A peer-reviewed version of this preprint was published in PeerJ on 18 June 2015.

View the peer-reviewed version (peerj.com/articles/1017), which is the preferred citable publication unless you specifically need to cite this preprint.

Perrineau M, Price DC, Mohr G, Bhattacharya D. 2015. Recent mobility of plastid encoded group II introns and twintrons in five strains of the unicellular red alga *Porphyridium*. PeerJ 3:e1017 https://doi.org/10.7717/peerj.1017

Recent mobility of plastid encoded group II introns and twintrons in five strains of the unicellular red alga *Porphyridium*

Marie-Mathilde Perrineau[1], Dana C. Price[1], Georg Mohr[2], Debashish Bhattacharya[1,3*]

[1]Department of Ecology, Evolution and Natural Resources, Rutgers University, New Brunswick, New Jersey, 08901, USA

[2]Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, TX 78712, USA

[3]Institute of Marine and Coastal Science, Rutgers University, New Brunswick, New Jersey, 08901, USA

[*]Corresponding author: Debashish Bhattacharya, Department of Ecology, Evolution and Natural Resources, Rutgers University, 59 Dudley Road, 102 Foran Hall, New Brunswick, NJ 08901, USA. Telephone: +1 848-932-6218; Fax: +1 732-932-8746; E-mail: debash.bhattacharya@gmail.com

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.729v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

Abstract

Group II introns are closely linked to eukaryote evolution because nuclear spliceosomal introns and the small RNAs associated with the spliceosome are thought to trace their ancient origins to these mobile elements. Therefore, elucidating how group II introns move, and how they lose mobility can potentially shed light on fundamental aspects of eukaryote biology. To this end, we studied five strains of the unicellular red alga *Porphyridium purpureum* that surprisingly contain 42 group II introns in their plastid genomes. We focused on a subset of these introns that encode mobility-conferring intron-encoded proteins (IEPs) and found them to be distributed among the strains in a lineage-specific manner. The reverse transcriptase and maturase domains were present in all lineages but the DNA endonuclease domain was deleted in vertically inherited introns, demonstrating a key step in the loss of mobility. *P. purpureum* plastid intron RNAs had a classic group IIB secondary structure despite variability in the DIII and DVI domains. We report for the first time the presence of twintrons (introns-within-introns, derived from the same mobile element) in Rhodophyta. The *P. purpureum* IEPs and their mobile introns provide a valuable model for the study of mobile retroelements in eukaryotes and offer promise for biotechnological applications.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.729v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

Introduction

Nuclear genome evolution and eukaryotic cell biology in general are closely tied to the origin and spread of autocatalytic group II introns. These parasitic genetic elements are thought to initially have entered the eukaryotic domain through primary mitochondrial endosymbiosis (e.g., Rogozin et al. 2012; Doolittle 2014). Thereafter, group II introns presumably migrated to the nucleus and gave birth to the forerunners of nuclear spliceosomal introns and the small RNAs associated with the spliceosome (Cech, 1986; Sharp, 1991; Qu et al. 2014). This explanation of intron origin, although widely held to be true (e.g., Rogozin et al. 2012) is nonetheless shrouded in the mists of evolutionary time. Understanding more recent cases of group II intron gain and loss are vital to testing ideas about the biology of autocatalytic introns. Here we studied group II intron evolution in five closely related strains of the unicellular red alga *Porphyridium purpureum* (Rhodophyta) that surprisingly contain over 40 intervening sequences in their plastid genomes (Tajima et al. 2014). Red algae are not only interesting in their own account as a taxonomically rich group of primary producers (Ragan et al. 1994; Bhattacharya et al. 2013) but they also contributed their plastid to a myriad of chlorophyll *c*–containing algae such as diatoms, haptophytes, and cryptophytes through secondary endosymbiosis (Bhattacharya et al.*;* 2004; Archibald, 2009). Therefore, group II introns resident in red algal plastid genomes could also have entered other algal lineages through endosymbiosis.

With these ideas in mind, we explored the genetic diversity, secondary structure, and evolution of group II introns and their mobility-conferring intron-encoded proteins (IEPs; Lambowitz and Zimmerly, 2011) in the plastid genome of five strains of *P. purpureum*, four of which were determined for this study. Phylogenetic analyses show

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.729v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

that the *P. purpureum* IEPs and their introns are monophyletic, suggesting a shared evolutionary history (Toro and Martínez-Abarca, 2013). Analysis of IEPs reveals key traits associated with intron mobility and loss, and analysis of secondary structures uncover unique features of red algal group II introns. We also report for the first time the presence of twintrons (introns-within-introns) in Rhodophyta plastid genomes and deduce their recent origins from existing IEPs that targeted heterologous DNA sites. In summary, our study identifies a promising red algal model for the study of group II intron biology and evolution and suggests these mobile elements could potentially be harnessed for biotechnological applications (Enyeart et al. 2014).

