department of Energy, Joint Genome Institute,
Walnut Creek, CA 94598, USA

E. V. Armbrust
School of Oceanography, University of Washington,
Seattle, WA 98195, USA

C. Bowler
Cell Signalling Laboratory, Stazione Zoologica,
Villa Comunale, 80121 Naples, Italy

C. Bowler
CNRS FRE2910, Département de Biologie,
Ecole Normale Supérieure, 46 rue d'Ulm, 75230 Paris, France

## Introduction

Diatoms are a group of marine algae that make up an important part of marine food webs and contribute significantly to drawdown of atmospheric $CO_2$ (Field et al. 1998; Falkowski et al. 2004). They are also important from the evolutionary point of view, since they represent a lineage of photosynthetic eukaryotes that acquired chloroplasts by secondary rather than primary endosymbiogenesis (Gibbs 1981). It is now well established that all chloroplasts are the descendants of a primary endosymbiotic relationship in which a cyanobacterium was engulfed by (or invaded) a heterotrophic eukaryote, and eventually lost most of its genes, many of which were transferred to the host nucleus. This ancestral photosynthetic lineage diversified into all the modern groups with "primary" chloroplasts: the

rhodophyte (red) algae, the glaucophyte algae, the green algae and land plants. Subsequently, there were a number of secondary endosymbioses where a non-photosynthetic eukaryotic host acquired a eukaryotic endosymbiont that already had a chloroplast. The host kept the chloroplast and some of the endosymbiont's nuclear genes, which were incorporated into the host nuclear genome. Because the secondary chloroplasts were now surrounded by four membranes, this required considerable adaptation at the cellular level in terms of intracellular transport and the coordination of cellular activities (Cavalier-Smith 2000; McFadden 2001).

Diatoms belong to the photosynthetic heterokonts, one of the four large groups of algae that acquired their chloroplasts from a red algal endosymbiont. The first "secondary" algal chloroplast genome sequenced was that of the centric diatom *Odontella sinensis* (Kowallik et al. 1995). There are now complete chloroplast genome sequences from representatives of two other groups that acquired their chloroplasts from red algae: the cryptophyte *Guillardia theta* (Douglas and Penny 1999) and the haptophyte *Emiliania huxleyi* (Sánchez Puerta et al. 2005). Chloroplast genomes of all three groups (collectively referred to as chromists) have fewer genes than those of red algae, but more than those of green algae or plants. The dinoflagellates, the fourth group to have acquired red algal chloroplasts, appear to have taken gene loss to the extreme: only 14–16 genes remain in the chloroplast genome, and they are generally found on individual DNA minicircles (Green 2004).

The global importance of diatoms was recognized by the complete sequencing of the nuclear genome of the centric diatom *Thalassiosira pseudonana* (Armbrust et al. 2004) and the recently completed sequencing of the pennate diatom *Phaeodactylum tricornutum* (http://www.genome.jgi-psf.org/Phatr1/Phatr1.home.html). We are now in a position to compare the chloroplast genome sequences of both these diatoms with that of the centric diatom *O. sinensis* and those of other members of the red lineage. By "red lineage", we refer to the red algae and the four groups that acquired their chloroplasts from a red algal ancestor by secondary endosymbiosis, namely the heterokonts, haptophytes, cryptophytes and dinoflagellates.

## Methods

### Sequencing and assembly

The monoclonal culture of *P. tricornutum* chosen for sequencing was CCAP1055/1, derived from a single cell of strain CCMP632. The whole genome was obtained by sequencing three whole genome libraries of insert size 3, 8 and 35 kb (unpublished data). It was assembled using the JAZZ assembler (Aparicio et al. 2002; Chapman, unpublished data), and scaffolds containing the chloroplast genome were identified by similarity to other chloroplast genomes. The chloroplast genome of *T. pseudonana* was identified and assembled as previously described (Armbrust et al. 2004).

To perform finishing, initial read layouts from the chloroplast scaffolds were converted into the Phred/Phrap/Consed pipeline (Gordon et al. 1998). All sequences were manually inspected and repeats resolved using Orchid (http://www-shgc.stanford.edu/informatics/orchid.html). Both chloroplast genomes are circular and have an estimated error rate of less than 1 error in 100,000 bp. Their sequences are available in Genbank under accession numbers EF067920 and EF067921. Finished versions of both nuclear genomes can be accessed at http://www.genome.jgi-psf.org/.

### Gene finding and analysis

The annotation of the plastid genomes was done manually in Artemis (Rutherford et al. 2000). Open reading frames (ORFs) longer than 100 amino acids were identified by BLAST (Altschul et al. 1997) and /or direct alignment with the *O. sinensis* plastid genome using BioEdit (Hall 1999). Genes and gene clusters conserved in the three diatom genomes were identified using PipMaker (Schwartz et al. 2000) and MAUVE (Darling et al. 2004), and the alignments were refined manually with BioEdit. The GRIMM webserver (Tesler 2002) and the GRAPPA-IR program (Cui et al. 2006a) were used to infer the putative rearrangements of the clusters in the large single copy (LSC) region of the three diatom genomes. The dataset was made up of the thirteen clusters plus the six genes found between those clusters in the three genomes.

A number of genes and conserved open-reading frames (*ycf's*) have been identified and named in the 10 years since the *O. sinensis* genome was published (Kowallik et al. 1995). However, nine very small AT-rich ORFs in that genome were omitted from the analysis because they have no counterparts in the other diatoms. An updated list of chloroplast genes in all five chromists is given in Supplementary Table 1.

To identify transfer RNA genes, the sequences were searched for the motif GTTCRANYC using the application fuzznuc, from the EMBOSS suite (Rice et al. 2000). Motifs were then mapped on the sequence in Artemis, and the putative clover-leaf folding around the motif was checked to determine the presence of

tRNA genes. ARAGORN 1.1 (Laslett and Canback 2004) and tRNAscan-SE 1.21 (Lowe and Eddy 1997) search servers were used for confirmation and to find the trnE gene, which has a modified motif. Ribosomal RNAs were roughly localized and identified by BLAST, and precisely delimited through manual alignment with other plastid, bacterial and mitochondrial genes, taking the secondary structure into account.

The program PipMaker (Schwartz et al. 2000) was used to run each genome against itself, to have a global view of the existence of direct repeats. Four applications of the EMBOSS suite were used to detect tandem and inverted repeats (equicktandem, etandem, palindrome and einverted). The default parameters were used with the following modifications: equicktandem: maximum size 250 bp, threshold 20; etandem: 8–30 bp, 30–60 bp, 60–100 bp and 100–250 bp size ranges; palindrome: minimum size 12 bp, maximum size 20 bp, maximum gap between 30, one mismatch allowed, no overlapping. After manual verification, results with less than 80% identity were discarded.

