Mol Gen Genet (2001) 264: 682–690
DOI 10.1007/s004380000355

ORIGINAL PAPER

N. A. Doetsch · M. D. Thompson · M. R. Favreau
R. B. Hallick

Comparison of *psbK* operon organization and group III intron content in chloroplast genomes of 12 Euglenoid species

Received: 13 April 2000 / Accepted: 14 August 2000 / Published online: 21 October 2000

Abstract A novel mixed operon has been identified in the photosynthetic protist *E. gracilis*. The genes for *psbK*, *ycf12*, *psaM*, and *trnR* are co-transcribed. The resulting tetracistronic transcripts are processed through endonucleolytic cleavage of the intergenic spacers and intron splicing to form three mature monocistronic mRNAs and a tRNA. A group III twintron and a group III intron are located in *psbK*. Another group III intron is found in *ycf12*. The *psbK* operon has been cloned by PCR amplification from nine related Euglenoid species. In each species, the gene order and content of the *psbK* operon is conserved. The *psbK* operons contain phylogenetically conserved eubacterial promoter, translational, and 3′ processing elements. Intron content varies significantly from species to species. Based on a comparison of the intron content with the results of phylogenetic analysis, group III intron evolution within the Euglenoid lineage is much more complex than previously believed.

Key words *Euglena* · *psbK* · Intron · Evolution · Chloroplast

Introduction

Chloroplast genomes contain genes for photosynthetic proteins and genes that encode components of prokaryotic-like translation and transcription machinery. A feature that is conserved between chloroplasts and prokaryotes is the organization of functionally related genes into operons, which are co-transcribed and subsequently processed by endonucleolytic cleavage into monocistronic transcripts. Many chloroplast operons contain transcriptional initiation sites similar to the prokaryotic consensus "–35" and "–10" elements (Stern et al. 1997). Likewise, many translational initiation sites have eubacterial-type Shine-Dalgarno sequences (GGAGG) located about 10 nt upstream of the start codon. As in prokaryotes, the 3′ untranslated regions (UTRs) of many chloroplast coding sequences contain an inverted repeat. In prokaryotes this 3′ stem-loop structure is often utilized as a translation termination signal. However, chloroplast 3′ stem-loops are not efficient translational terminators, but instead appear to be essential for mRNA stability and/or 3′ processing (Stern and Gruissem 1987; Rott et al. 1996).

The *Euglena gracilis* chloroplast genome contains an extraordinarily high number of introns (Hallick et al. 1993). There are at least 88 group II introns, 65 group III introns, and 15 twintrons (introns inserted into other introns). Group III introns are abbreviated versions of group II introns, lacking at least four of the six conserved domains (Copertino and Hallick 1993), including the catalytic domain V necessary for group II intron self-splicing. Key features of group III introns are a relaxed group II-like 5′-boundary sequence, a narrow size distribution of 100 ± 25 nt, and a functionally and phylogenetically conserved domain VI (Copertino et al. 1994; Doetsch et al. 1998). Domain VI contains the branch A nucleotide, an unpaired adenine residue located 7–8 nt before the splice site, that initiates the two-step splicing reaction with a nucleophilic attack on the 5′ splice boundary, yielding a lariat intermediate.

The large number of introns in the *E. gracilis* genome and the existence of twintrons have been used to predict the evolutionary history of Euglenoid introns. The insertion of a mobile intron into another, already fixed intron is the most likely method of twintron formation.

Communicated by R. G. Herrmann

N. A. Doetsch · M. D. Thompson
M. R. Favreau · R. B. Hallick (✉)
Department of Biochemistry,
University of Arizona,
1041 E. Lowell Street,
Tucson, AZ 85721-0088, USA
E-mail: hallick@u.arizona.edu
Tel.: +1-520-6213026
Fax: +1-520-6211697

N. A. Doetsch
Department of Molecular and Cellular Biology,
University of Arizona, Tucson, AZ 85721, USA

Intron-encoded maturases, capable of promoting intron mobility in yeast, have been found in the *Euglena* genome within both a group II and a group III intron (Zhang et al. 1995; Doetsch et al. 1998). Group II introns are relatively late evolutionary additions to Euglenoid chloroplast genomes (Thompson et al. 1995). In the most highly derived species, *E. gracilis*, the number of *rbcL* group II introns is much greater (9) than that found in basally branching (0) or intermediate species (7). The most parsimonious interpretation of these data is that the ancestral genome was either intron-free or contained very few introns. The ability of once-mobile introns to insert themselves into novel positions, most probably with the help of intron-encoded proteins, apparently accounts for the high numbers present in the extant *E. gracilis* genome.