Materials and Methods

Porphyridium purpureum strains and plastid genomes

Four *Porphyridium purpureum* strains (SAG 1380-1a, SAG 1380-1b, SAG 1380-1d [obtained from the Culture Collection of Algae, Göttingen University] and CCMP 1328 [from the National Center for Marine Algae and Microbiota, East Boothbay, ME]) were grown under sterile conditions on Artificial Sea Water (Jones et al. 1963) at 25°C, under continuous light (100µmol photons$\cdot m^{-2} \cdot s^{-1}$) on a rotary shaker at 100 rpm (Innova 43, New Brunswick Eppendorf, Enfield, CT). Cells were pelleted *via* centrifugation and DNA was extracted from ca. 100 mg of material with the DNeasy Plant Mini Kit (Qiagen) following the manufacturer's protocol. Sequencing libraries were prepared for each strain using the Nextera DNA Sample Preparation Kit (Illumina Inc, San Diego, CA) and sequenced on an Illumina MiSeq sequencer using a 300-cycle (150x150 paired-end) MiSeq Reagent Kit v2 (Illumina, Inc.). Sequencing reads were quality and adapter

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.729v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

trimmed (Q limit cutoff = 0.05) and overlapping pairs were merged at 3' end overlap using the CLC Genomics Workbench 6.5.1 (CLC Bio, Aarhus, Denmark).

Mapping, polymorphism detection and analysis

The reads from each strain were mapped to the *P. purpureum* plastid reference genome (strain NIES 2140; Tajima et al. 2014) with a stringency of 90% sequence identity over a 90% read length fraction using the CLC Genomics Workbench (CLC Bio, Aarhus, DK). SNPs were called using the Genomics Workbench 6.5.1 quality-based variant detection (≥10x base coverage, quality score >30 and ≥50% frequency required to be called). An uncorrected distance phylogeny was constructed using a matrix of DNA polymorphisms detected between the five plastid genomes with the program MEGA6.06 (Tamura et al. 2013; 100 bootstrap replicates).

Group II intron and IEP identification

Novel group II introns in the plastid genomes of the four *P. purpureum* strains were identified by aligning de-novo assembled (using the CLC Genomics Workbench v.6.5.1 de-novo assembler) plastid contigs from each strain to the NIES 2140 reference. Insertions were annotated as putative introns, and further confirmed by mapping the raw short read data to the contigs and manually inspecting for assembly artifacts. The group II intron/IEP sequences described here are accessible using NCBI accession numbers KKJ826367 to KKJ826395 and the *P. purpureum* plastid genome under NC_023133 (Tajima et al. 2014).

Intron encoded proteins (IEPs) were identified within the novel introns by ORF detection using the bacterial/plastidic genetic code. The four domains that constitute an IEP (i.e., reverse transcriptase [RT], maturase [X], DNA-binding [D], and endonuclease [En] [Mohr et al. 1993; San Filippo and Lambowitz, 2000]) were identified by sequence alignment using ClustalX (Larkin et al. 2007) to known IEPs of the prokaryote CL1/CL2 group and to those from the Rhodophyta, Viridiplantae, Cryptophyta, Euglenozoa, and stramenopiles (listed in Table S1) obtained from NCBI and the Group II intron database. To examine the phylogeny of these mobile elements, the IEP peptide sequences were aligned with the RT-domain alignment of Toro and Martínez-Abarca (2013) and maximum likelihood phylogenies were inferred under the WAG amino acid substitution model with 100 bootstrap replicates using MEGA6.06.

Intron structure and evolution

Intron secondary structures were predicted using sequence alignment, manual domain identification, and automatic structure conformation in comparison with previously predicted structures of group IIB introns using the Mfold web server (Zuker, 2003; Table S1). A detailed secondary structure model was generated based on the *rpoC1* intron and *mat1g* IEP (Fig. 1). This was then used as a guide to predict draft structures using PseudoViewer3 (Byun and Han, 2009) for all other group II introns. A domain alignment was then performed against the group II intron structures derived for the cryptophyte *Rhodomonas salina* (Maier et al. 1995; Khan et al. 2007) using ClustalX2.1, and a maximum-likelihood phylogeny was generated using intronic nucleotide sequence data under the GTR + I + Γ model with 100 bootstrap replicates using MEGA6.06 (Tajima et

al. 2014). Prior to this, the IEPs or IEP remnants were removed to avoid potential long-branch attraction artifacts. Additionally, conserved motifs within the basal DI, DIV, DV and DVI domains were used as a BLASTN (Altschul et al. 1990) query to the five aligned plastid genomes to identify additional group II intron structures present in all strains (and thus not identified via length heterogeneity upon initial assessment).

The twintrons present in the *P. purpureum* plastid genome were aligned and compared to the other introns to allow identification of the outer and inner introns, exon binding sites, to describe their secondary structures, and potentially to understand their mode of origin.

Results and Discussion

A phylogenetic tree of the five studied *P. purpureum* strains inferred on the basis of 332 single nucleotide polymorphisms (SNPs) present in their plastid genomes demonstrates the monophyly of the four strains reported here (SAG 1380-1a/b/d, CCMP-1328) with respect to strain NIES 2140 (Fig. 2A; Tajima et al. 2014). By examining length heterogeneity within these plastid genome sequence alignments, we identified four novel group II intron/IEP combinations (*mat1f, 1g, 1h, 1i;* Fig. 2B) in addition to the five previously reported (*mat1a, 1b, 1c, 1d, 1e*) by Tajima et al. (2014). These novel elements were found to exhibit lineage-specific distributions on the phylogeny, whereas *mat1a, b, c* and *e* were recovered from all strains (Fig. 2B). Using conserved structural motifs (see Fig. S1 and Materials & Methods) as the basis for a homology search within remaining intronic and intergenic *P. purpureum* plastid sequence, we were able to define three additional group II introns (int *mntA*, int.a *rpoB*, and an intergenic *psbN-psbT*) present in

all four strains and containing remnant (or 'ghost') ORFs that have lost their IEPs. These introns were subsequently included in our analyses.

