Lateral transfer of introns in the cryptophyte plastid genome

Hameed Khan and John M. Archibald*

Canadian Institute for Advanced Research, Program in Integrated Microbial Biodiversity, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada

Received October 17, 2007; Revised January 17, 2008; Accepted February 16, 2008

ABSTRACT

Cryptophytes are unicellular eukaryotic algae that acquired photosynthesis secondarily through the uptake and retention of a red-algal endosymbiont. The plastid genome of the cryptophyte Rhodomonas salina CCMP1319 was recently sequenced and found to contain a genetic element similar to a group II intron. Here, we explore the distribution, structure and function of group II introns in the plastid genomes of distantly and closely related cryptophytes. The predicted secondary structures of six introns contained in three different genes were examined and found to be generally similar to group II introns but unusually large in size (including the largest known noncoding intron). Phylogenetic analysis suggests that the cryptophyte group II introns were acquired via lateral gene transfer (LGT) from a euglenid-like species. Unexpectedly, the six introns occupy five distinct genomic locations, suggesting multiple LGT events or recent transposition (or both). Combined with structural considerations, RT–PCR experiments suggest that the transferred introns are degenerate ‘twintrons’ (i.e. nested group II/group III introns) in which the internal intron has lost its splicing capability, resulting in an amalgamation with the outer intron.

INTRODUCTION

Group II introns are a type of retroelement found in bacterial and eukaryotic organellar genomes, and are generally believed to be the ancestors of spliceosomal introns and non-long terminal repeat (non-LTR) retrotransposons (1–3). These introns are transcribed into catalytic RNAs that are capable of splicing themselves from pre-mRNA with the assistance of proteins (4). The transcribed intron RNA forms a secondary structure comprised of six domains (D1–D6) that extend from a central core (5). Domain I (D1) is the largest noncoding domain and is believed to be involved in RNA catalysis, while domain V (D5) is thought to be the catalytic core of the ribozyme and is highly conserved in sequence (6,7). The function of Domain II (D2) remains unclear, although it is thought to have tertiary interactions with Domain VI (D6) and D1 (8,9). Domain III (D3) appears to play a role in splicing, since deletion of D3 has been shown to impair in vitro catalysis in cis (10). If delivered in trans, D3 strongly interacts with other parts of the intron and increases splicing efficiency (10). Group II introns possess unique conserved boundaries of 5’-GYGYG and 3’-AY (11).

Splicing of group II introns involves two sequential transesterification reactions. Initially, the 2’ OH of the unpaired and highly conserved bulged adenosine in D6 acts as a nucleophile, attacking the phosphodiester bond of the 5’ splice site to form a lariat intermediate (12). The second reaction uses the available 3’ OH of the 5’ exon to attack the phosphodiester bond of the 3’ end, resulting in ligation of the exons and release of the intron lariat (4). Several tertiary interactions within the intron secondary structure are believed to assist intron RNA stabilization and splicing (6,13). Due to conserved structural and sequence differences amongst group II introns, these genetic elements are divided into subgroups IIA, IIB and IIC. Each of these subgroups is further divided into subfamilies A1, A2, B1 and B2 (5).

A significant fraction of group II introns encode a protein known as an intron-encoded protein (IEP), whose ORF is invariably located in the loop of D4. The protein assists in the splicing and mobility of the intron (14–16). A typical group II IEP has four distinct protein domains, a reverse transcriptase (RT), maturation (X), nonconserved DNA binding (D) and endonuclease (En) domain (17).
The RT domain is subdivided into eight subdomains (0–7, with subdomain 0 corresponding to an N-terminal extension) (18). Domain X immediately follows RT subdomain 7 and spans ~100 amino acids (19). Although the function of domain X is unclear, mutational studies suggest it plays a role in RNA splicing (19–21). The D and En domains appear to play a critical role in reverse transcription and intron mobility (15). About a quarter of organellar and most bacterial IEPs lack the En domain, and phylogenetic analysis suggests that this domain has been lost multiple times in organellar and bacterial lineages (16,22).

Self-splicing introns nested within existing introns have been observed in several protist lineages and are generally referred to as ‘twintrons’ (23). In such cases, evidence suggests that the internal intron is spliced first, ligating the external intron, followed by external intron splicing and exon ligation (23). Twintrons can be comprised of two or more group II introns nested within one another or a combination of group II and group III introns, the latter being a miniaturized version of the former (24).