Here we describe the identification and evolutionary analysis of a new polycistronic operon from *E. gracilis*. The genes for *psbK*, *ycf12*, *psaM*, and *trnR* are co-transcribed, yielding a mixed operon encoding genes for proteins of photosystems I and II, as well as a tRNA-Arg gene. The *psbK* operon is very suitable for an evolutionary study of group III intron and twintron content, comparable to earlier studies on *Euglena* group II introns and twintrons. Of particular interest is the question whether group II and group III introns might have a common or independent evolutionary histories. To further investigate this question, the *psbK* operon was cloned and sequenced from nine related Euglenoid species.

Materials and methods

Euglena cultures and nucleic acid extraction

The following Euglenoid strains were obtained from the University of Texas Culture Collection: *E. gracilis* var. Z strain (UTEX 753), *E. stellata* (UTEX 327), *E. viridis* (UTEX 85), *E. myxocylindracea* (UTEX 1989), *E. deses* (UTEX LB 370), *E. anabaena* (UTEX 373), *Phacus accuminata* (UTEX LB 1288), *E. spirogyra* (UTEX LB1307), *E. pisciformis* (UTEX 1604), and *E. sanguinea* (UTEX 2345). *E. mutabilis* was obtained from the laboratory of Dr. Rich Triemer.

E. gracilis liquid cultures were maintained in either heterotrophic (*Euglena* broth; Sigma) or photoautotrophic medium (Hallick et al. 1982) under continuous illumination as described. *E. myxocylindracea*, *E. stellata*, and *E. mutabilis* were also grown in liquid heterotrophic *Euglena* broth, although growth rates were not as high as those of *E. gracilis*, and culture densities were lower. *E. stellata* was also grown on solid slants of Proteose medium (Starr and Zeikus 1993).

Total nucleic acid extracts were prepared as previously described (Thompson et al. 1995), either directly from cultures obtained from the UTEX collection or following additional growth in our laboratory.

Northern hybridization analysis

^{32}P-labeled RNA probes were synthesized in vitro as previously described (Hong et al. 1995), using T7 RNA polymerase and clones of fully spliced *psbK* operon cDNAs as template. Aliquots (5 μg) of total nucleic acid extracted from *E. gracilis* grown either photoautotrophically or heterotrophically were fractionated on a 3% polyacrylamide gel containing 7 M urea, and transferred to Gene Screen membrane by electroblotting (Ausubel et al. 1995). Prehybridization was carried out for 2 h in hybridization buffer (5 × SSC, 2 × Denhardt's reagent, 1% SDS, 50% formamide, 1 mg/ml salmon sperm). RNA probes specific to either fully spliced *psbK* alone (pEZC1073 linearized with *Hin*dIII), a co-transcript of fully spliced *psaM* and *trnR* (pEZC1097 linearized with *Hin*dIII), or a transcript of *ycf12*, *psaM*, and *trnR* (pEZC 1097 linearized with *Hin*cII) were added to a concentration of 10^6 dpm/ml of hybridization buffer and incubated for 16 h at 50 °C. Filters were washed twice at room temperature with 0.1% SDS, 0.1% SSC, and twice at 60 °C, and exposed to X-ray film.

PCR amplification, cDNA synthesis, and sequencing

The *psbK* operon was isolated from the Euglenoid species *E. gracilis*, *E. deses*, *E. mutabilis*, *E. viridis*, *E. myxocylindracea*, *E. sanguinea*, *E. stellata*, *Lepocinclis beutschlii*, and from *P. accuminata*, by PCR amplification of 1 μl of total nucleic acid extract (total nucleic acid, approximately 0.1 μg/μl), with the synthetic oligonucleotides P1 and C1 (all primer sequences and co-ordinates are listed in Table 1). Primer P1 anneals to the *trnT* gene directly upstream of the *psbK* operon, and is identical to the mRNA-like strand of *E. gracilis* corresponding to co-ordinates 30,806–30,829 (all co-ordinates refer to Genbank X70810, unless otherwise specified). Primers C1 and C2 are identical to the cDNA-like strand of *E. gracilis trnR* (coordinates 28,698–29,718 and 29,718–29,738 respectively). Amplification was done in an Eppendorf MasterCycler Gradient with the reaction parameters: 94 °C 2′; 35 cycles of 94 °C for 1 min, 50 °C for 1 min and 72 °C for 1 min; and 72 °C for 10 min. The *E. anabaena* sequence was amplified under identical conditions except that the primers used were P1 and C2 at an annealing temperature of 40.7 °C. The resulting PCR products were either cloned into the *Eco*RV-digested, ddT-tailed Bluescript vector pKS+ by the method of Holton and Graham (1991), or were directly ligated into the pGEM T-easy vector (Promega) following the manufacturer's instructions. Each clone was sequenced