We identified six new intron/IEP insertion sites in our *P. purpureum* strains (*mat1fa, 1fb, 1fc, 1g, 1h, 1i*) in addition to the five previously described in the NIES 2140 strain (*mat1a, 1b, 1c, 1d, 1e*; Tajima et al. 2014). Among the nine intron/IEP combinations present, only four occur at the same insertion site in all strains (*mat1a*, *1b*, *1c*, and *1e*), whereas four are unique to individual strains (*mat1d*, *1g*, *1h*, and *1i*). The *mat1fa* and *mat1fb* intron/IEP combinations are identical at the nucleotide level and form twintrons (see below), whereas *mat1fc* is a novel insertion in the *atpB* gene and contains a single SNP.

A maximum-likelihood phylogeny was constructed using an alignment of the 42 group II introns present in NIES 2140 and the novel introns described in this study (with IEP sequences removed from the alignment; Fig. 3). This analysis demonstrates that twelve IEP/IEP-remnant containing introns in *P. purpureum* are monophyletic (88% bootstrap support), whereas two introns (*mat1a* and *mat1b)* are evolutionarily distinct. Despite partial nucleotide sequence conservation (Fig. S1), the intergenic region containing *mat1a* was unable to be folded into a functional group II intron structure (only domains DIV-DVI could be identified (Fig. S2), and we were unable to identify any group II intron secondary structural homology within the *mat1b*-containing intron (see Fig. S1 and the section below entitled, 'Group IIB intron secondary structure'). In addition, the group II introns with remnant or ghost ORFs recovered in our analysis grouped with those that maintained functional IEPs. These results are consistent with the evolutionary model widely accepted for group II introns (Toor et al. 2001; Simon et al.

2009) that predicts co-evolution of IEPs and self-splicing RNAs, and suggests that IEP-lacking (remnant) introns derive from introns that once contained a functional mobility-conferring enzyme. Here, we show for the first time examples of recent intron mobility and putative stability; the latter being represented by plastid-encoded IEPs that lack the endonuclease domain due to mutation and/or sequence degeneration.

Intron-encoded proteins

Intron-encoded proteins present at the same insertion site are nearly identical among the strains (98.9-100% amino acid identity), except for the *mat1b* IEP in strain NIES 2140 which has an apparent truncation of 27 amino acids due to a premature stop-codon. All nine IEPs contain two fully conserved reverse transcriptase (RT) and maturase (X) domains (Fig. S3), whereas four of the five elements present in all five *P. purpureum* strains (*mat1a, 1b, 1c, 1e*) are either truncated or have completely lost the DNA-binding (D) and endonuclease (EN) domains responsible for conferring mobility (Simon et al. 2009). These latter group II introns thus appear to have lost mobility, and exhibit vertical inheritance. Additionally, *mat1a* and *mat1b* lack the YADD motif crucial for reverse transcriptase activity at the active site (Fig. S3; Moran et al. 1995). The remaining five group II introns are distributed in a lineage-specific pattern on the *P. purpureum* phylogeny (Fig. 2A) and likely remain mobile because they retain all functional domains (Fig. S1).

Phylogenetic analysis using the IEP peptide alignment shows that seven of the nine *P. purpureum* IEPs form a monophyletic clade that is sister to cryptophyte plastid IEPs, the cyanobacterial CL2B clade, and Euglenozoa plastids (Fig 4). The *mat1a* and

199 *mat1b* IEPs, derived from group II introns found to lack typical secondary structure,
200 create a paraphyletic assemblage within the cryptophytes (*mat1a*) or group outside of the
201 CL2B clade (*mat1b*).
202
203 ## Group IIB intron secondary structure
204 Self-splicing group II introns are dependent on a conserved secondary and tertiary RNA
205 structure. These autocatalytic genetic elements are composed of six distinct double-
206 helical domains (DI to DVI) that radiate from a central wheel with each domain having a
207 specific activity (Lambowitz and Zimmerly, 2011). As illustrated by the *rpoC1* intron
208 that contains *mat1g* (Fig. 1), the introns studied here have group IIB intron secondary
209 structures following this model. Annotated sequence alignments and draft secondary
210 structures for the remaining introns are presented in the supplementary information (Figs.
211 S4-S16). As expected, the *P. purpureum* IEPs are located in domain IV (DIV), which is
212 integral for ribozyme activity. DIVa (including the protein-binding site) and DV contain
213 conserved regions (96±4% identity), whereas DVI is highly variable (37±17% identity;
214 length range 44-162bp; see Fig. S4).
215 The DVI domain contains a conserved, bulged adenosine that serves as a
216 nucleophile during lariat generation upon splicing (Peebles et al. 1986; Robart et al.
217 2014), however most *P. purpreum* group II intron models described here maintain an
218 additional unpaired guanine in an AG bulge. The effect this has on the splicing reaction
219 remains unknown. Structural analysis reveals a novel and unusual bipartite DIII domain
220 configuration predicted for the intron containing *mat1c* and the IEP-lacking structures
221 (int.b *rpoC2*, *psbN-psbT*, int.a *rpoB,* int *mntA;* Figs. S12-S16). The DIII domain

222 contributes an adenosine pair to a base stack that serves to reinforce DV opposite the
223 catalytic site, and stabilizes the entire structure (Robart et al. 2014). Modification of this
224 domain in the *P. purpureum* group II intron structures that have lost mobility may reflect
225 the lack of an IEP and thus need for reinforcement.