Group III introns are believed to be the descendents of group II introns but only retain D1 and D6, and have been found and studied only in euglenids, a eukaryotic group with secondary plastids of green algal ancestry (25,26).

Although group II introns have been studied in detail in bacterial genomes, some fungal mitochondrial genomes and eukaryotic organelar genomes of the green plastid lineage (1,14,27–31) have not been found in the plastids of red algae (32–36). An interesting exception is in the recently sequenced red algal-derived plastid genome of the cryptophyte alga *Rhodomonas salina* CCMP1319, in which a group II intron was found in the *psbN* gene (37). Cryptophytes are a remarkable group of unicellular eukaryotes that acquired photosynthesis via secondary endosymbiosis (38–41). This occurs when a nonphotosynthetic eukaryotic phagotroph ingests a photosynthetic eukaryote and retains its photosynthetic machinery. In addition to the presence of a group II intron, the *R. salina* genome is unusual in that it encodes a noncyanobacterial type DNA polymerase acquired by lateral gene transfer (LGT), the first instance of putative DNA replication machinery encoded in plastid DNA (37). While LGT is believed to be extremely rare in plastid genomes (42), cryptophyte plastids appear somewhat prone to the acquisition of foreign DNA (37,42).

Here, we present the sequence and predicted structure of six group II introns in the plastid genomes of the cryptophytes *Hemiselmis anderseni*, *Chroomonas paucli-plastida* and several species within the genus *Rhodomonas* (43). Phylogenetic analysis of IEPs suggests that the introns were acquired by LGT, most likely from a euglenid species. All six cryptophyte introns are unusually large and may be the product of an ancient amalgamation between two group II introns, with the majority of the internal intron deleted except for the ORF. Interestingly, the cryptophyte introns exist in a variety of genomic locations, suggesting recent transposition or multiple independent LGT events.

### MATERIALS AND METHODS

#### Cell culturing and nucleic acid extractions

Cryptophyte cultures were obtained from public culture collections and grown under conditions described previously (44). Total cellular RNA was isolated from 500 ml of cell culture harvested by centrifugation. The cell pellet was resuspended in 5 ml of TRI-REAGENT® (Invitrogen, Carlsbad, California) and 1 ml of chloroform and centrifuged for 30 min at 4°C. The supernatant was subjected to three rounds of phenol/chloroform extraction. RNA was precipitated using isopropanol, centrifuged and washed with 80% ethanol. DNA was isolated as described previously (44).

#### Gene amplification, cloning and sequencing

*GroEL* genes were amplified by PCR using a combination of exact-match and degenerate PCR primers. Intron-containing loci were too large to amplify in a single PCR reaction: genes were thus amplified in distinct overlapping fragments. Partial coding sequence of the 5′ region of the *groEL* gene for *Rhodomonas* sp. CCMP1178 and *Rhodomonas* sp. CCMP2045 were retrieved from GenBank (37) and exact match primers were designed to the 5′ ends of these sequences (2045.groEL.F1 GCACGG TTCCTATGAAAGATACCC, 1178.groEL.F1 GTACG GTTCGGACGAGAGGTATC). A degenerate forward primer was used for *Rhodomonas* sp. CCMP1170 and *Rhodomonas baltica* RCC350 (groEL.F1 GTCACTCTAGGNCCNAANGG), and a reverse degenerate primer was used for all of the *Rhodomonas* species (groEL.R5 CCTCTCTGTACDATNCYCCTCYTC). PCR products were purified using the MinElute Gel Extraction Kit (Qiagen, Valencia, California) and cloned using the Topo TA Cloning Kit (Invitrogen) according to the manufacturer’s protocol. At least five independent bacterial colonies were grown in LB broth overnight and the plasmids were extracted using the QuickLyse Miniprep Kit (Qiagen). Plasmid inserts were sequenced using the CEF Dye Terminator Cycle Sequencing Kit (Beckman Coulter, Inc., Fullerton, CA, USA) and a Beckman Coulter CEQ8000. Sequences determined in this study have been submitted to GenBank under the following accession numbers: EU305620-EU305621.