The application fuzznuc (EMBOSS) was used to search for putative phage T7-like and bacterial promoters as well as Shine-Dalgarno sequences. The 150 nt upstream of each gene were searched for the motif ACTCACTA allowing one mismatch for the T7 phage-type promoter (motif chosen from Chen and Schneider 2005), and two possible −35 boxes (TTTAAA and TTGACA) separated by 16–20 nt from the canonical sequence TATAAT for the −10 box, allowing one mismatch. For translation, 50 nt upstream of the start codon was searched for sub-motifs of the Shine-Dalgarno sequence, i.e., AAGG, AGGA, GGAG, GAGG (Hagopian et al. 2004).

To detect evidence of recent gene transfer from chloroplast to nucleus, the chloroplast proteomes of all three diatoms, *G. theta* and one or more red algae were used to search the nuclear genomes of *P. tricornutum* and *T. pseudonana* using the program BLASTP (Altschul et al. 1997). Hits with scores over 100 and e-values less than $e^{-10}$ were examined manually and compared to all publicly available sequences. Proteins targeted to the chloroplast were identified by their ER targeting presequence, detected by SignalP-version 3.0 (Bendtsen et al. 2004), and by the following FxP motif (Kilian and Kroth 2005).

## Results

### General features of the chloroplast genomes

The chloroplast genomes of both diatoms were assembled from whole-genome sequence data and finished as described in "Methods" section. The *T. pseudonana* chloroplast genome was briefly described in Armbrust et al. (2004), but its analysis was not completed at that time. The general features of these two diatom chloroplast genomes are compared with the previously sequenced genome of the centric diatom *O. sinensis* in Table 1. All three genomes map as a single circle with two inverted repeats (IRs) dividing the genome into large single copy (LSC) and small single copy (SSC) regions as in many other plastid genomes. The

**Table 1** General features of diatom, cryptophyte and haptophyte chloroplast genomes

| | Organism | | | | |
|---|---|---|---|---|---|
| | *Phaeodactylum tricornutum* | *Thalassiosira pseudonana* | *Odontella sinensis* | *Emiliania huxleyi* | *Guillardia theta* |
| Size (bp) | 117,369 | 128,814 | 119,704 | 105,309 | 121,524 |
| Inverted repeat (IR) | 6,912 | 18,337 | 7,725 | 4,841 | 4,967 |
| Small single-copy region (SSC) | 39,871 | 26,889 | 38,908 | 11,183 | 15,421 |
| Large single-copy region (LSC) | 63,674 | 65,250 | 65,346 | 84,444 | 96,309 |
| Total G + C content (%) | 32.56 | 30.66 | 31.82 | 36.81 | 32.97 |
| Gene content (total)[a] | 162 | 159 | 160 | 144 | 177 |
| % Coding sequence | 87.5 | 85.2 | 84.1 | 86.3 | 87.9 |
| Protein-coding genes (%GC) | 130 (32.9) | 127 (31.5) | 128 (32.7) | 113 (37.3) | 144 (33.3) |
| rRNA genes (%GC) | 3 (47.2) | 3 (47.0) | 3 (46.6) | 3 (48.2) | 3 (48.4) |
| tRNA genes (%GC) | 27 (53.0) | 27 (52.6) | 27 (53.2) | 27 (54.2) | 28 (54.5) |
| Other RNAs (%GC) | 2 (26.0) | 2 (25.6) | 2 (27.7) | 1 (29.9) | 2 (31.7) |
| No. of overlapping genes | 4 | 4 | 4 | 1 | 5 |
| No. of introns | 0 | 0 | 0 | 0 | 0 |
| Average size intergenic spacer in bp (%GC) | 88.4 (18.8) | 108.2 (16.3) | 115.7 (18.1) | 97.6 (24.6) | 82.9 (18.4) |
| Start codons: ATG | 124 | 121 | 123 | 105 | 136 |
| Start codons: GTG | 5 | 5 | 5 | 6 | 6 |
| Start codons: other | 1 ATT | 1 ATA | None | None | 2 TTG |

[a] Genes duplicated in the IR are only counted once

*T. pseudonana* chloroplast genome is the largest of the three diatom genomes, and is slightly larger than that of the cryptophyte *G. theta* (Douglas and Penny 1999). Its larger size is due to the fact that the IRs encompass more genes than in the other two diatoms (Fig. 1). It should be noted that the *T. pseudonana* plastid genome appeared as concatemers in the optical map (Armbrust et al. 2004), suggesting that at least a fraction of the molecules may not be circular (Bendich 2004).

All five chromist chloroplast genomes are compact (Table 1). Their most striking feature is a complete lack of introns. Intergenic spacers tend to be small with average sizes of 91–118 nt. There are four identical cases of overlapping genes in the three diatoms. The genes *sufC-sufB* overlap by 1 nt, *atpD-atpF* by 4 nt, *rpl4-rpl23* by 8 nt and *psbD-psbC* by 53 nt. Although the *E. huxleyi* plastid genome is the smallest (Sánchez Puerta et al. 2005), only the *psbD* and *psbC* genes overlap, also by 53 nt, whereas they overlap by 95 nt in *G. theta*. At least one gene pair in each genome has no intergenic spacer, i.e., the stop codon is immediately followed by the start codon of the next gene.

All three diatom plastid genomes encode the three rRNA subunits (5S, 16S and 23S) in the IRs. The 27 tRNAs are sufficient to satisfy all the requirements for in organello protein synthesis. All protein-coding genes use the standard plastid/bacterial genetic code (code table 11), which differs from the universal code only in the use of additional start codons (Table 1). In each genome there are five genes that have GTG rather than ATG start codons, as well as one gene that initiates with an ATT in *P. tricornutum* and one starting with ATA in *T. pseudonana*. The most common stop codon is TAA.

It is very difficult to detect small RNAs because their function depends more on three-dimensional structure than on primary sequence conservation (e.g., Bullerwell et al. 2003). However, we found the gene for a tmRNA (transfer-messenger RNA, Gueneau et al. 1999; Gueneau de Novoa and Williams 2004) as well as the gene *ffs* encoding the chloroplast signal recognition particle RNA. The latter was predicted in *O. sinensis* and several algae on the basis of a probabilistic model taking into account both sequence and secondary structure (Rosenblad and Samuelsson 2004). An RNA gene present in three red algae (*Gracilaria tenuistipitata*, *Cyanidioschyzon merolae, Porphyra purpurea*) but missing in the chromists is *rnpB*, which encodes a small RNA that is part of ribonuclease P (de la Cruz and Vioque 2003; Seif et al. 2003). Because of the conserved position of this gene in the three red algal plastid genomes, our inability to find it in the diatom genomes probably means it has been lost from them (see below).