226 Group II intron RNAs self-splice *via* base-pairing interactions between exon-
227 binding sites (EBS1 & EBS2) on the ribozyme and intron-binding sites (IBS1 & IBS2) at
228 the 5' exon region (Lambowitz and Zimmerly, 2011). Despite a common origin, the *P.*
229 *purpureum* introns that encode an IEP appear to have a highly variable EBS (Fig. S17)
230 that may explain their ability to spread to novel sites in these plastid genomes. Each
231 EBS/IBS pairing is uniquely associated with an intron/IEP combination, and
232 complementarity between both is present. EBS1 and/or EBS2 were not identified for the
233 *mat1a*, *mat1b*, and *mat1c* introns. Interestingly, EBS1 is located at the same site in the
234 nucleotide alignment, whereas the EBS2 position is variable due to length heterogeneity
235 between introns. Understanding how variation in these binding sites affects ability of
236 group II introns to self-splice and bind target DNA is paramount for 'targetron'
237 development (Enyeart et al. 2014) and application of these mobile elements to
238 biotechnology.

239 Finally, sequence alignment of the *P. purpureum* introns described here with the
240 five *Rhodomonas salina* introns presented in Khan and Archibald (2008) (Fig. S4)
241 demonstrates that the domain organization and secondary structure of these elements in
242 both species are similar. We were thus able to derive amended secondary structures for
243 the cryptophyte models proposed by Maier et al. (1995) and Khan and Archibald (2008)
244 using *P. purpureum* as a guide. In doing so, we identified a cryptophyte domain IVa

similar to that in *P. purpureum* that contains the IEP and has modified domains DII and DIII (e.g., Fig. S18). We propose that the non-canonical features described by Khan and Archibald (2008) in *R. salina* and *H. andersenii* (i.e., domain insertions, ORF relocation, absence of internal splicing) can be explained by degeneration of the endonuclease domain between the protein C-terminus and domain IVa. Amended structures for the remaining cryptophyte introns are presented herein (Figs. S18-S22).

Red algal twintrons

Introns nested within other introns (or twintrons) were first reported in the *Euglena gracilis* plastid (Copertino and Hallick, 1991). Since then, group II/III twintrons have been reported at multiple sites in complete Euglenozoa plastid genomes (*E. gracilis* and *Monomorphina aenigmatic*a; Pombert et al. 2012) and from the plastid genomes of the cryptophytes *Rhodomonas salina* and *Hemiselmis andersenii* (Maier et al. 1995; Khan et al. 2007 [however see discussion, above]). Twintrons have also been described in the prokaryotes *Thermosynechococcus elongatus,* a thermophilic cyanobacterium (Mohr et al. 2010) and in *Methanosarcina acetivorans,* an archaebacterium (Dai and Zimmerly, 2003). Here we provide the first description of twintrons in rhodophyte plastid genomes, and the first known report of an inner intron (*mat1f*) found nested within two different outer introns (while also inserted in a third gene). The plastid genomes of three *P. purpureum* strains each contain two twintrons encoding *mat1fa* and *mat1fb* (Fig. 2A, 2B) that are bounded by different outer introns inserted in the *rpoC2* and *atpI* genes, respectively. Two strains contain a copy of the inner intron/IEP inserted singly within the *atpB* gene as *mat1fc*. Alignment of the outer and inner twintron regions together with the

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.729v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

other introns shows that the two different twintrons have very similar structures (Fig. S3). Despite partial sequence similarity (78.2% sequence identity in pairwise comparisons), the two outer introns have similar IEP remnants. The IEPs are truncated at the same site, likely due to a partial protein deletion. Approximately 130 nt and 555 nt, respectively, remain in the 5' and 3' regions of the former IEP in the external introns. Presumably, the later insertion of the inner intron happened at the same binding site (85 nt further downstream from the excision site). Our analyses show that the closely related outer introns int.b (atpI) and int.b (rpoC2; Fig. 3) in *P. purpureum* retain IEP remnants that have been truncated in the same region due to inner intron insertion at the same DIV target site. Of future interest is to study the splicing of these red algal twintrons to confirm that excision occurs in consecutive steps as in other chloroplast twintrons (Copertino et al. 1992).

Conclusions

In summary, our results support a relatively simple explanation for the origin of a complex family of group II introns in the plastid genome of different *P. purpureum* strains (see Fig. 2A). We suggest that the common ancestor of these five strains contained several IEP-encoding group II introns that may trace their origin to the cyanobacterial primary plastid endosymbiont. In turn, the Cryptophyta may have acquired these group II introns during the secondary endosymbiosis of a red alga potentially related to a *Porphyridium*-like donor. These hypotheses require testing with additional plastid genome data from red algae and cryptophytes. Regardless of the time or mode of origin

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.729v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

our data suggest that seeds for nuclear spliceosomal introns exist in red algae vis-à-vis organelle encoded group II introns.

It is also clear that during evolution, some mobile group II introns lose their IEP either by complete deletion, partial degeneration (i.e., loss of the YADD motif), or by point mutations that resulted in-frame stop codons (as in the En domain). All of these events create mobility-impaired introns that are stably inherited in descendant lineages. However, some *P. purpureum* IEPs recovered here have not undergone deleterious change and apparently retain mobility. These mobile introns are inserted in different genes in the plastid genomes, including the intron encoding the *mat1f* IEP that created two different twintron combinations. We suggest that *P. purpureum* is a potentially valuable eukaryote model for understanding the evolution of recently mobile group II introns. The presence of active IEPs in the *P. purpureum* plastid genome also makes this species a good candidate for biotechnological applications, for example *via* the insertion of IEP encoded foreign genes in plastid genomes (Enyeart et al. 2014). In this regard, *P. purpureum* synthesizes compounds of interest such as unsaturated fatty acids and photosynthetic pigments (Lang et al. 2011) and plastid transformation is stable, which is rare for red microalgae (Lapidot et al. 2002).

