J Mol Evol (1999) 48:236–244

JOURNAL OF MOLECULAR EVOLUTION

The Plastid Genome of the Cryptophyte Alga, *Guillardia theta:* Complete Sequence and Conserved Synteny Groups Confirm Its Common Ancestry with Red Algae

Susan E. Douglas, Susanne L. Penny

Institute for Marine Biosciences, 1411 Oxford Street, Halifax, Nova Scotia, Canada B3H 3Z1

Received: 16 June 1998 / Accepted: 20 August 1998

Abstract. The plastid genome of the cryptophyte alga *Guillardia theta* (121,524 bp) has been completely sequenced. The genome is 33% G+C and contains a short, nonidentical inverted repeat (4.9 kb) encoding the two rRNA cistrons. The large and small single-copy regions are 96.3 and 15.4 kb, respectively. Forty-six genes encoding proteins for photosynthesis, 5 genes for biosynthetic function, 5 genes involved in replication and division, 30 tRNA genes, 44 ribosomal protein genes (26 large subunit and 18 small subunit), 3 translation factors, 8 genes encoding components of the transcriptional machinery including 3 *ycfs* (hypothetical chloroplast frames), and 26 additional *ycfs* have been identified. There are eight ORFs larger than 50 amino acids, 3 of which have homologues on the plastid genome of the rhodophyte, *Porphyra purpurea* (Reith and Munholland 1995) and/or the *Synechocystis* genome (Kaneko et al. 1996) and can be designated new *ycfs.* Intergenic spacers are very short, no introns have been detected, and several genes overlap, all resulting in a very compact genome. In addition, large clusters of genes (such as those for the ribosomal proteins) are organized into single transcriptional units (Wang et al. 1997), again resulting in an economically organized genome. The cryptophyte plastid genome is almost completely comprised of clusters of genes that are found on the rhodophyte *Porphyra purpurea,* confirming its common ancestry with red algae. Furthermore, recombination events involving both tRNA genes and the rRNA cistrons appear to have been responsible for the structure of the cryptophyte plastid genome, including the formation of the inverted repeat.

Key words: Algae — Cryptophyte — DNA sequence — Endosymbiosis — Evolution — Genome — Plastid

Introduction

Cryptophytes are an enigmatic group of small biflagellate algae that share pigment characteristics with two distinct algal groups, the rhodophytes (phycobiliproteins) and the chromophytes (chlorophyll *c*). Like chromophytes, they harbour complex plastids surrounded by four membranes, rather than the two surrounding rhodophyte and chlorophyte plastids. However, they differ from chromophytes in possessing a small nucleus-like organelle (the nucleomorph) in the reduced space between the inner and outer plastid membrane pairs (Greenwood et al. 1997).

It has been proposed that organisms containing complex plastids arose by endosymbiosis of a photosynthetic eukaryote and a phagotrophic host with subsequent loss or reduction of eukaryotic features of the endosymbiont, such as the nucleus and cytoplasm (Gillott and Gibbs 1980). The resulting plastids would contain four membranes. Cryptophytes, by possessing vestiges of these eukaryotic features in the form of a nucleomorph and periplastidal space between the inner and the outer plastid membrane pairs, could be thought of as representing an intermediate en route to complex plastids.

Correspondence to: S.E. Douglas; *e-mail:* susan.douglas@nrc.ca

Ultrastructural data (Gibbs 1981), combined with molecular sequence data (Douglas et al. 1991; Douglas and Murphy 1994), provide strong evidence that cryptophyte algae arose by secondary endosymbiosis of a primitive eukaryotic rhodophyte. Recent phylogenetic analyses have reinforced the sister-group relationship between rhodophytes and nucleomorphs and, also, demonstrated an affiliation between cryptophyte hosts and glaucocystophytes (Bhattacharya and Medlin 1995; Van De Peer et al. 1996).

Plastid gene sequences have been utilized in phylogenetic analyses aimed at determining the relationships among eukaryotic photosynthetic lineages. However, problems encountered with substitutional bias caused by the relatively high A+T content of plastid genes (Lockhart et al. 1992), as well as varying mutational rates of different plastid genes or different sites within genes (Van De Peer et al. 1996) and the possibility of lateral gene transfers (Delwiche and Palmer 1996), have yielded conflicting results that may not reflect the true evolution of these lineages. Comparisons of gene order, on the other hand, offer a means of determining relationships among plastids that are not affected by these phenomena (Kowallik 1989, 1997; Wang et al. 1997).

The recent acquisition of complete genome sequences from the plastids of a number of green (Sugiura 1992; Hallick et al. 1993; Wakasugi et al. 1997) and nongreen photosynthetic eukaryotes (Kowallik et al. 1995; Reith and Munholland 1995; Stirewalt et al. 1995), as well as the cyanobacterium *Synechocystis* PCC6803 (Kaneko et al. 1996), allows new approaches to elucidating evolutionary relationships between algal lineages. In this context, it is of special interest to investigate the coding potential of the plastid of a chlorophyll *c*- and phycobilin-containing alga that may represent an intermediate stage in the evolution of complex plastids that have arisen by secondary endosymbiosis.

Analysis of the content of the *G. theta* plastid genome reveals strong similarities with that from the rhodophyte, *Porphyra purpurea* (Reith and Munholland 1995). Large stretches of DNA are conserved in gene order between the two plastids, although reduced in size in the cryptophyte. In many cases, tRNA genes are at the borders of the conserved stretches and adjacent to genes that have been deleted in the cryptophyte plastid, indicating that rearrangements have arisen through recombination between nonhomologous tRNA genes, as described in rice (Hiratsuka et al. 1989). In addition, recombination between the directly repeated rRNA cistons of the ancestral rhodophyte plastid appears to have resulted in the formation of the inverted repeat of the present-day cryptophyte plastid.