#### PCR and RT–PCR

RT–PCR was performed using the Qiagen OneStep RT–PCR kit according to the manufacturer’s protocol. Exact-match primers were designed to exons flanking the introns and to conserved group II intron features (D5 and the first 20–25 nt from the 5′ end). Primer sequences corresponding to the introns and exons were as follows: Rhodo 1178intron.F1 GTCCGATTTCCCTTTAAGT AAACAG, Rhodo1178intron.R1 CGTACGGTCGAT TTTCCACC, Rhodo1178intron.F3 GTGGTGAAAG TCCAACCATGCA, Rhodo1178intron.R3 CCTGCTACATCATTTGCTTGAAGC, C.paucipl. intron.F1 GTGCACTCATGCTTTGCTTGAAGC, C. pau cipl. intron.R1 CTCTCTGGATCTGCGGCGTGC,
C. paucipl. exon.F1 CAGGACCTGCTCATATAGGAAC
G, C. paucipl. exon.R1 GGTTGTTCTTCAATATCCTTT
CTGG, R. sal. 1319.intron.R1 CTCCTCATTCAGATCC
GTACGTC, R. sal. 1319.intron.F1 GCGATTCGTTTCT
TAGTACAATGG, R. sal. 1319.psbN.R1 CTCTGGCC
CATGTCTTTTTTAATC, R. sal. 1319.psbN.F1 GGA
AAGCGACAGTTTTTTAGCC.

PCR reactions were performed using reagents supplied with the Qiagen RT–PCR kit but with the use of Invitrogen Hi-Fidelity Taq polymerase. RNA template was treated with DNase supplied by Promega; DNase treatments were performed at 37°C for 30 min. DNase activity was terminated by the addition of stop solution (Promega, Madison, Wisconsin) and incubation at 65°C for 12 min. Thermal cycling conditions for products under 2 kb were as follows: 50°C for 30 min, 95°C for 15 min, followed by 40 cycles of 94°C for 30 s, 52°C for 30 s and 72°C for 2 min, with a final extension at 72°C for 10 min. For products above 2 kb, the initial extension temperature was reduced from 50°C to 45°C and the Taq polymerase extension temperature was changed to 68°C.

Phylogenetic analysis and intron structure prediction

A set of 84 IEP sequences from a wide range of prokaryotic and organellar genomes were retrieved from GenBank using a combination of BLASTP (45) and genome-specific searches. Protein sequences were aligned using ClustalX (46) with manual adjustment performed in MacClade 4.06 (47). Preliminary phylogenetic analyses were performed on the full dataset in order to (i) eliminate highly similar/redundant sequences, (ii) detect obvious outliers whose evolutionary position would not bear on the question of the origin of the cryptophyte introns and (iii) eliminate extremely divergent/long branching sequences, to minimize the impact of long-branch attraction artifacts. We settled on an alignment of 50 sequences and 318 sites (available upon request) that was suitable for more rigorous analysis. Maximum likelihood analysis of IEPs was performed using PhyML (48) and IQPNNI (49). The WAG, JTT and RtREV amino acid substitution matrices were used with a gamma distribution estimated by four categories to model site rate heterogeneity. Statistical support for each node was determined by bootstrap analysis (100 replicates).

Intron secondary structure predictions were initially performed using MFOLD (50,51). Manual manipulation and rearrangement of various domains was performed by eye using the generic structures of group IIB introns presented by Toor et al. (52) as reference.

RESULTS AND DISCUSSION

Cryptophyte plastid intron diversity

Preliminary investigations suggest that group II introns are a prominent feature of the plastid genomes of cryptophytes. Maier et al. (53) first described an unusual self-splicing intron in the groEL gene of the cryptophyte Pyrenomonas salina (now R. salina) and we recently identified introns in the psbN gene of R. salina strain CCMP1319 (37) and the B subunit gene of light-independent protochlorophyllide oxidoreductase (chlb) from H. anderseni and C. pauciplastida (54). In order to better understand the diversity and structure of cryptophyte group IIB introns, groEL genes were sequenced from four additional cryptophytes (Rhodomonas sp. CCMP1178, Rhodomonas sp. CCMP2045, Rhodomonas sp. CCMP1170 and Rhodomonas baltica RCC350) and a detailed sequence and secondary structure analysis was performed on the introns contained therein.