The protein-coding gene complement of the three diatom species is almost identical, with three notable exceptions (Table 2). The *P. tricornutum* plastid retains three genes, *acpP* (acyl carrier protein), *syfB* (Phe tRNA synthetase) and *tsf* (EF-Ts homolog), which have all been lost in *T. pseudonana*. The *tsf* gene is found on the chloroplast genome of another heterokont, *Fucus vesiculosus* (accession no. DQ307683), suggesting that the loss in *T. pseudonana* may be relatively recent. The *syfB* gene is found on the plastid genomes of three red algae and in cyanobacteria; the only other significant homologs are in two insects.

*O. sinensis* ORFs 148, 355, and 380 are conserved in all three diatom chloroplast genomes, but are not found in any other chloroplast genome. After consultation with annotators at SwissProt, they have been named *ycf88* (ORF148), *ycf89* (ORF355) and *ycf90* (ORF380) (A. Auchincloss, personal communication). None of the AT-rich orfs annotated in the *O. sinensis* plastid genome was found in the *T. pseudonana* and *P. tricornutum* genomes. Because the chloroplast genome of *O. sinensis* was published more than 10 years ago (Kowallik et al. 1995), a number of genes and conserved open-reading frames (*ycf's*) had not been identified and named at that time. An updated list of protein-coding genes in the five chromist plastid genomes is therefore provided in Supplementary Table 1.

## Gene transfer during primary and secondary endosymbioses

All five red algal chloroplast genomes that have been sequenced carry more genes than any of the chromist genomes, suggesting that plastid gene transfer to the nucleus continued after secondary endosymbiosis. A comparison of the gene content of the five available chromist plastid genomes (Table 1; Supplementary Table 1) shows that the cryptophyte *G. theta* has 20 genes not found in all three diatoms, most of which are on the plastid genome of one or more red algae. *G. theta* also has a relict red nucleus (the nucleomorph) suggesting that the integration of the secondary endosymbiont has not reached the same stage in the cryptophyte lineage as it has in the heterokont or haptophyte algae.

The haptophyte *E. huxleyi* has the smallest number of plastid genes of the chromists so far sequenced, most notably 11 fewer ribosomal protein genes (Sánchez Puerta et al. 2005). We searched for nuclear homologs in the available haptophyte EST databases at NCBI (www.ncbi.nlm.nih.gov) and found ESTs for a number of them, suggesting that the genes may have been

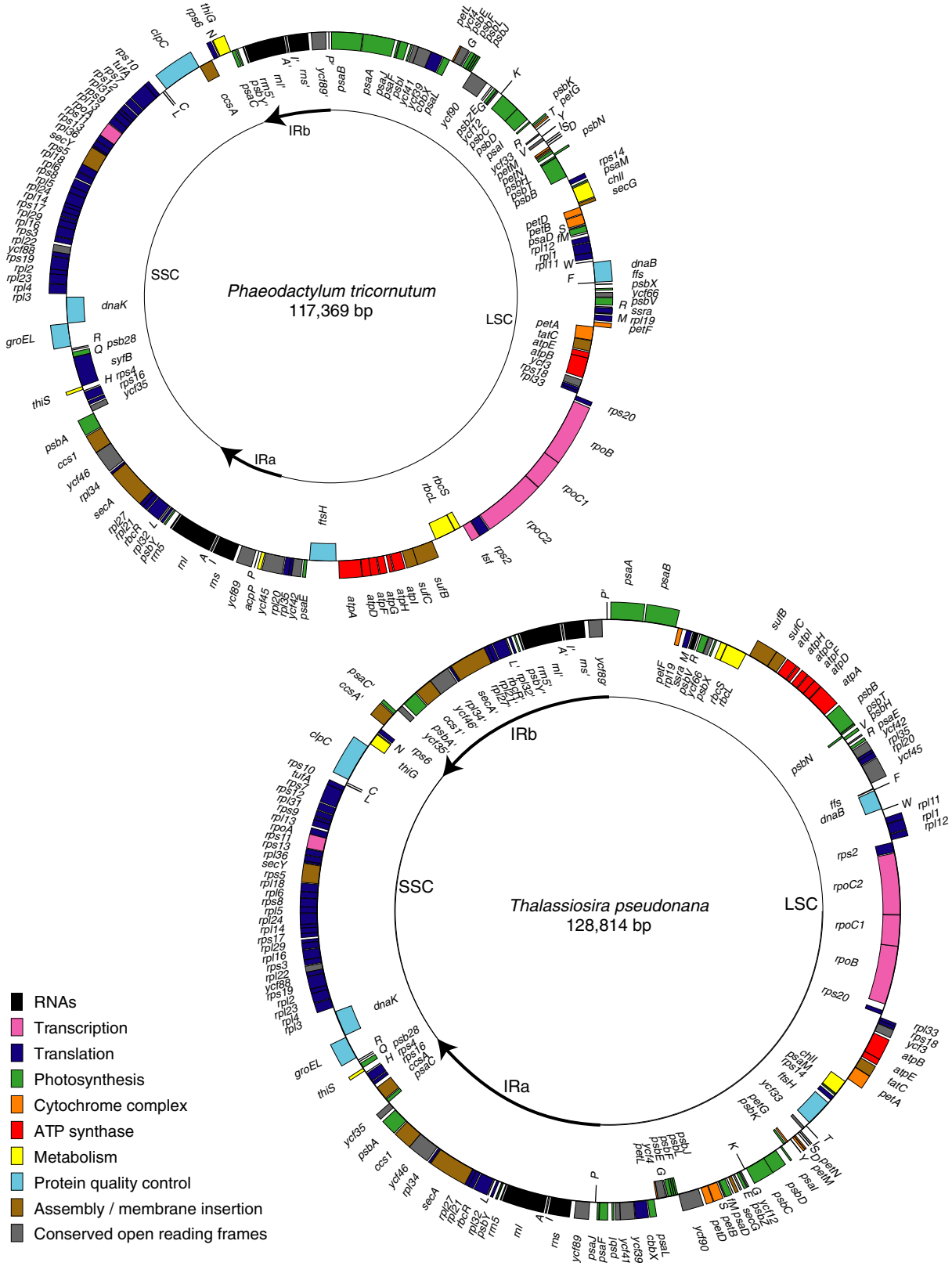


**Fig. 1** Chloroplast genome maps of *P. tricornutum* and *T. pseudonana*. Genes on the outside are transcribed clockwise; those on the inside counterclockwise. Genes are colour-coded by functional category as shown at *bottom left*. The tRNA genes are indicated by the *single-letter code* of the corresponding amino acid. *IRa* and *IRb* inverted repeats; *LSC* large single-copy region; *SSC* small single-copy region