Materials and Methods

Guillardia theta, formerly designed *Cryptomonas* Φ (Hill and Wetherbee 1990), was cultivated as described (Douglas 1988), and plastid DNA was isolated by cesium chloride equilibrium centrifugation in the presence of Hoechst 33258 (Douglas 1988). The majority of the chloroplast genome was subcloned into pUC19 (Pharmacia) using a variety of restriction enzymes. A small portion of the genome that could not be cloned into pUC19 due to a lack of appropriate restriction enzyme sites was amplified by PCR using the high accuracy polymerase Pfu (Stratagene) and cloned into the vector pCR2.1 (Invitrogen). At least three clones of each amplification product were sequenced to reduce the possibility of PCR-generated artefacts. Template DNA was prepared using the Nucleobond AX kit (Machery Nagel) and sequencing was performed using an ABI 373A automated sequencer and the AmpliTaqFS dye terminator cycle sequencing ready reaction kit (Perkin Elmer). Specific oligonucleotide primers were used to fill gaps and complete the sequence of both strands. Sequence analysis and contig assembly was performed using Sequencher (Gene Codes, Inc.). Coding regions were identified by BLAST searches of GenBank (Gish and Gates 1993) and automated database searches were performed using MAGPIE (Gaasterland and Sensen 1996). Codon usage (based on known coding sequences including *ycf*s) was calculated using DNA Strider (Marck 1988).

Results and Discussion

Genome Organization. The circular plastid DNA of *Guillardia theta* is 121,524 bp, contains two small (approximately 4.9-kb) rRNA-containing inverted repeats, and encodes 183 genes (including the duplicated rRNA cistron genes) that are equally distributed on both strands (Fig. 1). It is the epitome of compactness (90% is coding sequence), exhibiting short A+T-rich intergenic spacers, no pseudogenes or introns, and four cases of overlapping genes. This is similar to the situation in the rhodophyte *P. purpurea* and the chromophyte *Odontella sinensis* but contrasts with green plants such as rice, where only 68% of the plastid genome is coding sequence (Hiratsuka 1989), the inverted repeats are much larger, and pseudogenes and introns are commonly found (see Sugiura 1992). All except six of the open reading frames (ORFs) have homologs on at least one other plastid genome. The ORFs without clear plastid homologues include *hlpA*, which encodes a histone-like protein (Wang and Liu 1991; Grasser et al. 1997) and appears to be unique to *G. theta*, and ORFs 53, 62, 65, 125, and 252. Other than the ribosomal RNAs, no genes for structural RNAs, such as the RNA component of RNase P (*rnpB*) that is present in *Cyanophora paradoxa* (Stirewalt et al. 1995) and *P. purpurea* (Reith and Munholland 1995), have been detected.

Like most plastids, the G+C content is low (33%), and interestingly, identical to that of *P. purpurea* (Reith and Munholland 1995). The codon usage reflects this bias, with codons ending in G and C comprising only 19% of the total. However, the highly expressed genes *psbA* and *rbcL* have a different codon bias that may be a result of selection for increased translation efficiency (Morton 1998). Alternatively, the distinct codon usage of these two genes could reflect their different origin resulting from horizontal gene transfer events. Codon usage of the five unidentified ORFs is similar to that of known coding regions, indicating that they are bona fide reading frames. The ochre termination codon TAA is used in 77% of cases, with amber and opal codons being used 15