A total of six introns from three different genes (groEL, psbN and chlb) were analyzed. Unexpectedly, introns found in the same gene in different cryptophytes were often present in distinct locations. The H. anderseni and C. pauciplastida chlb introns are separated by 60 bp and while the H. anderseni intron possesses an ORF, the C. pauciplastida intron does not (Figure 1). The Rhodo-

monas sp. CCMP2045 groEL gene was found to contain two introns (groEL-1 and groEL-2) and Rhodomonas sp. CCMP 1178 groEL contains a single intron (groEL-3) in

Figure 1. Location of group II introns in cryptophyte plastid genes. Genes are represented as shaded boxes (roughly to scale) with the intron locations highlighted by gray triangles. Introns that possess an ORF contain an additional box (dark gray) on top of the triangle.
the same location as groEL-1. No obvious nucleotide sequence similarity exists between any of these introns, except for short stretches of similarity at the 5' and 3' ends of groEL-1 and groEL-3 (the latter intron contains an ORF while the former does not). The original ORF-containing R. salina groEL intron published by Maier et al. (53) is located in yet a third position (Figure 1).

In contrast, the groEL gene in the completely sequenced plastid genome of R. salina CCMP1319 (37) is intron-lacking, as are the groEL genes PCR-amplified from Rhodomonas sp. 1170 and R. baltica. All four of the IEPs possess features in common with typical group II IEPs, including RT domains 0–7 and an X domain, but lack the D and En domains, as is typical for organellar introns. Like the proteins encoded in euglenid plastid introns, as well as several yeast and plant mitochondrial introns, the cryptophyte plastid IEPs lack the highly conserved YADD motif in subdomain 5 of the RT domain (16). Although bacterial introns have been shown to transpose despite lacking a recognizable En domain (14), this has not been observed in organellar introns. Together with the lack of the D and En domains, the absence of a YADD motif in the cryptophyte IEPs would seem to suggest that these introns are immobile (see below).

**Intron secondary structure**

The cryptophyte plastid introns possess many of the features found in group IIB introns (52), including a highly conserved sequence matching that of a typical D5,
an unpaired adenosine residue in D6 and many of the predicted tertiary interactions such as exon–intron-binding sites (EBS1-IBS1, EBS2-IBS2), and β–β’, α–α’, ε–ε’ and λ–λ’ interactions (5,7,13) (Figures 2 and 3, Supplementary Data Figures 1–4). However, they also have several features not previously seen in group II introns, most notably insertions between D2 and D5. For example, 206–315 bp of sequence separates the beginning of D5 from the end of the ORF in most of the ORF-containing introns (Figures 2 and 3, Supplementary Data Figures 1 and 3). Furthermore, in *Rhodomonas* sp. CCMP1178 and *R. salina* CCMP1319, the ORF is not located in D4 as is normally the case in group II introns, but is instead located in a novel domain present immediately downstream of D2 (Figure 3 and Supplementary Data Figure 1). The ORF present in the groEL-2 intron of *Rhodomonas* sp. CCMP2045 resides in a distinct domain adjacent to D3 (Supplementary Data Figure 3), while the *H. andersenii* chlB intron ORF resides in the loop of D4 (Figure 2). The *H. andersenii* intron is also unlike the other cryptophyte introns in that it has a canonical D3 consisting of conserved nucleotide base pairing specific to group IIB introns (52), with 143 bp (plus a 1290-bp ORF) separating the end of D3 and D5. The predicted α–α’ interaction residues were found, whereas the ε–ε’ and λ–λ’ interactions appear to be absent (Figure 2).

Remarkably, the *C. pauciplastida* chlB intron (Supplementary Data Figure 4) does not encode an ORF yet is 1121 bp in size, the largest noncoding intron found to date. We were unable to reliably fold 428 bp of sequence present between D2 and D3 of the *C. pauciplastida* intron, as multiple distinct structures were predicted by MFOLD, none of which showed similarity to known group II intron domain structures. BLAST analysis (45) did not detect the presence of a degenerate ORF in this region. Several of the groEL introns (e.g. groEL-1 from *Rhodomonas* sp. CCMP2045) were also difficult to fold.