**Table 2** Diatom plastid gene list

| | |
|---|---|
| RNAs | |
| Ribosomal | *rns*, *rnl*, *rrn5* |
| Transfer | *trnA*(*ugc*), *trnC*(*gca*), *trnD*(*guc*), *trnE*(*uuc*), *trnF*(*gaa*), *trnG*(*gcc*), *trnG*(*ucc*), *trnH*(*gug*), *trnI*(*cau*), *trnI*(*gau*), *trnK*(*uuu*), *trnL*(*caa*), *trnL*(*uaa*), *trnM*(*cau*), *trnN*(*guu*), *trnP*(*ugg*), *trnQ*(*uug*), *trnR*(*acg*), *trnR*(*ccg*), *trnR*(*ucu*), *trnS*(*gcu*), *trnS*(*uga*), *trnT*(*ugu*), *trnV*(*uac*), *trnW*(*cca*), *trnY*(*gua*) |
| Others | *ffs*, *ssra* |
| Transcription | *cbbX*, *rbcR*, *rpoA*, *rpoB*, *rpoC1*, *rpoC2*, *tsf*[a] |
| Translation | *syfB*[a], *tufA* |
| Ribosomal proteins | |
| Small subunit | *rps2*, *rps3*, *rps4*, *rps5*, *rps6*, *rps7*, *rps8*, *rps9*, *rps10*, *rps11*, *rps12*, *rps13*, *rps14*, *rps16*, *rps17*, *rps18*, *rps19*, *rps20* |
| Large subunit | *rpl1*, *rpl2*, *rpl3*, *rpl4*, *rpl5*, *rpl6*, *rpl11*, *rpl12*, *rpl13*, *rpl14*, *rpl16*, *rpl18*, *rpl19*, *rpl20*, *rpl21*, *rpl22*, *rpl23*, *rpl24*, *rpl27*, *rpl29*, *rpl31*, *rpl32*, *rpl33*, *rpl34*, *rpl35*, *rpl36* |
| Photosynthesis | |
| ATP synthase | *atpA*, *atpB*, *atpD*, *atpE*, *atpF*, *atpG*, *atpH*, *atpI* |
| Photosystem I | *psaA*, *psaB*, *psaC*, *psaD*, *psaE*, *psaF*, *psaI*, *psaJ*, *psaL*, *psaM* |
| Photosystem II | *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbN*, *psbT*, *psbV*, *psbX*, *psbY*, *psbZ*, *psb28* |
| Cytochrome complex | *petA*, *petB*, *petD*, *petF*, *petG*, *petL* (*ycf7*), *petM* (*ycf31*), *petN* (*ycf6*) |
| Metabolism | *acpP*[b], *chlI*, *rbcL*, *rbcS*, *thiG*, *thiS* |
| Protein quality control | *clpC*, *dnaB*, *dnaK*, *ftsH* (*ycf25*), *groEL* |
| Assembly, membrane insertion | *ccs1*, *ccsA*, *secA*, *secG*, *secY*, *sufB*, *sufC*, *tatC* |
| Conserved open reading frames | *ycf3*, *ycf4*, *ycf12*, *ycf33*, *ycf35*, *ycf39*, *ycf41*, *ycf42*, *ycf45*, *ycf46*, *ycf66*, *ycf88*, *ycf89*, *ycf90* |

[a] *P. tricornutum* only

[b] *P. tricornutum* and *O. sinensis* only

successfully transferred to the host nucleus and the products retargeted. These included *acpP*, *ftsH*, *petF*, *psaE*, *sufC*, *tsf*, *ycf66* and most of the missing ribosomal proteins. It appears that gene transfer has progressed further in the haptophyte lineage than in either the heterokont or the cryptophyte lineages. On the other hand, the *E. huxleyi* genome has nine genes not found in any of the diatom genomes (Supplementary Table 1).

To investigate the gene transfer process in more detail, the nuclear genomes of *T. pseudonana* and *P. tricornutum* were probed with their plastid proteomes using BLASTP (Altschul et al. 1997). Several chloroplast genes (*acpP*, *cbbX*, *psb28*, *secA*, *tsf*, *tufA*) had strongly predicted homologs in one or both nuclear genomes. We have previously reported (Armbrust et al. 2004) the case of the *psb28* genes[1] in *T. pseudonana*, where there are plastid and nuclear copies that are very closely related to each other, and the nuclear copy has a predicted ER signal sequence as expected for a chloroplast-targeted protein. This appeared to be a gene transfer in progress, where the chloroplast copy had not yet been deleted. To our surprise, we find that *P. tricornutum* has only a chloroplast copy of this gene, not a nuclear one. This suggests that the copying of this gene to the host nucleus in *T. pseudonana* may be a relatively recent event. The gene is present on the *O. sinensis* plastid genome but there is no available nuclear data for that alga.

*T. pseudonana*, which does not have the plastid *tsf* gene, has a nuclear gene with a conventional ER signal peptide followed by the FxP motif characteristic of many chloroplast-targeted sequences (Kilian and Kroth 2005). *P. tricornutum* has both the plastid gene and a nuclear homolog, which is closely-related to the nuclear *T. pseudonana* gene on phylogenetic trees (data not shown), although no signal peptide was predicted. Since the gene is still on the plastid genome in *P. tricornutum*, *O. sinensis* and the phaeophyte *Fucus vesiculosus*, as well as all the red algal plastid genomes, and no nuclear gene with a plastid-targeting sequence was found in the red alga *C. merolae* (Matsuzaki et al. 2004), transfer of this gene to the nucleus appears to be a work in progress in the diatom lineage.

This appears to be the case for another gene, the *acpP* (acyl carrier protein) gene. *P. tricornutum* has three genes: one plastid-encoded and two nuclear-encoded; *T. pseudonana* just has the two nuclear genes. In each diatom, one of the nuclear homologs has a predicted ER signal peptide followed by the FxP motif (Kilian and Kroth 2005) indicative of chloroplast targeting; the other is mitochondrial.

[1] The gene was incorrectly referred to as *psbW* (*psb28*) in Armbrust et al. 2004. The name *psbW* should be used only for a 6 kDa hydrophobic protein that is always nuclear-encoded (Shi and Schroeder 2004). The gene annotated as *psbW* in the *Synechocystis* 6803 genome and the plastid genomes of *O. sinensis*, *G. theta* and several red algae encodes a hydrophilic 13 kDa protein and should be called *psb28* (Kashino et al. 2002).