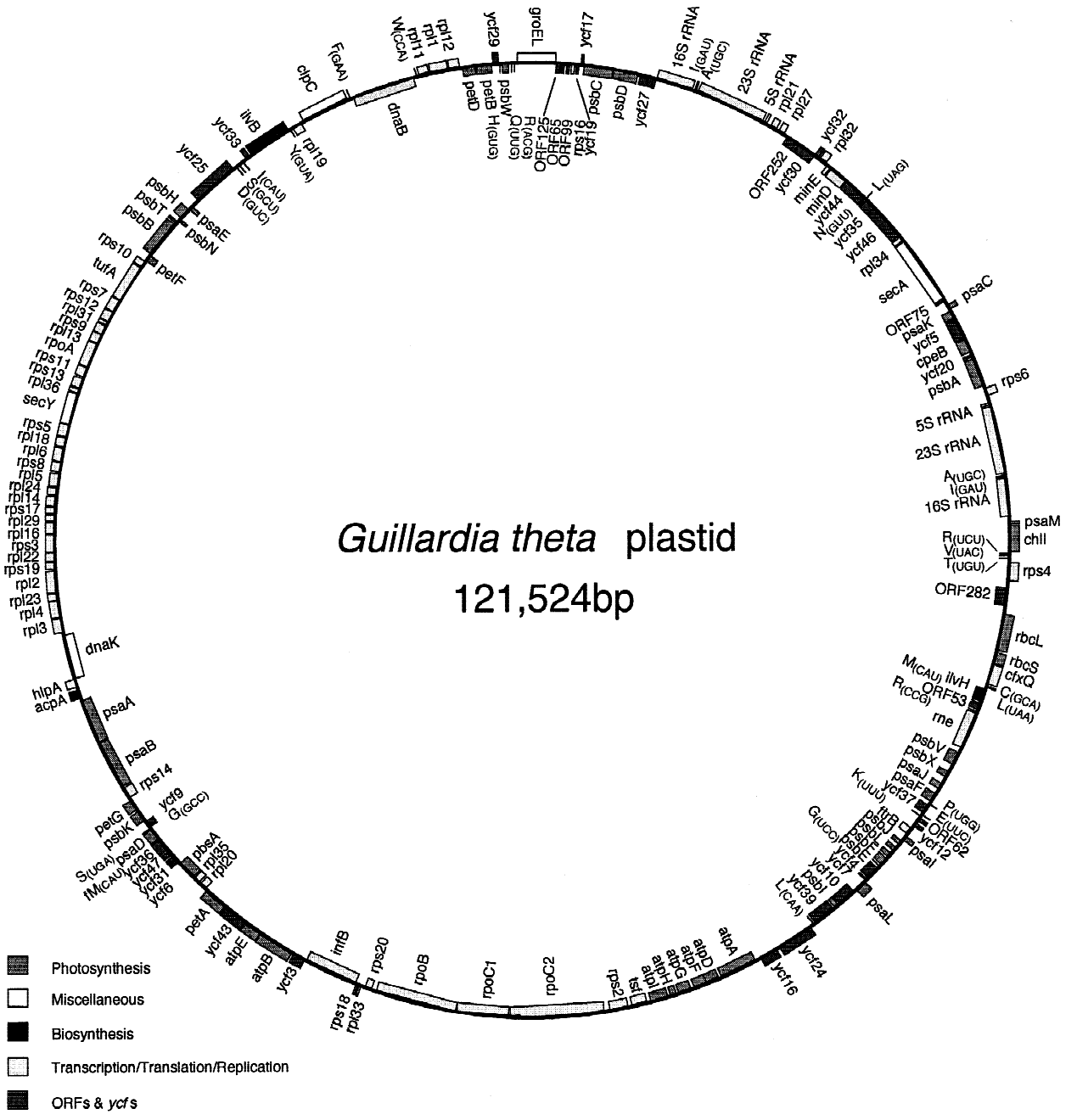

Fig. 1. Gene map of the plastid genome of *G. theta*. Genes transcribed from the plus strand are depicted outside the circle and those from the minus strand inside. Genes are shaded according to function as shown in the key.

and 8%, respectively. ATG is the predominant initiation codon, but GTG is used seven times and TTG once.

Transcription. Transcription in plastids is performed by two types of RNA polymerase, a multisubunit eubacteria-like plastic-encoded polymerase and a single-subunit phage T7-like nuclear-encoded polymerase (Allison et al. 1996). Four subunits of the eubacteria-like RNA polymerase (*rpoA, rpoB, rpoC1,* and *rpoC2*) are encoded on the *G. theta* plastid genome.

There is considerable evidence for polycistronic transcription of genes for related functions, not only from the presence of gene clusters on plastid genomes, but also from transcription studies (Douglas and Murphy 1994; Löffelhardt et al. 1997; Wang et al. 1997). Putative promoters similar to the canonical prokaryotic-type promoter sequences have been reported upstream of many *G. theta* plastid genes (Douglas et al. 1990; Douglas and Turner 1991; Douglas and Murphy 1994; Wang et al. 1997) and inspection of sequences upstream of polycistronic transcripts known from cyanobacteria and other plastids (see Löffelhardt et al. 1997) have revealed several more putative promoters. In all cases, the −10 TATAAT consensus is present, although the −35 con-

Table 1. Distribution of ycfs among photosynthetic lineages[a]

Name	Putative function	Synonym	Gr	*C.P.*	*O.s.*	*G.t.*	*P.p.*	*Syn*
ycf1	Hypothetical chloroplast RF1		+					
ycf2	Hypothetical chloroplast RF2	*ftsH* partially	+					
ycf3	Stable accumulation of PSI complex		+	+	+	+	+	+
ycf4	Stable accumulation of PSI complex		+	+	+	+	+	+
ycf5	Heme attachment to c-type cytochromes	*ccsA*	+	+	+	+	+	+
ycf6	Hypothetical chloroplast RF6		+	+	+	+	+	+
ycf7	Subunit of cytochrome b6f complex	*petI*	+	+	+	+	+	+
ycf8	PSII subunit req'd under stress	*psbT*	+	+	+	+	+	+
ycf9	Hypothetical chloroplast RF9		+	+	+	+	+	+
ycf10	Inorganic carbon uptake	*cotA/cemA*	+			+	+	+
ycf11	Acetyl CoA carboxylase beta subunit	*aceD/zfpA*	+				+	+
ycf12	Hypothetical chloroplast RF12		+	+	+	+	+	+
ycf13	Maturase-like protein	*matA*	+					
ycf14	Hypothetical chloroplast RF14 (intron)	*matK*	+					
ycf15	Hypothetical chloroplast RF15		+					
ycf16	ABC transporter subunit			+	+	+	+	+
ycf17	Similar to CAB/ELIP/HLIP protein			+		+	+	+
ycf18	Hypothetical chloroplast RF18	*nblA*					+	+
ycf19	Hypothetical chloroplast RF19					+	+	+
ycf20	Hypothetical chloroplast RF20					+	+	+
ycf21	Hypothetical chloroplast RF21			+			+	+
ycf22	Hypothetical chloroplast RF22						+	+
ycf23	Hypothetical chloroplast RF23			+			+	+
ycf24	ABC transporter subunit			+	+	+	+	+
ycf25	Homologous to *E. coli* protein *ftsH*		+		+	+	+	+
ycf26	*envZ* homolover putative His kinase	*dfr*					+	+
ycf27	*omp*R homolover, putative trp			+	+	+	+	+
ycf28	*ntcA* homolover, putative trp						+	+
ycf29	*tctD* homolover, putative up			+		+	+	+
ycf30	*lysR* homolover, putative trp	*rbcR*		+	+	+	+	+
ycf31	Cytochrome b6f complex subunit	*petM*		+	+	+	+	+
ycf32	Photosystem II thylakoid protein			+	+	+	+	+
ycf33	Hypothetical chloroplast RF33			+	+	+	+	+
ycf34	Hypothetical chloroplast RF34			+			+	+
ycf35	Hypothetical chloroplast RF35			+	+	+	+	+
ycf36	Hypothetical chloroplast RF36			+		+	+	+
ycf37	Hypothetical chloroplast RF37			+		+	+	+
ycf38	Hypothetical chloroplast RF38			+			+	+
ycf39	Hypothetical chloroplast RF39			+	+	+	+	+
ycf40	Hypothetical chloroplast RF40				+		+	+
ycf41	Hypothetical chloroplast RF41				+		+	+
ycf42	Hypothetical chloroplast RF42	*basI*			+		+	+
ycf43	Potential integral membrane protein	*yigU/ycbT*			+	+	+	+
ycf44	c-type holocytochrome formation	*ccs*			+	+	+	+
ycf45	Hypothetical chloroplast RF45				+		+	+
ycf46	Hypothetical chloroplast RF46				+	+	+	+
ycf47	Hypothetical chloroplast RF47				+	+	+	+
ycf61	Hypothetical chloroplast RF48	ORF75				+	+	+
ycf65	Hypothetical chloroplast RF49	ORF99b				+	+	+
ycf80	Hypothetical chloroplast RF50	ORF282				+	+	