Acknowledgments

We thank Nicolas Toro for sharing his RT domain-based IEP protein alignment. We are grateful to members of the Genome Cooperative at the Rutgers School of Environmental and Biological Sciences for supporting this research. The authors have no conflict of interest with respect to this work.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403-410.

Archibald, JM. 2009. The puzzle of plastid evolution. *Current Biology* 19:R81-88.

Bhattacharya D, Yoon HS, and Hackett JD. 2004. Photosynthetic eukaryotes unite: endosymbiosis connects the dots. *Bioessays* 26:50-60.

Bhattacharya D, Price DC, Chan CX, Qiu H, Rose N, Ball S. et al. 2013. Genome of the red alga *Porphyridium purpureum*. *Nature Communications* 4:1941.

Byun Y, Han K. 2009. PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots. *Bioinformatics* 25:1435-1437.

Cech TR. 1986. The generality of self-splicing RNA: Relationship to nuclear mRNA splicing. *Cell* 44:207-210.

Copertino DW, Hallick RB. 1991. Group II twintron: an intron within an intron in a chloroplast cytochrome b-559 gene. *The EMBO Journal* 10:433-442.

Copertino, DW, Shigeoka S, Hallick RB. 1992. Chloroplast group III twintron excision utilizing multiple 5'- and 3'-splice sites. *The EMBO Journal* 11:5041-5050.

Dai L, Zimmerly S. 2003. ORF-less and reverse-transcriptase-encoding group II introns in archaebacteria, with a pattern of homing into related group II intron ORFs. *RNA* 9:14-19.

Doolittle WF. 2014. The trouble with (group II) introns. *Proceedings of the National Academy of Sciences of the United States of America* 111:6536-6537.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.729v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

334 Enyeart PJ, Mohr G, Ellington AD, Lambowitz AM. 2014. Biotechnological applications
335 of mobile group II introns and their reverse transcriptases: gene targeting, RNA-
336 seq, and non-coding RNA analysis. *Mobile DNA* 5:2.
337 Jones RF, Speer HL, Kuyr W. 1963. Studies on the growth of the red alga *Porphyridium*
338 *cruentum*. *Physiologia Plantarum* 16:636-643.
339 Khan H, Archibald JM. 2008. Lateral transfer of introns in the cryptophyte plastid
340 genome. *Nucleic Acids Research* 36:3043-3053.
341 Khan H, Parks N, Kozera C, Curtis BA, Parsons BJ, Bowman S, Archibald J. 2007.
342 Plastid genome sequence of the cryptophyte alga *Rhodomonas salina*
343 CCMP1319: lateral transfer of putative DNA replication machinery and a test of
344 chromist plastid phylogeny. *Molecular Biology and Evolution* 24:1832-1842.
345 Lambowitz AM, Zimmerly S. 2011. Group II introns: mobile ribozymes that invade
346 DNA. *Cold Spring Harbor Perspectives in Biology* 3:a003616.
347 Lang I, Hodac L, Friedl T, Feussner I. 2011. Fatty acid profiles and their distribution
348 patterns in microalgae: a comprehensive analysis of more than 2000 strains from
349 the SAG culture collection. *BMC Plant Biology* 11:124.
350 Lapidot M, Raveh D, Sivan A, Arad SM, Shapira M. 2002. Stable chloroplast
351 transformation of the unicellular red alga *Porphyridium* species. *Plant Physiology*
352 129:7-12.
353 Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H,
354 Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins
355 DG. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947-2948.

356 Maier, UG, Rensing SA, Igloi GL, Maerz M. 1995. Twintrons are not unique to the
357 *Euglena* chloroplast genome: structure and evolution of a plastome cpn60 gene
358 from a cryptomonad. *Molecular & General Genetics* 246:128-131.
359 Mohr G, Ghanem E, Lambowitz AM. 2010. Mechanisms used for genomic proliferation
360 by thermophilic group II introns. *PLoS Biology* 8:e1000391.
361 Moran JV, Zimmerly S, Eskes R, Kennell JC, Lambowitz AM, Butow RA. et al. 1995.
362 Mobile group II introns of yeast mitochondrial DNA are novel site-specific
363 retroelements. *Molecular and Cell Biology* 15:2828-2838.
364 Pombert JF, James ER, Janouškovec J, Keeling PJ. 2012. Evidence for transitional stages
365 in the evolution of euglenid group II introns and twintrons in the *Monomorphina*
366 *aenigmatica* plastid genome. *PLoS One* 7:e53433.
367 Qu G, Dong X, Piazza CL, Chalamcharla VR, Lutz S, Curcio MJ. et al. 2014. RNA-RNA
368 interactions and pre-mRNA mislocalization as drivers of group II intron loss from
369 nuclear genomes. *Proceedings of the National Academy of Sciences of the United*
370 *States of America* 111:6612-6617.
371 Ragan MA, Bird CJ, Rice EL, Gutell RR, Murphy CA, Singh RK. 1994. A molecular
372 phylogeny of the marine red algae (Rhodophyta) based on the nuclear small-
373 subunit rRNA gene. *Proceedings of the National Academy of Sciences of the*
374 *United States of America* 91:7276-7280.
375 Robart AR, Chan RT, Peters JK, Kanagalaghatta RR, Toor N. 2014. Crystal structure of a
376 eukaryotic group II intron lariat. *Nature* 514:193-197
377 Rogozin IB, Carmel L, Csuros M, Koonin EV. 2012. Origin and evolution of
378 spliceosomal introns. *Biology Direct* 7:11.