**Figure 3.** Predicted secondary structure of the group IIB intron present in the *groEL* gene of *Rhodomonas* sp. CCMP1178. All six canonical group II intron domains are labeled with Roman numerals (I–VI). Tertiary interactions are labeled through the use of Greek letters and shaded gray. The unpaired adenosine residue is circled and enlarged. EBS and IBS refer to exon- and intron-binding sites, respectively.
with confidence in certain areas and given the unusual placement of most of the cryptophyte intron ORFs described earlier (i.e. outside D4), additional experiments such as X-ray crystallography will be required to determine their precise structures.

Intron splicing

Although the predicted secondary structures of the cryptophyte introns described earlier are distinct from one another (Figures 2 and 3; Supplementary Data Figures 1–4), they are all unusually large compared to typical group II introns and, with the exception of H. andersenii, the locations of their ‘insertion’ sequences are similar. This raises the possibility that they possess nested introns, as was proposed for the original groEL intron of R. salina (53). We tested this hypothesis using RT–PCR to detect the presence of nested splicing reactions. RT–PCR primers were designed to intron-flanking exonic sequence as well as the outermost region clearly identified as the putative ‘external’ group II intron, with the forward primer being specific to the 5’ end and the reverse primer to the conserved D5. RT–PCR experiments using the intron primers should detect the presence of internal splicing activity, if present. In order to eliminate the chance of amplification from DNA contamination, RNA samples were treated with DNase and additional controls were carried out in which DNase-digested template was used in RT–PCR reactions with only Taq DNA polymerase, instead of a combination of RT and Taq DNA polymerase.

RT–PCR results for four cryptophyte introns are shown in Figure 4. Amplicons generated using exon primers against groEL, chlB and psbN were 150 bp or less and in each case, cloning and sequencing confirmed that these products are ligated exons, i.e. derived from fully spliced RNA. When primers designed to intron sequences were used, RT–PCR and PCR reactions yielded products of identical size (e.g. compare lanes 2 and 3 to lanes 6 and 7 in Figure 4a), indicating that no additional internal splicing was taking place in any of the four introns tested. In sum, these results indicate that the unusually large cryptophyte introns are spliced as a single entity.

Phylogeny of IEPs

To gain insight into the origin(s) of the cryptophyte group II/III introns, we performed phylogenetic analyses using a large set of IEPs from mitochondrial and plastid genomes as well their homologs in diverse bacteria. Maximum likelihood phylogenies (Figure 5) show that the three Rhodomonas proteins encoded in the groEL introns and the psbN IEP of R. salina CCMP1319 form a monophyletic group with weak statistical support. In addition, these sequences are related to the group III intron ORFs present in the psbC gene of the Euglena longa, Euglena gracilis and Lepocinclis buetschlii plastid genomes. Statistical support for this relationship is strong (98% and 100% using the PhyML and IQPNNI methods of tree reconstruction, respectively; Figure 5). Intron density in the completely sequenced E. gracilis plastid genome is extraordinarily high (28) but out of the 160 group II/group III introns present, only three possess an ORF. One of these resides within a group III intron while the other two are group II intron-encoded (the latter two IEPs were too divergent to reliably include in our dataset). These results suggest that the Rhodomonas introns were acquired by LGT from a euglenid-type group III intron. Phylogenies of Rhodomonas intron-containing groEL proteins in the context of a diverse set of plant, algal and bacterial homologs indicate that these proteins are red algal in origin (data not shown), as would be predicted based on the evolutionary history of the cryptophyte plastid (37), suggesting that it was the intron (and its ORF)—not the groEL gene itself—that was transferred.

The mitochondrial encoded group II intron ORFs in the red alga Porphyra purpurea appear as the nearest outgroup to the Rhodomonas and euglenid ORFs in our phylogenies (Figure 5). A previous study suggested that these red algal introns are themselves the product of LGT from a cyanobacterial donor (55). The relationship between the cyanobacterial introns and the plastid-encoded introns of the green algae Euglena myxoxynidracea and Chlamydomonas sp. has also been suggested to be the result of multiple LGTs, involving cyanobacteria and these two eukaryotes (56,57). Our results are consistent with these hypotheses, although it should be emphasized that support for the backbone of the cyanobacterial/organellar portion of the phylogeny is very weak. It is also difficult to infer the directionality of putative LGTs with certainty. In the case of the groEL introns in Rhodomonas, it is formally possible that the LGT occurred in the direction of cryptophytes-to-euglenids and not the other way around. However, combined with secondary structural considerations (see below), the huge number of group II/group III introns present in the E. gracilis chloroplast genome compared to the cryptophyte plastid genomes makes this scenario unlikely.