[a] Lineages or their members are abbreviated as follows: green algae and land plants, Gr; *Cyanophora paradoxa, C.p.; Odontella sinensis, O.s.; Guillardia theta, G.t.; Porphyra purpurae, P.p.;* and *Synechocystis* PCC, 6803, *Syn.* Presence of a *ycf* is indicated by a + symbol. trp, transcriptional regulatory protein.

sensus is sometimes absent. It is possible that a nuclear-encoded polymerase transcribes those genes where canonical prokaryotic-type promoters are absent, although the 10-nucleotide consensus promoter sequence identified from tobacco (Hajdukiewicz et al. 1997) could not be detected.

Plastid gene expression in chloroplasts is regulated mainly at the posttranscriptional level (Danon 1997). However the presence of three potential genes with significant similarity to transcriptional regulatory proteins (*ycf27, ycf29, ycf30*) in *G. theta, P. purpurea, C. paradoxa*, and *O. sinensis* plastid genomes (Table 1) and a gene for an ATP-binding polypeptide involved in the expression of Rubisco (*cfxQ*) in all except *C. paradoxa* (Table 2) indicates that at least some gene expression occurs by transcriptional regulation. Ribonuclease E

Table 2. Distribution of genes among nongreen plastid genomes[a]

Name	*C.p.*	*O.s.*	*G.t.*	*P.p.*
ATP synthase				
atpA	+	+	+	+
atpB	+	+	+	+
atpD	+	+	+	+
atpE	+	+	+	+
atpF	+	+	+	+
atpG	+	+	+	+
atpH	+	+	+	+
atpI		+	+	+
Photosystem I				
psaA	+	+	+	+
psaB	+	+	+	+
psaC	+	+	+	+
psaD		+	+	+
psaE	+	+	+	+
psaF	+	+	+	+
psaI	+	+	+	+
psaJ	+	+	+	+
psaK			+	+
psaL		+	+	+
psaM	+	+	+	+
Photosystem II				
psbA	+	+	+	+
psbB	+	+	+	+
psbC	+	+	+	+
psbD	+	+	+	+
psbE	+	+	+	+
psbF	+	+	+	+
psbH	+	+	+	+
psbI	+	+	+	+
psbJ	+	+	+	+
psbK	+	+	+	+
psbL	+	+	+	+
psbN	+	+	+	+
psbT(ycf8)	+	+	+	+
psbV	+	+	+	+
psbW	+	+	+	+
psbX	+	+	+	+
Rubisco				
rbcL	+[b]	+	+	+
rbcS	+[b]	+	+	+
Phycobiliproteins				
apcA	+			+
apcB	+			+
apcD	+			+
apcE	+			+
apcF	+			+
cpcA	+			+
cpcB	+			+
cpcG	+			+
cpeA				+
cpeB			+	+
Electron transfer				
petA	+	+	+	+
petB	+	+	+	+
petD	+	+	+	+
petF	+	+	+	+
petG	+	+	+	+
petL (ycf7)	+	+	+	+
petM (ycf31)	+	+	+	+
ftrB			+	+
Miscellaneous				
clpC		+	+	+
dnaB		+	+	+

Table 2. Continued

Name	*C.p.*	*O.s.*	*G.t.*	*P.p.*
dnaK	+	+	+	+
groEL	+	+	+	+
secA		+	+	+
secY	+	+	+	+
Ribosomal proteins				
rpl1	+	+	+	+
rpl2	+	+	+	+
rpl3	+	+	+	+
rpl4		+	+	+
rpl5	+	+	+	+
rpl6	+	+	+	+
rpl9				+
rpl11	+	+	+	+
rpl12	+	+	+	+
rpl13		+	+	+
rpl14	+	+	+	+
rpl16	+	+	+	+
rpl18	+	+	+	+
rpl19	+	+	+	+
rpl20	+	+	+	+
rpl21	+	+	+	+
rpl22	+	+	+	+
rpl23		+	+	+
rpl24		+	+	+
rpl27		+	+	+
rpl28	+			+
rpl29		+	+	+
rpl31		+	+	+
rpl32		+	+	+
rpl33	+	+	+	+
rpl34	+	+	+	+
rpl35	+	+	+	+
rpl36	+	+	+	+
rps1				+
rps2	+	+	+	+
rps3	+	+	+	+
rps4	+	+	+	+
rps5	+	+	+	+
rps6	+	+	+	+
rps7	+	+	+	+
rps8	+	+	+	+
rps9	+	+	+	+
rps10	+	+	+	+
rps11	+	+	+	+
rps12	+	+	+	+
rps13	+	+	+	+
rps14	+	+	+	+
rps16	+	+	+	+
rps17	+	+	+	+
rps18	+	+	+	+
rps19	+	+	+	+
rps20	+	+	+	+
Transcription/RNA processing				
cfxQ		+	+	+
rne			+	+
rnpB	+			+
rpoA	+	+	+	+
rpoB	+	+	+	+
rpoCl	+	+	+	+
rpoC2	+	+	+	+
Translation				
infB			+	+

Table 2. Continued

Name	*C.p.*	*O.s.*	*G.t.*	*P.p.*
tsf			+	+
tufA	+	+	+	+
Biosynthesis				
acpA	+	+	+	+
chlB	+			+
chlI	+	+	+	+
chlL	+			+
ChlN	+			+
ilvB			+	+
ilvH			+	+
pbsA			+	+
preA	+			+
trpG	+			+

[a] Except for ribosomal proteins, genes unique to a single genome are not shown. Abbreviations are as in Table 1.
[b] The rubisco genes of the *C. paradoxa* plastid are not homologous to those of the other plastids.

(*rne*), also encoded on the plastid genome of *G. theta*, may participate in posttranscriptional degradation of mRNAs.