379 Sharp PA. (1991) Five easy pieces. *Science* 254:663.
380 Simon DM, Kelchner SA, Zimmerly S. 2009. A broadscale phylogenetic analysis of
381 group II intron RNAs and intron-encoded reverse transcriptases. *Molecular*
382 *Biology and Evolution* 26:2795-2808.
383 Tajima N, Sato S, Maruyama F, Kurokawa K, Ohta H, Tabata S. et al. 2014. Analysis of
384 the complete plastid genome of the unicellular red alga *Porphyridium purpureum*.
385 *Journal of Plant Research* 127:389-397.
386 Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular
387 Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution*
388 30:2725-2729.
389 Toro N, Martínez-Abarca F. 2013. Comprehensive phylogenetic analysis of bacterial
390 group II intron-encoded ORFs lacking the DNA endonuclease domain reveals
391 new varieties. *PLoS One* 8:e55102.
392 Toor N, Hausner G, Zimmerly S. 2001. Coevolution of group II intron RNA structures
393 with their intron-encoded reverse transcriptases. *RNA* 7:1142-1152.
394 Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction.
395 *Nucleic Acids Research* 31:3406-3415.
396
397
398
399
400
401

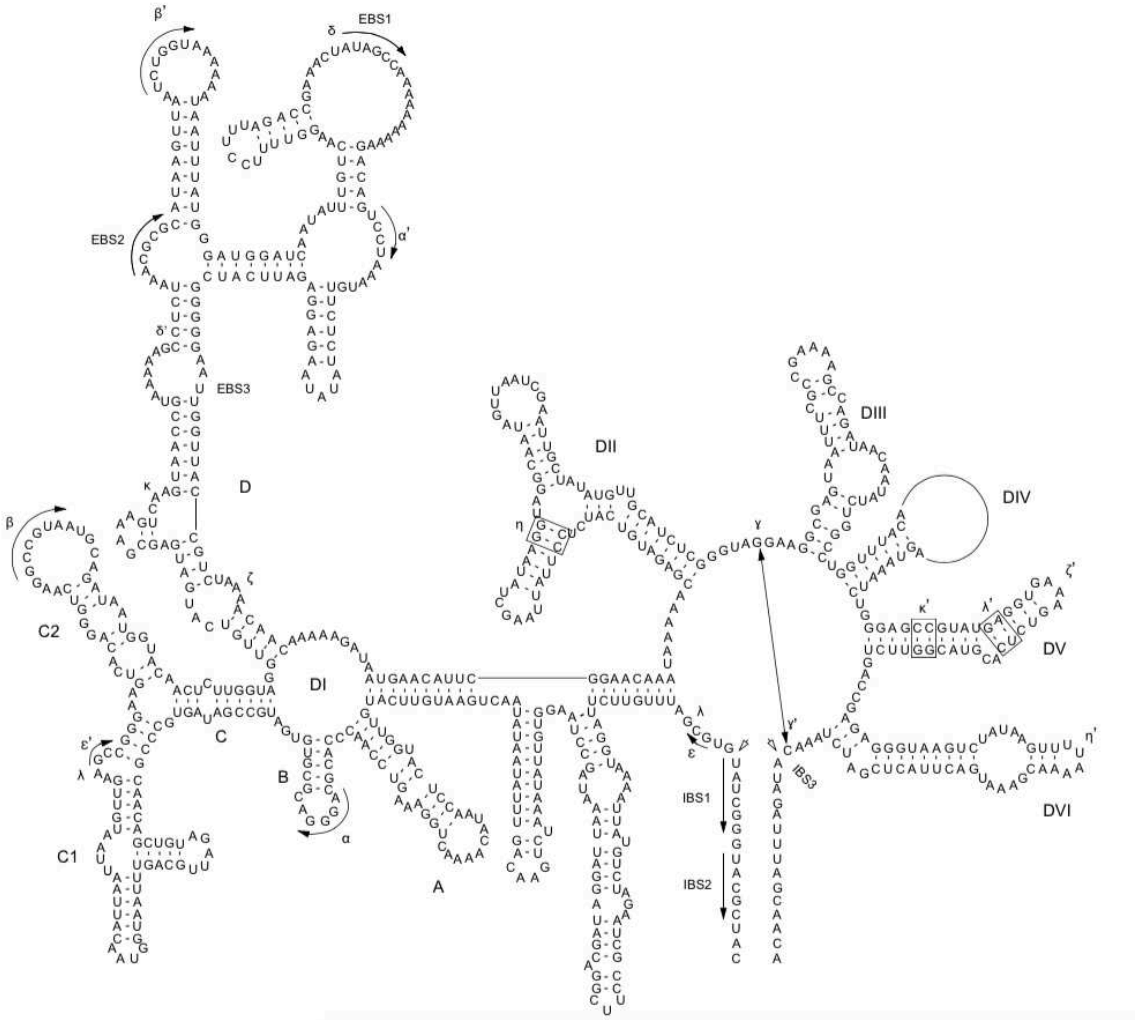
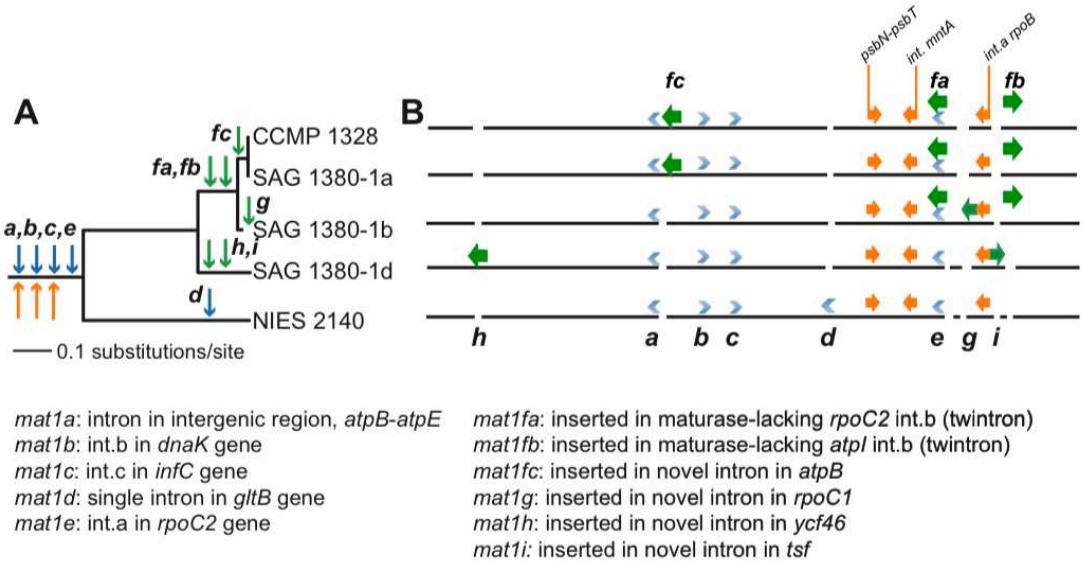