**Table 3** Gene clusters conserved in the three diatom plastid genomes

| Cluster name | No. of genes | Genes in cluster | Size range |
|---|---|---|---|
| 1 | 2 | *psaA, psaB* | 4.5–4.6 kb |
| 2 | 12 | (*rps2*, *rpoC2, rpoC1, rpoB,* rps20), *rpl33, rps18,* ycf3, *atpB, atpE,* tatC, *petA* | 16.7–17.6 kb |
| 3 | 8 | *psbX*, *ycf66*, *psbV*, *trnR2*, *ssra*, *trnM*, *rpl19*, *petF* | 2.6–2.8 kb |
| 4 | 7 | *ffs*, *trnF*, (*dnaB*), *trnW*, *rpl11*, *rpl1*, *rpl12* | 3.4–3.7 kb |
| 5 | 10 | (*petG*, *psbK*), *psaI*, *psbD, psbC,* (*trnK*), *ycf12*, *psbZ*, *trnG2*, *trnE* | 4.3–4.6 kb |
| 6 | 14 | *ycf90*, *(psbJ, psbL, psbF, psbE,* trnG1, *ycf4*, petL), *psaL*, *cbbX*, *ycf39*, *ycf41*, *psbI*, *psaF, psaJ* | 7.0–7.2 kb |
| 7 | 2 | *petN*, *petM* | 258 bp |
| 8 | 4 | *psbB, psbT, (psbN), psbH* | 2.1 kb |
| 9 | 10 | *(rbcS, rbcL),* sufB, *sufC*, *atpI, atpH, atpG, atpF, atpD, atpA* | 8.9–9.0 kb |
| 10 | 4 | (*ycf33*), *trnI2*, *trnS2*, *trnD* | 655–716 bp |
| 11 | 5 | *trnfM*, *psaD*, *trnS1*, *petB, petD* | 1.9–2.0 kb |
| 12 | 3 | *rps14*, *psaM*, *chlI* | 1.6–1.7 kb |
| 13 | 6 | *psaE*, *ycf42*, *rpl35*, *rpl20*, *ycf45*, *acp*[a] | 2.9–3.6 kb |
| 14 | 7 | *ycf89*, *rns, trnI1, trnA, rnl, rn5, psbY* | 6.7–7.0 kb |
| 15 | 10 | *rpl32*, *trnL2*, *rbcR*, *rpl21*, *rpl27*, *secA*, *rpl34*, *ycf46*, *ccs1*, *psbA* | 8.5–9.0 kb |
| 16 | 42 | (*rps16*, *rps4*, *trnH*), *thiS*, (*syfB*[b], *psb28*, *trnQ*, *trnR1*), *groEL*, *(dnaK), rpl3, rpl4, rpl23, rpl2, rps19,* ycf88[c], *rpl22, rps3, rpl16, rpl29, rps17, rpl14, rpl24, rpl5, rps8, rpl6, rpl18, rps5, secY, rpl36, rps13, rps11, rpoA, rpl13, rps9, rpl31, rps12, rps7, tufA, rps10,* (*trnL1*, *trnC*), *clpC* | 25.3–27.5 kb |
| 17 | 4 | (*ccsA*), *rps6*, *trnN*, *thiG* | 2.4–2.6 kb |

*Double underline* found in red algae and chromists. *Single underline* found in red algae and chromists except for *E. huxleyi*. ( ) indicates genes encoded on the reverse strand

[a] Not in *T. pseudonana*

[b] Only in *P. tricornutum*

[c] Diatom specific

Several plastid genes such as *groEL*, *dnaK*, *ftsH*, *ycf46* and *clpC* are members of multigene families encoding chaperones and proteases. They have homologs in several cell compartments or share conserved motifs (e.g., ATP-binding domains) with many nuclear genes. These genes and the ribosomal protein genes were not analyzed further at this time because they appear to represent ancient transfers and duplications.

## Gene order–gene clusters

Diatom gene clusters were defined using both MultiPipMaker (http://www.pipmaker.bx.psu.edu/cgi-bin/multipipmaker) and MAUVE (Darling et al. 2004) analysis. There are 17 gene clusters conserved among all three diatoms (Table 3). Only four tRNA genes and four or five protein coding genes are interspersed among them (Figs. 2, 3). In diatoms, the largest cluster (cluster 16) encompasses the ribosomal gene cluster (*rpl3* to *rps10*) as well as nine additional genes upstream and three genes downstream. The ribosomal gene cluster is one that is conserved in all or most plastid genomes and retains traces of the cyanobacterial gene order (Stoebe and Kowallik 1999). The diatom-specific *ycf88* is found in the middle of the ribosomal gene cluster. All three IRs include cluster 14 and *trnP*, but the latter is not in the same orientation in the three genomes. In *T. pseudonana*, the IR also includes cluster 15, while in *O. sinensis* only the first gene of cluster 15 (*rpl32*) is included in the IR.

Parts of the 17 clusters are found in all red algal and chromist plastid genomes (underlined in Table 3, mapped in Fig. 2). In some cases, the clusters in the red alga *P. purpurea* and the cryptophyte *G. theta* are larger because they contain additional genes (e.g. cluster 2). In other cases, the corresponding clusters appear to be dispersed into several smaller clusters (e.g. the cluster 16 fragments a–d). Out of the 17 diatom clusters, *G. theta* shares six complete ones and fragments of several others, although the cluster order is different. Some *G. theta* clusters lack genes present in the diatoms, e.g., *thiG* and *thiS* (Table 2). The plastid genome of *E. huxleyi* is extensively rearranged, to the extent that the only clusters still found are small groups of


**Fig. 2** Plastid gene clusters conserved among the five chromists and the red alga *P. purpurea*. Genes present in the diatoms are indicated as *gray boxes*; while genes present in the other algae are *white boxes*. Gene clusters are surrounded by *larger black boxes*. The direction of the cluster is the transcriptional direction of the majority of the genes, and is defined with respect to the corresponding diatom cluster (Table 3). *Letters* (a, b, c, d) after cluster numbers refer to multiple fragments of diatom clusters. *Star* indicates that the entire cluster is present

genes conserved in all rhodophyte and chromist plastid genomes.

All three diatom genomes are rearranged with respect to each other, and with a few exceptions these rearrangements are in the LSC (Fig. 3). We used the GRIMM server (http://www-cse.ucsd.edu/groups/bioinformatics/GRIMM) to infer the number of inversions that could have given rise to them. It calculated a total of 22 inversions between the three LSCs and predicted a common intermediate order that would minimize the number of rearrangements to produce the three diatom orders (A in Fig. 3). Surprisingly, the two centric diatoms (*O. sinensis* and *T. pseudonana*) are no closer to each other than they are to the pennate diatom (*P. tricornutum*). There are 19 proposed inversions between *T. pseudonana* and *O. sinensis*, 16 between *T. pseudonana* and *P. tricornutum*, but only 9 between *O. sinensis* and *P. tricornutum*. The same kind of analysis conducted with another program (GRAPPA-IR, Cui et al. 2006a) resulted in a somewhat different inter-