Translation. Many components of the translational apparatus are present, including the 3 rRNA molecules, 1 initiation factor (*infB*), two elongation factors (*tsf* and *tufA*), 26 genes for 50S ribosomal subunit proteins, and 18 genes for 30S ribosomal subunit proteins. Most of these ribosomal protein genes are found in a large cluster which is conserved to different degrees in different photosynthetic lineages and has been a found to be a useful character for phylogenetic reconstruction (Sugita et al. 1997; Wang et al. 1997). In addition, several other highly conserved ribosomal protein gene clusters are present (*rpl11/1/12*, *rpl33/rps18*, *rpl20/35*, and *rpl21/27*). Thirty tRNAs are present, two of which (*trnI* and *trnA*) are duplicated in the inverted repeats. This suite of tRNAs allows the decoding of all 61 sense codons.

Photosynthesis. All of the components of the ATP synthase with the exception of *atpC*, which was transferred to the nucleus very early in the evolution of plastids (Pancic et al. 1992; Kowallik 1997), are present on the *G. theta* plastid genome. As in both *P. purpurea* and *O. sinensis*, there is an overlap of four nucleotides between the *atpF* and the *atpD* genes. Interestingly, these two genes overlap by a single nucleotide in the cyanobacterium *Synechococcus* PCC 6301 (Cozens and Walker 1987).

Seven components of the electron transfer chain are also found (*petA, B, D, F,* and *G*), including the recently identified *petL* [formerly designated *ycf7* (Naithani et al. 1997)] (Table 1) and *petM* [formerly designated *ycf31* (de Vitry et al. 1996)] (Table 1). With the exception of *psbM*, which has been identified only on the *C. paradoxa* plastid genome, the complete suite of 28 photosystem I and II genes is present on the *G. theta* plastid genome. Also present is the beta subunit of ferredoxin thioreductase (*ftrB*), which participates in electron transfer and regulates several photosynthetic enzymes. In addition, genes for both subunits of Rubisco (*rbcL* and *rbcS*), the beta subunit of phycoerythrin (*cpeB*), and two genes involved in chlorophyll biosynthesis—a magnesium chelatase (*chlI*) and heme oxygenase (*pbsA*)—are found on the *G. theta* plastid genome. *ycf17*, which encodes a protein similar to members of the CAB/ELIP/HLIP family, is also present.

Replication and Cell Division. Although plastids lack histones, there is evidence for chromatin-associated proteins (see Grasser et al. 1997). One such protein, encoded by *hlpA* (Wang and Liu 1991), is thought to perform an architectural role in the plastid nucleoid (Grasser et al. 1997). In addition, *dnaB*, encoding a DNA helicase, and three other genes that participate in cell division (*minD, minE, ftsH*) that were recently reported from the plastid genome of the green alga *Chlorella vulgaris* C-27 (Wakasugi et al. 1997) have been identified on the *G. theta* plastid genome. It is interesting that *minD* and *minE* have not been identified in any other nongreen algae and *hlpA* is unique among all sequenced plastids. This may indicate that the cryptophyte endosymbiont represents a primitive stage in the evolution of the nongreen lineages, just as *C. vulgaris* represents a primitive stage in the green lineage.

Miscellaneous Functions. A number of genes involved in protein metabolism or transport are encoded on the plastid genome of *G. theta.* These include *secA* and *secY*, which are components of the sec protein translocation system, *groEL* and *dnaK* (chaperonin subunits), and *clpC* (the ATP binding subunit of the Clp protease).

Conserved Reading Frames (ycfs). Of the 47 reading frames that are conserved between at least two plastid genomes and are designated by the Commission on Plant Gene Nomenclature (Hallick and Bairoch 1994) as *ycfs* (hypothetical chloroplast frames), 29 have been identified on the plastid of *G. theta.* The distribution of these among various plastid groups and the functions of those that are known are listed in Table 1. In addition, *G. theta* ORFs 76, 99, and 282 are homologous to ORFs 75, 99b, and 450 of *P. purpurea* and are now designated *ycfs* 61, 65, and 80, respectively (Stoebe, personal communication).

A Conserved Intein. The *dnaB* genes from both *G. theta* and *P. purpurea* contain an additional stretch of protein-encoding sequence that is spliced out of the mature polypeptide, much as an intron is spliced out of mature mRNA. Inteins in the *dnaB* gene are relatively rare, being found only in two eubacteria, *Synechocystis* PCC6803 and *Rhodothermus marinus* (Liu and Hu 1997). The only other inteins known to occur in plastids are in the *clpP* genes of *Chlamydomonas reinhardtii* and *C. eugametos* (Huang et al. 1994). The *G. theta dnaB* intein is 160 amino acids long, whereas the *P. purpurea*

intein is 150 amino acids long. In both organisms, the codon usage of the intein is very similar to that of the exteins, indicating that it is quite ancient and the codon usage has become homogenized over time.

The Inverted Repeat. Examination of the regions flanking the cryptophyte *rrnB* cistron shows that it contains genes from the upstream region of *rrnB* (*trnV, trnR, chlI,* and *psaM*) and the downstream region of *rrnA* (*rps6*) of *P. purpurea* (Fig. 2). The reciprocal arrangement (involving 36.9 kb of sequence upstream of the *P. purpurea rrnA* cistron) is evident in the crytophyte *rrnA* cistron. It is highly likely that the inverted repeat structure of the cryptophyte plastid has resulted from a reciprocal recombination event within the rRNA cistron.