Figure 1. *P. purpureum* group IIB intron structure. Predicted structure of the *rpoC1* intron containing the *mat1g* IEP. The structure is composed of six conserved domains (DI-DVI). Exon and intron binding site (EBS and IBS) and Greek letters indicate nucleotide sequences involved in long-range tertiary interactions. The IEP is located in the DIV domain



412

413

414 **Figure 2. Evolution of group II introns and IEPs in *Porphyridium* strains**.

415 (A) Neighbor-joining phylogenetic tree (uncorrected *p*-distance, 100 bootstrap replicates)

416 built using 332 SNPs identified in these plastid genomes. Blue arrows illustrate the

417 distribution of group II introns described by Tajima et al. (2014), green arrows denote

418 group II introns described in this study, and orange arrows denote IEP-lacking (or

419 degenerate) group II intron structures defined here. (B) Location of group II introns/IEPs

420 in the plastid genomes. Blue chevrons illustrate introns described by Tajima et al. (2014),

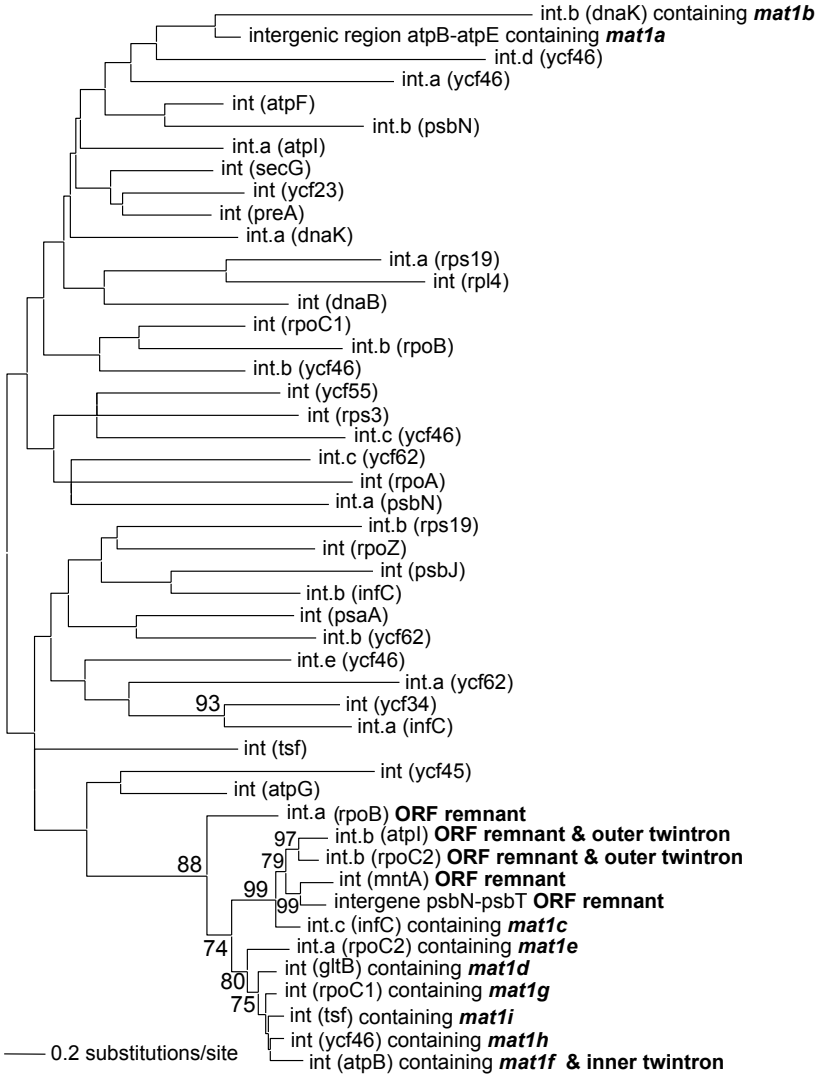
421 green arrows denote introns newly described in this study, and orange arrows illustrate