Interestingly, the chlB IEP in H. andersenii does not branch with the other cryptophyte sequences but instead clusters with a sequence from the mitochondrial genome of the moss Physcomitrella patens, two green algal plastid sequences and diverse cyanobacterial homologs. The placement of this sequence outside of the cryptophyte clade raises the intriguing possibility that the H. andersenii chlB intron was acquired independently of the psbN and groEL introns, a hypothesis that is further supported by the fact that its secondary structure is very different from the others. However, combined with secondary structural considerations (see below), the huge number of group II/group III introns present in the E. gracilis chloroplast genome compared to the cryptophyte plastid genomes makes this scenario unlikely.

With respect to the evolution of the Rhodonomas introns themselves, it is significant that the groEL introns are very distinct from one another and reside in different
locations in different organisms. This either means that intron transposition (and rapid sequence divergence) has occurred very recently during cryptophyte plastid genome evolution or that multiple independent LGTs have given rise to the observed complement of \textit{groEL} introns. As noted earlier, both the euglenid and cryptophyte IEPs lack a recognizable YADD motif. Regardless of whether the loss of this motif occurred independently in euglenids and cryptophytes or in euglenids prior to LGT, based on what is known about the function of the YADD motif in other systems (16), the euglenid and cryptophyte introns should be impaired in terms of their mobility. Yet, the shear abundance of introns in the chloroplast genome of \textit{Euglena} (28) would seem to suggest that these introns are in fact mobile and the variation of intron location in cryptophytes is also consistent with mobility. However, at present, it is not possible to assess the relative contributions of LGT versus transposition in giving rise to the spectrum of self-splicing introns in the cryptophyte plastid. In combination with detailed
biochemical analyses, a much broader sampling of complete cryptophyte plastid genome sequences would be useful in this regard.

Cryptophyte plastid introns—amalgamation or degeneration?

Our phylogenetic analyses suggest a specific evolutionary connection between the groEL and psbN introns of *Rhodomonas* species and group III introns present in euglenid plastid genomes (24–26), the latter being derived versions of group II introns. While the cryptophyte intron secondary structures possess additional domains that could correspond to parts of internal introns nested within group II introns (Figure 2 and 3, Supplementary Data Figures 1 and 3), RT–PCR experiments indicate that they are spliced as a single entity. Therefore, the unusually large cryptophyte introns could simply be highly degenerate group II introns that have greatly expanded in size in particular regions (Figures 2 and 3; Supplementary Data Figures 1–4), but nevertheless still retain the ability to splice (Figure 4). A more intriguing possibility is that the cryptophyte introns were originally twintrons—as seen in

the plastid genomes of euglenid species—that have degenerated and amalgamated to produce a single splicing entity. Consistent with amalgamation is (i) the large size of the *Rhodomonas* introns, (ii) the fact that canonical features of group II introns (e.g. D3, D5 and D6) can be found flanking the intron insertions and (iii) the atypical placement of their ORFs, i.e. upstream of D4. In theory, the intron amalgamation(s) could have occurred in the cryptophyte plastid genome after LGT or in the euglenids (or an intermediate species) prior to LGT. Unfortunately, a significant amount of sequence divergence has taken place between the euglenid and cryptophyte introns (and between the cryptophyte introns themselves), and extra group II intron-like domains cannot be identified within the above-mentioned insertions. Regardless, if our secondary structure predictions are correct, the *Rhodomonas* introns represent the first described instances of group II intron ORFs being located outside of D4.