The inverted repeats of *G. theta* are not identical in sequence. There is one transition substitution in the SSU rRNA gene, two in the $tRNA^{Ala}$–LSU rRNA intergenic-spacer, three in the LSU rRNA gene, one in the LSU rRNA–5S rRNA intergenic spacer, and two in the 5S rRNA gene. None of the substitutions affect secondary structure of the mature rRNA molecules. The regions upstream of the 16S rRNA genes are well conserved (86% similar) for 111 bp, presumably to preserve the region surrounding the promoter. However sequence similarity stops immediately downstream of the 5S rRNA gene. In *P. purpurea,* the coding sequences are much more variable than in *G. theta,* with 41 of 4820 positions differing (Reith and Munholland 1995), and the flanking sequences diverge within seven nucleotides of the 16S rRNA and within two of the 5S rRNA (Reith and Munholland 1993). This implies that the copy-correction mechanism that ensures identity of repeats in land plant chloroplasts, but is apparently absent in *P. purpurea,* may be only partially developed in *G. theta.* Similarly, the expansion of the inverted repeat by gene conversion, which has occurred to differing extents in other lineages, may have occurred to a limited extent in *G. theta* since the region upstream of the SSU rRNA is conserved for a short distance.

Synteny Groups. The entire plastid genome of *G. theta* is comprised of synteny groups that are present in *P. purpurea* (Fig. 3; junctions marked by arrowheads). Three of these synteny groups are very large (two are over 30 kb and one is 17 kb). In all cases, the gene order and transcriptional orientation are conserved, but some genes present on the plastid genome of *P. purpurea* (usually those involved in biosynthesis, phycobiliprotein synthesis or ORFs of unknown function) have been deleted from the plastid genome of *G. theta* (Fig. 4). In fact there are only five genes involved in biosynthesis remaining on the *G. theta* plastid genome (*ilvB, ilvH, pbsA, chlI,* and *acpA*), a single subunit of phycoerythrin (*cpeB*), and five ORFs of unknown function.

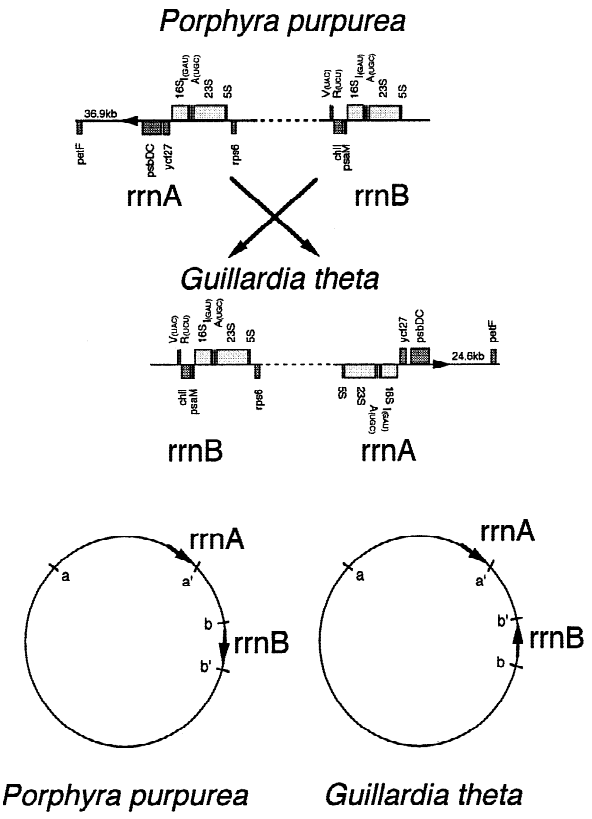

Fig. 2. Recombination between ribosomal RNA cistrons. A reciprocal crossover event between *rrn*A and *rrn*B of *P. purpurea* results in the exchange of flanking sequences and the resulting arrangement seen in *G. theta.* Deletion of several genes has resulted in the reduction in size of the *rrn*A flanking region from 36.9 kb in *P. purpurea* to 24.6 kb in *G. theta.* Borders of *rrn*A and *rrn*B are represented by a/a′ and b/b′, respectively.

In many cases, tRNA genes are present at the junctions of the synteny groups (asterisks; Fig. 3), suggesting that they may have participated in the deletion of gene sequences, possibly by acting as recognition signals (Hiratsuka et al. 1989). In addition, there are sixteen instances where tRNA genes are also found adjacent to *P. purpurea* genes that have been deleted from *G. theta.* Figure 4B shows one such region of the *P. purpurea* genome where three deletions have occurred relative to *G. theta,* all of which are adjacent to tRNA genes.

Evolutionary Implications. This is the first plastid genome to be sequenced form a nucleomorph-containing organism and as such it is of interest to compare its coding capacity with that of other algae that have arisen by secondary endosymbiosis but do not contain a nucleomorph. The diatom *O. sinensis* is one such example that has been completely sequenced (Kowallik et al. 1995). Although some synteny groups are shared between *G. theta* and *O. sinensis* (large ribosomal protein, *atpA, rpoBC1C2, psbBTNH* gene clusters), none are as large or as striking as those shared between *G. theta* and *P. purpurea.* There has been much more rearrangement (Fig. 4), indicating either that a longer period of evolution has

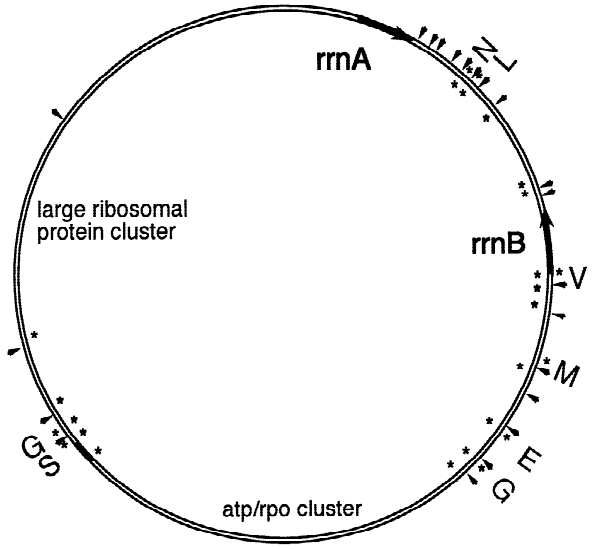

Fig. 3. Synteny groups present on the *G. theta* plastid genome. Junctions between conserved gene clusters are indicated by arrows. Asterisks inside the circle represent tRNA genes present at the junctions on the *P. purpurea* plastid genome and those outside the circle (with letters) represent those at the junctions on the *G. theta* plastid genome. The two rRNA cistrons are represented by shaded arrows and an internal inversion of the cluster *ycf*6/31/47/36/*trn*fMet within the larger synteny group is indicated by shading.