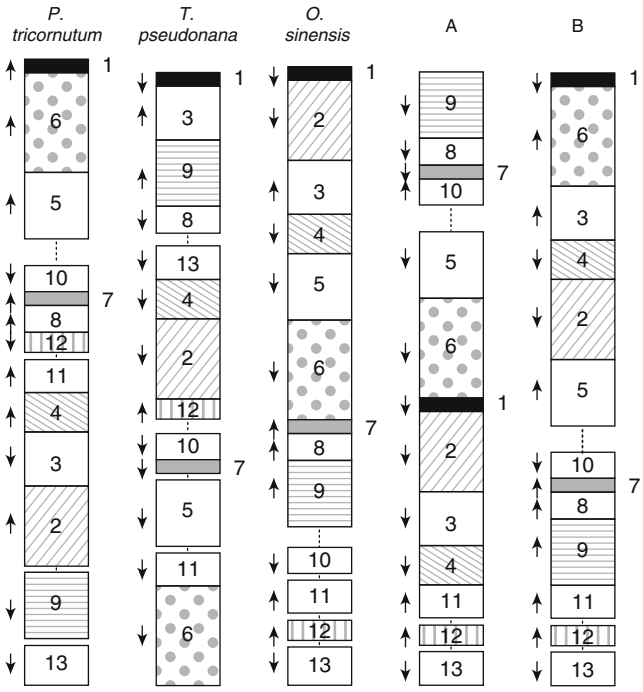


**Fig. 3** Schematic representation of diatom LSC region gene cluster arrangements, with two putative intermediate orders that could have give rise to them with the smallest number of inversions. The *dots* represent individual genes that are not part of a cluster and are not found in the same position in all the genomes. *A* Order derived by GRIMM; *B* order derived by GRAPPA-IR. *Cluster numbers* as in Table 3. *Arrows* show the directions of the clusters

mediate arrangement (B in Fig. 3) and similar numbers of inversions. However, it should be noted that these programs only consider inversion and no other mechanisms such as intermolecular recombination.

## Repeats, palindromes, 5′ and 3′ upstream sequences

All five chromist plastid genomes and four red algal plastid genomes were analyzed with the program PipMaker to find repeat sequences. In marked contrast to the chloroplast genomes of most green algae (Pombert et al. 2006; Maul et al. 2002), no direct repeats were found. With the programs etandem, palindrome and einverted of the EMBOSS suite, some regions of putative tandem repeats were identified, but most had less than 80% identity. The only kind of repeats we found were small inverted repeats, either true palindromes (no loop) or stem-loops (hairpins). These were frequently located at the 3′-ends of genes, where they are probably involved in transcription termination. A similar lack of repeats is also true of mesophilic red algal genomes. The one exception appears to be the chloroplast genome of *C. merolae* (Ohta et al. 2003), which contains two regions of tandem repeats, some of which are nested (data not shown).

**Table 4** Numbers of putative transcriptional and translational signals[a]

| Alga | Phage T7-like promoters | Bacterial promoters | Shine-Dalgarno sequences |
|---|---|---|---|
| *O. sinensis* | 3 | 23 | 53 |
| *P. tricornutum* | 3 | 20 | 41 |
| *T. pseudonana* | 3 | 35 | 59 |
| *E. huxleyi* | 9 | 11 | 30 |
| *G. theta* | 5 | 28 | 45 |
| *G. tenuistipitata* | 11 | 37 | 71 |
| *P. purpurea* | 13 | 35 | 83 |
| *C. caldarium* | 15 | 15 | 59 |
| *C. merolae* | 8 | 3 | 73 |

[a] See methods

The 150 nt upstream of each gene was searched for putative transcriptional promoters of the phage T7 type (ACTCACTA) and the bacterial type −35 (TTTAAA and TTGACA) and −10 (TATAAT) motifs. Both types of promoters were found (Table 4). In addition, motifs derived from the Shine-Dalgarno sequence were found within 50 nt upstream of a number of protein-coding genes. Table 4 also shows a similar analysis carried out on four red algae: *Gracilaria tenuistipitata* (Hagopian et al. 2004), *Porphyra purpurea* (Reith and Munholland 1995), *Cyanidium caldarium* (Glöckner et al. 2000) and *Cyanidioschyzon merolae* (Ohta et al. 2003). Alignments of some of the upstream sequences from all five chromists and the four red algae show conserved blocks at the nucleotide level in highly AT-rich sequences where small ORFs were annotated in *O. sinensis* (data not shown). We are aware that this is only a preliminary analysis, but it nonetheless reveals opportunities for testing hypotheses about gene expression by reverse genetics, because transformation of two diatom species has been achieved (Falciatore et al. 1999; Poulsen and Kröger 2005).

## Discussion

The plastid genomes of three diatoms, a cryptophyte and a haptophyte share many common features such as their size and gene complements. All the chromist and red algal genomes are very compact, lacking introns and short dispersed repeats. In contrast, genomes of the green lineage have much larger intergenic spacers, may have introns, and some have very large numbers of repeats, e.g., in *Chlamydomonas reinhardtii* there are “islands of genes in a sea of repeats” (Maul et al.

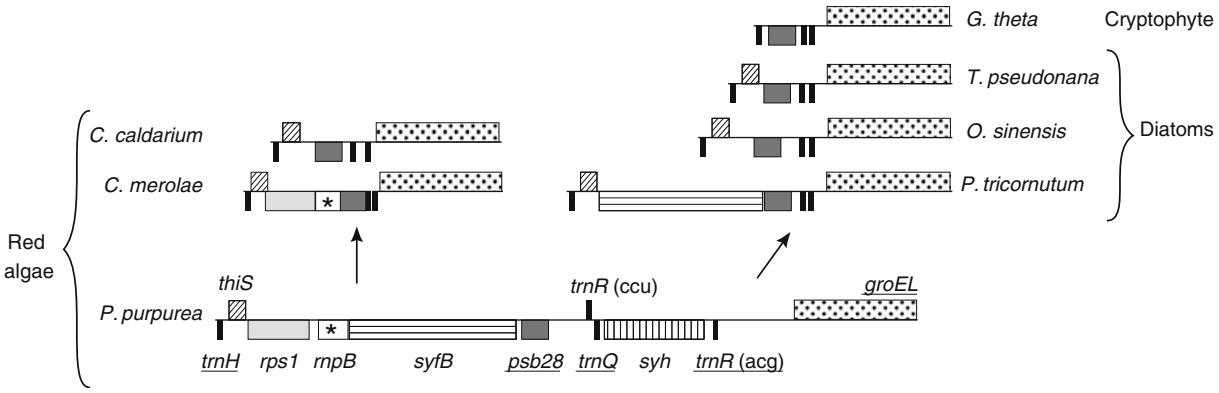


**Fig. 4** Independent losses of genes in a syntenic cluster. Genes are depicted as blocks with *different shadings*. Genes above the *horizontal black line* are transcribed toward the *right*, and those below the line to the *left*. The genes *syfB* and *rnpB* have been lost independently in the red algae and the diatoms

2002; Pombert et al. 2006). In terms of genome organization, the chromist chloroplast genomes have the quadripartite architecture usually found in the green lineage. In this they differ from the red algae, which have no IRs (Hagopian et al. 2004).