What circumstances could have led to amalgamation(s) in the cryptophyte plastid introns? When considering this possibility, it is important to consider that in addition to the maturase activity provided by the X domain of the IEPs, nucleus-encoded protein factors are often essential
for group II intron splicing in mitochondria and plastids. For example, approximately 18 nucleus-encoded proteins are needed for the splicing of a mitochondrial group II intron in the yeast *Saccharomyces cerevisiae*, and 14 nucleus-encoded proteins are required for proper splicing of a plastid group II intron in *Chlamydomonas reinhardtii* (22,58–61). Due to their peculiar distribution in nature, these proteins are thought to have been ‘co-opted’ to function in intron splicing relatively recently (22). Traditional group II introns (non-twintrons) that are present in the mitochondria of fungi, and plastids of green algae, presumably splice with the assistance of the intron-encoded matarase in conjunction with nucleus-encoded factors. In the case of euglenids, it seems likely that a distinct set of nucleus-encoded proteins are required for nested intron splicing, since twintrons have only been found and shown to splice individually in members of this lineage. If, as our phylogenetic analyses suggest, the *Rhodomonas* introns are the product of LGT from plastid-encoded twintrons in a euglenid-like organism, it seems unlikely that the full complement of necessary genes for nucleus-encoded splicing factors would be transferred at the same time to the cryptophyte nucleus, and even if they were, their protein products may or may not contain N-terminal targeting signals that would function in a cryptophyte cell. Without such factors, the transferred twintron would be functionally impaired or nonfunctional, and if the internal intron were particularly reliant on the presence of nucleus-encoded factors, there would presumably be strong selective pressure for deletions and compensatory changes to the innermost intron, such that splicing activity of the outermost intron is maintained.

The group III introns of euglenids, which are likely shrunken versions of group II introns, are a potentially important link to the cryptophyte introns, as are the ‘mini’-group II introns of the euglenid *Lepocinclis beutschlii* (24). Two ‘mini’-group II introns in *L. beutschlii* are a mere 224 and 258 bp in size, in between group II and group III introns, and their internal introns have been shown to splice independently (24). The mini-group-II intron secondary structure is composed of short D1, D5 and D6 domains, with two other small domains that are not found in canonical group II introns. This novel structure could represent an intermediate in the transition from group II to group III. It seems likely that the nuclear genomes of euglenids encode protein factors that are essential for the splicing of these highly unusual introns.

The *groEL* introns of *Rhodomonas* sp. CCMP2045 (*groEL*-1) and *Rhodomonas* sp. CCMP1178 (*groEL*-3) are interesting in that while they are located in the same position and are presumably the product of a single insertion, they are extremely different from one another. Only *groEL*-3 contains an ORF (Figures 1 and 3) and sequence similarity between the two is limited to ~200 bp at the 5’ end and ~100 bp at the 3’ end. Given that the protein product of one intron ORF should be able to provide splicing activity in *trans* to the remaining introns in the genome, it is possible that *groEL*-1 originally possessed an ORF that subsequently acquired mutations that led to its eventual degeneration. In each of the cryptophyte introns examined (except for *H. andersenii*), the remnants of an internal intron, along with regions of the external intron, could have given rise to several new domains and a novel D3 and D4. Following the loss of its splicing ability, the majority of the internal intron would have been deleted, presumably in order to stabilize the group II intron structure. We predict that the position of the internal intron was between D2 and D5, and random deletion of the internal intron could have resulted in the concomitant deletion of the outer D3 and D4. Amalgamation of the two introns would have formed a stable secondary structure, with a novel D3 and D4 that now represent the remnants of the internal intron and regions of the external intron. Where present, the *Rhodomonas* intron ORFs are located within the loop of a domain, and this particular domain probably replaced the function of a D4, hence the presence of the original (i.e. external) D4 is not essential.

## CONCLUSION

We have demonstrated a highly complex evolutionary history for the group II-like introns in the plastid genomes of cryptophyte algae. These introns exhibit an unusual secondary structure that could be the result of amalgamations between group II introns that were laterally transferred to the cryptophyte plastid, possibly multiple times independently. While a more complete picture of intron evolution in cryptophytes will require additional plastid genome sequences and biochemical experimentation, it would appear that LGT has played an important role in shaping the structure and composition of the cryptophyte plastid.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank C. Lane and W. F. Doolittle for comments on an earlier version of this article, as well as those of two anonymous reviewers. M. Schnare is also thanked for comments and generous assistance with intron secondary structure predictions. This work was supported by an NSERC discovery grant awarded to J.M.A. J.M.A. is a Scholar of the Canadian Institute for Advanced Research, and a NSERC discovery grant awarded to J.M.A. J.M.A. is a Scholar of the Canadian Institute for Advanced Research, Program in Integrated Microbial Biodiversity. Funding to pay the Open Access publication charges for this article was provided by NSERC.

*Conflict of interest statement*. None declared.

## REFERENCES