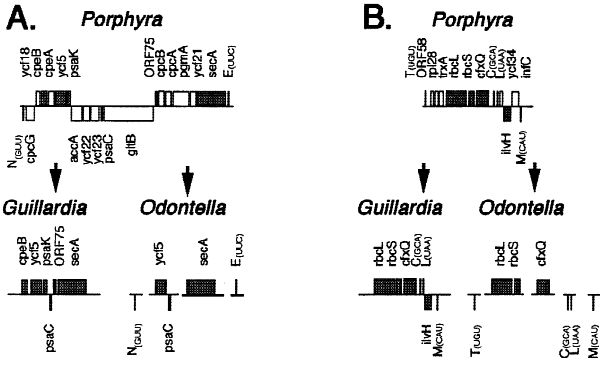

Fig. 4. **(A)** Gene arrangement between $tRNA^N$ and $tRNA^E$ genes of *P. purpurea*, *O. sinensis*, and *G. theta*. **(B)** Gene arrangement between $tRNA^T$ and $tRNA^M$ genes of *P. purpurea*, *O. sinensis*, and *G. theta*. Genes present in both *P. purpurea* and *G. theta*. Genomes are represented by shaded boxes and those that have been deleted from *G. theta* by empty boxes. Additional genes have been deleted from *O. sinensis*. Genes transcribed from the plus strand are depicted above the line and those from the minus strand below.

passed since the *O. sinensis* endosymbiont was established than for the *G. theta* endosymbiont or that the *O. sinensis* and *G. theta* endosymbionts were different. Alternatively, the presence of genes essential for plastid function in the nucleomorph have stabilized the plastid genome such that rearrangements do not occur to as great an extent in nucleomorph-containing organisms. That possibility is now under investigation although preliminary results indicate that there are very few genes for plastid-localized products in the nucleomorph (McFadden et al. 1997).

With the availability of complete plastid genome sequences, analysis of the distribution of gene clusters has increasingly been used for phylogenetic reconstruction (Kowallik 1997). Of particular interest are clusters that are widely separated on cyanobacterial genomes but appear to have fused subsequent to endosymbiosis and are present in plastid genomes from several lineages. Such arrangements provide very strong evidence for the monophyletic origin of plastids. For such similar organization in the plastid genomes of separate lineages to result from different endosymbionts, an extreme degree of convergent evolution would have to be invoked (Douglas 1994). Well-studied examples that have helped elucidate phylogenetic relationships between algal lineages include the large ribosomal protein cluster (Wang et al. 1997), the rRNA cistrons (Reith and Munholland 1993), the *rpoBC/atpA* cluster (Pancic et al. 1992; Kowallik 1997), and the *psbBTNH* cluster (Douglas 1994).

Three significant features suggest that the ancestor of the *G. theta* plastid closely resembled a rhodophyte plastid like that of *P. purpurea*. First, the conserved synteny groups, which are identical in gene order but reduced in gene content to stretches of the *P. purpurea* plastid genome, give strong evidence for a common ancestry. Second, both *G. theta* and *P. purpurea* contain an intein in their plastid *dnaB* genes. Given the rarity of inteins in plastid genes in general, and *dnaB* genes in particular, this is a significant shared character. Third, the inverted repeat of *G. theta* appears to have arisen by reciprocal recombination from the nonidentical, directly repeated rRNA cistrons of *P. purpurea* (interpreted by the authors as being a primitive feature) (Reith and Munholland 1993). Our results greatly strengthen their suggestions that the ancestral plastid may have had two direct, nonidentical rRNA repeats that were either reorganized into the inverted pattern seen in many land plants, glaucophytes, rhodophytes, cryptophytes, and chromophytes or reduced to a single copy in some chlorophytes and rhodophytes.

Acknowledgments. The sequence reported in this paper has been deposited in the GenBank database (accession No. AF041468) and is also available as a Magpie project at http://niji.imb.nrc.ca/magpie/plastid. The authors thank Michael Reith and Mark Ragan for comments on this manuscript and Gertraud Buger (Organelle Genome Megasequencing Project, University of Montreal) for assistance in annotation of the sequence for submission to Genbank. We are very grateful for the help of Paul Gordon, Christoph Sensen and Terry Gaasterland in the implementation of Magpie. This is NRCC publication No. 39789.

References

Allison LA, Simon LD, Maliga P (1996) Deletion of *rpoB* reveals a second distinct transcription system in plastids of higher plants. EMBO J 15:2802–2809

Bhattacharya D, Medlin L (1995) The phylogeny of plastids: A review based on comparisons of small-subunit ribosomal RNA coding regions. J Phycol 31:489–498

Cozens AL, Walker JE (1987) The organization and sequence of the genes for ATP synthase subunits in the cyanobacterium *Synechococcus* 6301. J Mol Biol 294:359–383

Danon A (1997) Translational regulation in the chloroplast. Plant Physiol 115:1293–1298

Delwiche CF, Palmer JD (1996) Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids. Mol Biol Evol 13:873–882

de Vitry C, Breyton C, Pierre Y, Popot JL (1996) The 4-kDa nuclear-encoded PetM polypeptide of the chloroplast cytochrome b6f complex. Nucleic acid and protein sequences, targeting signals, transmembrane topology. J Biol Chem 271:10667–10671

Douglas SE (1988) Physical mapping of the plastid genome from the chlorophyll c-containing alga, *Cryptomonas* Φ. Curr Gen 14:591–598

Douglas SE (1994) Chloroplast origins and evolution. In: Bryant DA (ed) The molecular biology of cyanobacteria. Kluwer, Amsterdam, pp 91–118

Douglas SE, Durnford DG, Morden CW (1990) Nucleotide sequence of the gene for the large subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase from *Cryptomonas* Φ: Evidence supporting the polyphyletic origin of plastids. J Phycol 26:500–508