There have been substantial rearrangements in the diatom genomes, but they are restricted to either the LSC or the IR-SSC-IR regions and do not involve gene exchange between regions. The rearrangements of the conserved gene clusters are almost entirely restricted to the LSC. It is worth noting that the two centric diatoms are no more similar to each other than to the pennate diatom in terms of their cluster order. We used the programs GRIMM and GRAPPA-IR (Tesler 2002; Cui et al. 2006a), which calculate the number of inversions that would change one cluster order into another. The algorithms also predict an intermediate structure, which would require the fewest rearrangements to give the present order of the three genomes, although the authors are very careful not to imply that this is an ancestral order. The hypothetical intermediate orders calculated by the two programs were not the same, and did not resemble any gene order in red algal genomes, except for the association of clusters 4 and 11 that exists in *P. purpurea* but not in any of the three diatoms.

The clusters defined from the diatom plastid genomes are an assortment of remnants of well-known gene clusters, e.g., the ribosomal protein genes and the ATP cluster (Stoebe and Kowallik 1999) and diatom-specific rearrangements. They are likely the result of two possible mechanisms of genome evolution: deletion and relocation. Figure 4 illustrates an example of multiple deletions of the same genes from a cluster found in red algal and chromist genomes. This cluster consists of eleven genes in *P. purpurea*, but is much smaller in the red algae *C. caldarium* and *C. merolae*, due to several non-contiguous deletions (*syfB*, *trnR*(ccu), *syh*). *C. caldarium* has also lost *rps1* and *rnpB*. In the red alga *G. tenuistipitata* only one gene (*trnR*) has been deleted, but the cluster has been split and one half relocated 7 kb away on the other strand (Hagopian et al. 2004). In the diatoms, *rps1*, *rnpB*, *trnR* (ccu) and *syh* have been lost, and only *P. tricornutum* has retained *syfB*. Only five genes remain in the cryptophyte *G. theta* (*trnH*, *psb28*, *trnQ*, *trnR* and *groEL*). The *E. huxleyi* plastid genome does encode four genes of this cluster (the three *trns* and *groEL*), but only *trnQ* and *trnR* are still together (data not shown).

Figure 5 shows an example of relocation involving the gene *ffs*, which encodes the signal recognition particle RNA (Rosenblad and Samuelsson 2004). The large cluster to which it belongs in the red algae has been split up and the segment containing *ffs* and downstream genes (*psbX*, *psbV*) relocated next to *dnaB* and *trnF* in *O. sinensis* and *P. tricornutum*. There has been a second split in the line leading to *T. pseudonana*, leaving *ffs* as part of cluster 4 and losing the segment containing *psbX* and *psbV*.

It has been suggested that gene order can be used in wide-range phylogenetic studies (Cui et al. 2006b). However, our analysis of the diatom genomes suggests that the pathways of gene rearrangement and loss have been so complex that only a wide sampling of chromist plastid genomes would make rigorous analysis possible and allow estimation of the degree of saturation of this signal. Unfortunately, this means that gene content and gene order cannot currently be used to determine whether the chromists are indeed the result of a single secondary endosymbiotic event (Cavalier-Smith 2000; Yoon et al. 2002). One thing we can say is that there is no evidence for the involvement of repeat sequences in facilitating these rearrangements, because there are

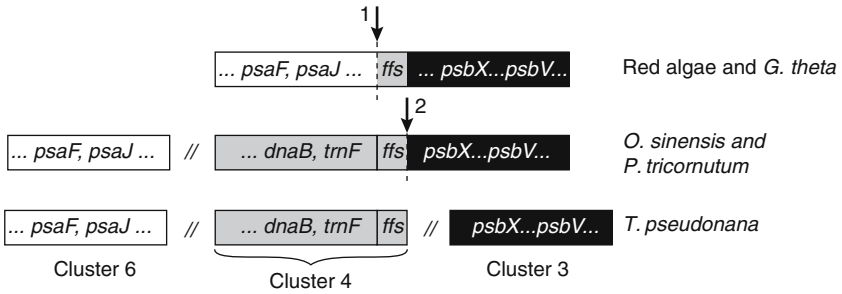


**Fig. 5** Two-step fragmentation and relocation of the *ffs* gene. Diatom gene clusters are depicted as blocks with *different shading*. The large cluster found in the red algae and *G. theta* has been split up once (*dotted line*, arrow 1) giving rise to the diatom cluster 4 (*gray*). A second split occurred in the line leading to *T. pseudonana* (*dotted line*, arrow 2), and cluster 3 (*black*) is no longer next to cluster 4. Double diagonal lines indicate that the clusters are physically separated. The region encompassing *psbX* and *psbV* in red algae includes genes lost in the diatom plastid genomes

few repeats in the diatom and red algal chloroplast genomes. This is completely different from the situation in the green lineage where repeats do appear to have played an important role in rearrangements (Maul et al. 2002; Pombert et al. 2006). Indeed, the number of gene rearrangements necessary to go from one genome order to another in green algae is much higher for fewer genes (Pombert et al. 2006).

Gene losses and rearrangements have obviously occurred independently in the three chromist lineages (Figs. 4, 5). Gene losses also occurred multiple times in the chlorophyte and streptophyte lineages (Turmel et al. 2006). Genes lost from plastid genomes may have been completely lost because they are no longer needed or they may have been replaced by duplicates of host genes that have acquired plastid targeting sequences. Alternatively, they may have been copied to the host nucleus, acquired plastid targeting presequences and subsequently been lost from the plastid genome. However, the successful integration of a gene into the nuclear genome does not necessarily lead to the loss of the plastid copy. For example, copies of the rubisco expression protein gene *cbbX* are found in both compartments in red algae and diatoms, and in the nucleomorph of *G. theta* (Maier et al. 2000). The case of the *groEL*/*cpn60* gene family is even more complex, involving several gene duplications and transfer of copies to the nucleus after the primary endosymbiosis (Zauner et al. 2006). It is clear that a variety of different processes are involved in determining the present-day gene complement of any plastid genome and that endosymbiogenesis is an on-going process in all photosynthetic eukaryotes.

**Acknowledgements** Financial support was provided by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) to BRG and the Centre National de Recherche Scientifique (CNRS) to CB. We thank Drs. I. Grigoriev, A. Kuo and R. Otillar at the Joint Genome Institute (Walnut Ceek, CA, USA) for help with sequence analysis, and Dr. Jijun Tang for the GRAPPA-IR analysis.