422 IEP-lacking (or degenerate) group II intron structures defined here.

423

424

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.729v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2014, publ: 19 Dec 2014

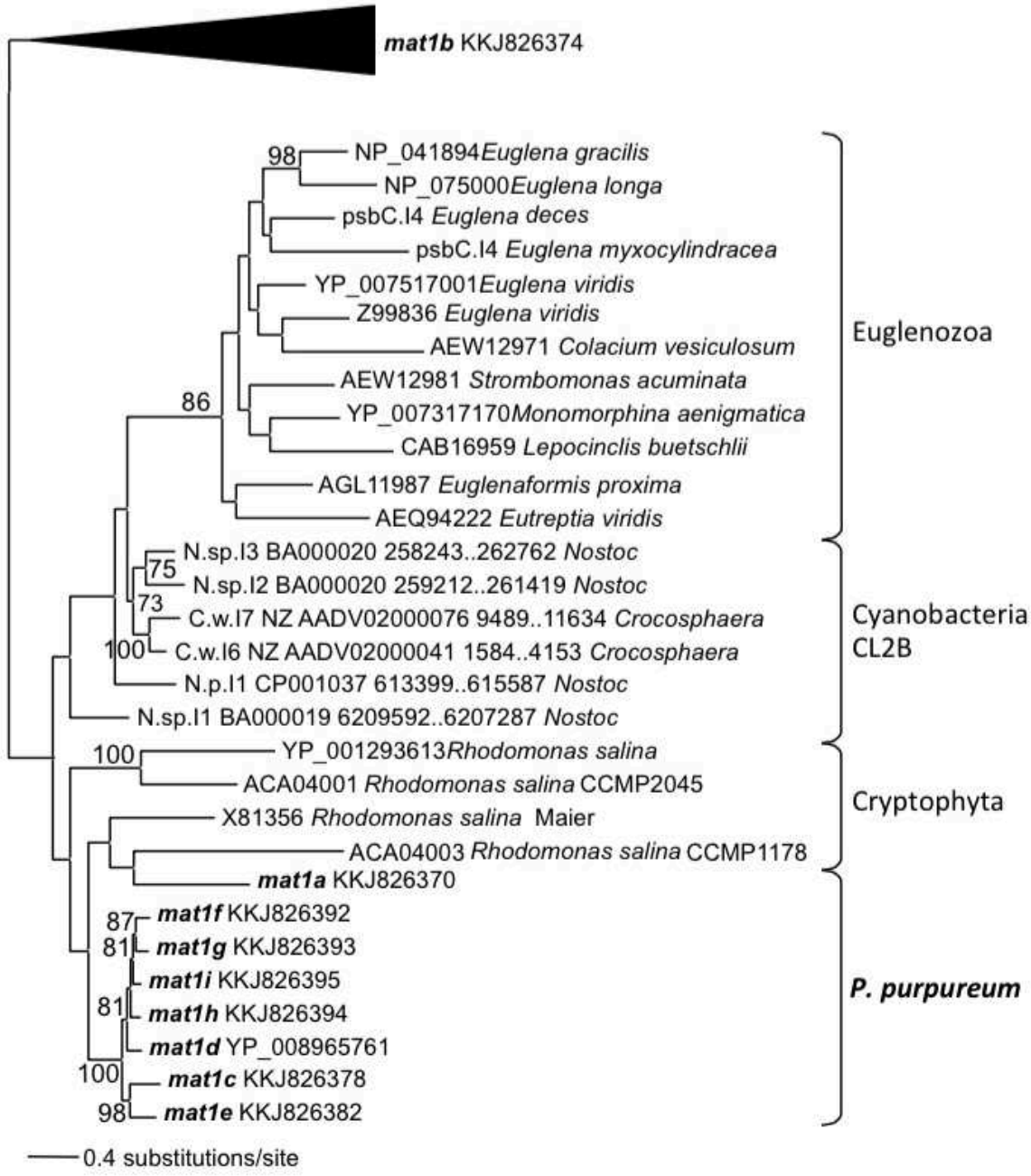


425

426 **Figure 3. Phylogeny of *P. purpureum* group II introns**. Maximum likelihood tree; only

427 bootstrap values >70% are shown. To avoid long-branch attraction, the IEP or IEP

428 remnant sequences (indicated in bold) were removed from the alignment.



429

430 **Figure 4. Phylogeny of CL2B group II IEPs**. The nine plastid-encoded IEP sequences

431 from *P. purpureum* were added to selected sequences from the bacterial group II intron

432 database, together with Cryptophyta and Euglenozoa IEPs (ML, bootstrap >70%). The

433 tree is rooted with proteins from the CL2A, CL1A, and CL1B groups (including the

434 *mat1b* IEP).