Douglas SE, Murphy CA (1994) Structural, transcriptional, and phylogenetic analyses of the *atpB* gene cluster from the plastid of *Cryptomonas* Φ (Cryptophyceae). J Phycol 30:329–340

Douglas SE, Murphy CA, Spencer DF, Gray MW (1991) Molecular evidence that cryptomonad algae are evolutionary chimaeras of two phylogenetically distinct unicellular eukaryotes. Nature 350:148–151

Douglas SE, Turner S (1991) Molecular evidence for the origin of plastids from a cyanobacterium-like ancestor. J Mol Evol 33:267–273

Gaasterland T, Sensen C (1996) Fully automated genome analysis that reflects user needs and preferences: A detailed introduction to the MAGPIE system architecture. Biochimie 78:302–310

Gibbs SP (1981) The chloroplast endoplasmic reticulum: structure, function, and evolutionary significance. Int Rev Cytol 72:49–99

Gillott MA, Gibbs SP (1980) The cryptomonad nucleomorph: Its ultrastructure and evolutionary significance. J Phycol 16:558–568

Gish W, Gates DJ (1993) Identification of protein coding regions by database similarity search. Nature Genet 3:226–272

Grasser KD, et al. (1997) The recombinant product of the *Cryptomonas* Φ plastid gene, *hlpA* is an architectural HU-like protein that promotes the assembly of complex nucleoprotein structures. Eur J Biochem 249:70–76

Greenwood AD, Griffiths HB, Santore UJ (1977) Chloroplasts and cell compartments in Cryptophyceae. Brit Phycol J 12:119

Hajdukiewicz PTJ, Allison LA, Maliga P (1997) The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids. EMBO J 16: 4041–4048

Hallick RB, Bairoch A (1994) Proposals for naming of chloroplast genes. III. Nomenclature for open reading frames encoded in chloroplast genomes. Plant Mol Biol Reptr 12:S29–S30

Hallick RB, et al. (1993) Complete sequence of *Euglena gracilis* chloroplast DNA. Nucleic Acids Res 21:3537–3544

Hill DRA, Wetherbee R (1990) *Guillardia theta* new-genus new-species Cryptophyceae. Can J Bot 68:1873–1876

Hiratsuka J, et al. (1989) The complete sequence of the rice (*Oryza sativa*) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of cereals. Mol Gen Genet 27:185–194

Huang C, et al. (1994) The *Chlamydomonas* chloroplast *clpP* gene contains translated large insertion sequences and is essential for cell growth. Mol Gen Genet 244:151–159

Kaneko T, et al. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res 3:109–136.

Kowallik KV (1989) Molecular aspects and phylogenetic implications of plastid genomes of certain chromophytes. In: Green JC et al. (ed) The chromophyte algae—problems and perspectives. Clarendon Press, Oxford, pp 101–124

Kowallik KV (1997) Origin and evolution of chloroplasts: Current status and future perspectives. In: Schenk HEA, et al. (eds) Eukaryotism and symbiosis. Springer Verlag, Berlin, Heidelberg, pp 3–23

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

Liu XQ, Hu Z (1997) A DnaB intein in *Rhodothermus marinus:* Indication of recent intein homing across remotely related organisms. Proc Natl Acad Sci USA 94:7851–7856

Lockhart PJ, Howe CJ, Bryant DA, Beanland TJ, Larkum AWD (1992) Substitutional bias confounds inference of cyanelle origins from sequence data. J Mol Evol 34:153–162

Löffelhardt W, Bohnert HJ, Bryant DA (1997) The complete sequence of the *Cyanophora paradoxa* cyanelle genome (*Glaucocystophyceae*). In: Bhattacharya D (ed). Springer, Wien, New York, pp 149–162

Marck C (1988) ''DNA Strider'': A ''C'' program for the fast analysis of DNA and protein sequences on Apple Macintosh family of computers. Nucleic Acids Res 16:1829–1835

McFadden GI, et al. (1997) Bonsai genomics: Sequencing the smallest eukaryotic genome. Trends Genet 13:46–49

Morton BR (1998) Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. J Mol Evol 46:449–459

Naithani S, Trivedi PK, Sane PV (1997) Characterization of the *orf31-petG* gene cluster from the plastid genome of *Populus deltoides.* Biochem Mol Biol Int 43:433–442

Pancic PG, Strotmann H, Kowallik KV (1992) Chloroplast ATPase genes in the diatom *Odontella sinensis* reflect cyanobacterial characters in structure and arrangement. J Mol Biol 224:529–536

Reith ME, Munholland J (1993) The ribosomal RNA repeats are nonidentical and directly oriented in the chloroplast genome of the red alga *Porphyra purpurea.* Curr Genet 24:443–450

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

Stirewalt VL, Michalowski CB, Loffelhardt W, Bohnert HJ, Bryant DA (1995) Nucleotide sequence of the cyanelle genome from *Cyanophora paradoxa.* Plant Mol Biol Rptr 13:327–332

Sugita M, et al. (1997) Organization of a large gene cluster encoding ribosomal proteins in the cyanobacterium *Synechococcus* sp. strain PCC 6301: Comparison of gene clusters among cyanobacteria, eubacteria and chloroplast genomes. Gene 195:73–79

Sugiura M (1992) The chloroplast genome. Plant Mol Biol 19:149–168

Van De Peer Y, Rensing SA, Maier UG, De Wachter R (1996) Substitution rate calibration of small subunit ribosomal RNA identifies chlorarachniophyte endosymbionts as remnants of green algae. Proc Natl Acad Sci USA 93:7732–7736

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

Wang SL, Liu X-Q (1991) The plastid genome of *Cryptomonas* Φ encodes an hsp70-like protein, a histone-like protein, and an acyl carrier protein. Proc Natl Acad Sci USA 88:10783–10787

Wang SL, Liu X-Q, Douglas SE (1997) The large ribosomal gene cluster of a cryptomonad plastid: Gene organization, sequence and evolutionary implications. Biochem Mol Biol Int 41:1035–1044