## References

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

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MD, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJ, Doggett N, Zharkikh A, Tavtigian SV, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, Tan YH, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, Brenner S (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297:1301–1310

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kroger N, Lau WW, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306:79–86

Bendich AJ (2004) Circular chloroplast chromosomes: the grand illusion. Plant Cell 16:1661–1666

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795

Bullerwell CE, Schnare MN, Gray MW (2003) Discovery and characterization of *Acanthamoeba castellanii* mitochondrial 5S rRNA. RNA 9:287–292

Cavalier-Smith T (2000) Membrane heredity and early chloroplast evolution. Trends Plant Sci. 5:174–182

Chen Z, Schneider TD (2005) Information theory based T7-like promoter models: classification of bacteriophages and differential evolution of promoters and their polymerases. Nucleic Acids Res 33:6172–6187

Cui L, Yue F, dePamphilis C, Moret BME, Tang J (2006a) Inferring ancestral chloroplast genomes with inverted repeats. In: Proceedings of the 2006 international conference on bioinformatics and computational biology (Biocomp'06), Las Vegas, NV, pp 75–81

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

Darling ACE, Mau B, Blattner FR, Perna NT (2004) Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14:1394–1403

de la Cruz J, Vioque A (2003) A structural and functional study of plastid RNAs homologous to catalytic bacterial RNase P RNA. Gene 321:47–56

Douglas SE, Penny SL (1999) The plastid genome of the cryptophyte alga, *Guillardia theta*: Complete sequence and conserved synteny groups confirm its common ancestry with red algae. J Mol Evol 48:236–244

Falciatore A, Casotti R, Leblanc C, Abrescia C, Bowler C (1999) Transformation of nonselectable reporter genes in marine diatoms. Mar Biotechnol 1:239–251

Falkowski PG, Katz ME, Knoll AH, Quigg A, Raven JA, Schofield O, Taylor FJR (2004) The evolution of modern eukaryotic phytoplankton. Science 305:354–360

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. Science 281:237–240

Gibbs SP (1981) The chloroplasts of some algal groups may have evolved from endosymbiotic eukaryotic algae. Ann NY Acad Sci 361:193–208

Glöckner G, Rosenthal A, Valentin K (2000) The structure and gene repertoire of an ancient red algal plastid genome. J Mol Evol 51:382–390

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8:195–202

Green BR (2004) The chloroplast genome of dinoflagellates—a reduced instruction set? Protist 155:23–31

Gueneau de Novoa P, Williams KP (2004) The tmRNA website: reductive evolution of tmRNA in plastids and other endosymbionts. Nucleic Acids Res 32:D104–D108

Gueneau P, Loiseaux-De Goër S, Williams KP (1999) The GC-rich region and *TψC trn* arm found in the *petF* region of the *Thalassiosira weissflogii* plastid genome encode a tmRNA. Eur J Phycol 34:533–535

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

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acid Symp Ser 41:95–98

Kashino Y, Lauber WM, Carroll JA, Wang Q, Whitmarsh J, Satoh K, Pakrasi HB (2002) Proteomic analysis of a highly active photosystem II preparation from the cyanobacterium *Synechocystis* sp. PCC6803 reveals the presence of novel polypeptides. Biochemistry 41:8004–8012

Kilian O, Kroth PG (2005) Identification and characterization of a new conserved motif within the presequence of proteins targeted into complex diatom plastids. Plant J 41:175–183

Kowallik KV, Stoebe B, Schaffran I, Kroth-Pancic P, Freier U (1995) The chloroplast genome of a chlorophyll a + c-containing alga, *Odontella sinensis*. Plant Mol Biol Rep 13:336–342

Laslett D, Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res 32:11–16

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

Maier U-G, Fraunholz M, Zauner S, Penny S, Douglas S (2000) A nucleomorph-encoded CbbX and the phylogeny of RuBisCo regulators. Mol Biol Evol 17:576–583

Matsuzaki M, Misumi O, Shin-I T, Maruyama S, Takahara M, Miyagishima SY, Mori T, Nishida K, Yagisawa F, Yoshida Y, Nishimura Y, Nakao S, Kobayashi T, Momoyama Y, Higashiyama T, Minoda A, Sano M, Nomoto H, Oishi K, Hayashi H, Ohta F, Nishizaka S, Haga S, Miura S, Morishita T, Kabeya Y, Terasawa K, Suzuki Y, Ishii Y, Asakawa S, Takano H, Ohta N, Kuroiwa H, Tanaka K, Shimizu N, Sugano S, Sato N, Nozaki H, Ogasawara N, Kohara Y, Kuroiwa T (2004) Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature 428:653–657

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

McFadden GI (2001) Primary and secondary endosymbiosis and the origin of plastids. J Phycol 37:951–959

Ohta N, Matsuzaki M, Misumi O, Miyagishima SY, Nozaki H, Tanaka K, Shin IT, Kohara Y, Kuroiwa T (2003) Complete sequence and analysis of the plastid genome of the unicellular red alga *Cyanidioschyzon merolae*. DNA Res 10:67–77

Pombert JF, Lemieux C, Turmel M (2006) 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 Biology 4:3

Poulsen N, Kröger N (2005) A new molecular tool for transgenic diatoms: control of mRNA and protein biosynthesis by an inducible promoter–terminator cassette. FEBS J 272:3413–3423

Reith M, Munholland J (1995) Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. Plant Mol Biol Rep 13:333–335

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

Rosenblad MA, Samuelsson T (2004) Identification of chloroplast signal recognition particle RNA genes. Plant Cell Physiol 45:1633–1639

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B (2000) Artemis: sequence visualization and annotation. Bioinformatics 16:944–945

Sánchez Puerta MV, Bachvaroff TR, Delwiche CF (2005) The complete plastid genome sequence of the haptophyte *Emiliania huxleyi*: a comparison to other plastid genomes. DNA Res 12:151–156

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

Seif ER, Forget L, Martin NC, Lang BF (2003) Mitochondrial RNase P RNAs in ascomycete fungi: lineage-specific variations in RNA secondary structure. RNA 9:1073–1083

Shi L-X, Schröder WP (2004) The low molecular mass subunits of the photosynthetic supracomplex, photosystem II. Biochim Biophys Acta 1608:75–96

Stoebe B, Kowallik KV (1999) Gene-cluster analysis in chloroplast genomics. Trends Genet 15:344–347

Tesler G (2002) GRIMM: genome rearrangements web server. Bioinformatics 18:492–493

Turmel M, Otis C, Lemieux C (2006) The chloroplast genome sequence of *Chara vulgaris* sheds new light into the closest green algal relatives of land plants. Mol Biol Evol 23:1324–1338

Yoon HS, Hackett JD, Pinto G, Bhattacharya D (2002) The single, ancient origin of chromist plastids. Proc Natl Acad Sci USA 15:15

Zauner S, Lockhart P, Stoebe-Maier B, Gilson P, McFadden GI, Maier UG (2006) Differential gene transfers and gene duplications in primary and secondary endosymbioses. BMC Evol Biol 6:38