Table 1 Sequences and accession coordinates of primers used in PCR and RT-PCR reactions

Oligo	Sequence (5′ → 3′)	Coordinates	Accession No.
P1	GCTCTACCACTGAGCTAAAAAGGC	30,829–30,806	X70810
P2	GGGAAAATAAAATGTC	30,707–30,721	X70810
P3	TTACCAGAA(CG)(AC)(AT)TATGCTCC	30,523–30,552	X70810
P4	GTTTCAAATTAGTTTAGATATG	81–103	AF241283
P5	GATATCTAATATGAGAATAACG	103–124	AF241283
P6	CTAACAAAAGTTTTAATAAC	57–76	AF241281
C1	GTCACAGATAGGATTCGAACCT	29,698–29,719	X70810
C2	TACACAAAACAACTTAGAAGG	29,718–29,738	X70810
C3	5′-TTTACTGAGGCCTGCC-3′	30,247–30,263	X70810
C4	CCTAGTTT(AT)A(AT)G(CG)TGAA-GAAAGC(AT)GG	29,842–29,868	X70810
C5	ATTAATTTGTTTCTTACAATG	570–590	AF241283
C6	ACTTTATAGATTTAACGG	538–555	AF241283

completely on both strands by the Sequencing Facility at the University of Arizona (Tucson, Ariz.).

Reverse-transcriptase reactions using 1 µl of total nucleic acid as templates were carried out at 37 °C for 1.5 h as described (Copertino and Hallick 1991). PCRs on the cDNA were done as described above, with the following annealing temperatures for each primer set: P1 + C1 and P3 + C4, 50.0 °C; P3 + C1, P5 + C6, and P6 + C4, 53.6 °C; P1 + C2, 40.7 °C; P2 + P3, 45.6 °C; P4 + C6, 48.0 °C.

Computer analysis

Localization of ORFs and determination of putative amino acid sequences were done with the computer program DNAStrider. Nucleotide sequence alignments were carried out using the PILEUP program (GCG Sequence Analysis Package, version 8.0, Madison, Wis.). Promoter alignments were then adjusted manually. Protein alignments were also done with PILEUP.

The intergenic regions *psbK-ycf12*, *ycf12-psaM*, and *psaM-trnR* were analyzed for RNA secondary structures using the FOLDRNA program in the GCG package. The *ycf12* group II introns of *E. stellata* and *E. sanguinea* were folded with FOLDRNA and then adjusted manually according to the group II intron consensus determined by Michel and Ferat (1995), and the *Euglena* group II intron consensus of Thompson et al. (1997) (not shown). Domain VI structures of all group III introns were folded manually. Analysis of group III domain I folding was also done with FOLDRNA using intron sequences lacking domain VI.

Results

The *psbK* operon is comprised of the *psbK, ycf12*, and *psaM* cistrons

In the *E. gracilis* chloroplast genome the coding region of the *psbK* gene is followed by the genes for the hypothetical chloroplast protein *ycf12*, the photosystem I gene *psaM*, and *trnR*. The tRNA gene, *trnT*, lies directly upstream of the 5′-end of the *psbK* operon, on the opposite strand of DNA. The region upstream of the *psbK* gene contains the transcriptional initiation site for the *psbK* operon (Stevenson 1994). Furthermore, *psbK* is the third most highly expressed mRNA in the *Euglena* chloroplast (Stevenson 1994). To determine if the downstream *ycf12*, *psaM*, and *trnR* cistrons are co-transcribed with *psbK*, transcripts were analyzed by Northern hybridization. The results are shown in Fig. 1. Probes specific for the fully spliced *psbK* gene bind to three major RNAs. These RNAs correspond in size to fully spliced transcripts containing *psbK-ycf12-psaM*, *psbK-ycf12*, and the monocistronic *psbK*. Probes specific for *ycf12*, *psaM* and *trnR* RNAs, or just *psaM* and *trnR* also hybridized to fully spliced tri-, di- and monocistronic transcripts (data not shown). Therefore, *psbK*, *ycf12*, and *psaM* are in a polycistronic operon. Only monocistronic *trnR* transcripts were evident. However, there is a precedent for failure to detect mixed mRNA-tRNA transcripts by Northern hybridization. The *trnK* gene is cotranscribed with the *psaA* operon, but is processed so rapidly that full-length transcripts could not be identified by Northern blotting, S1-nuclease protection assays, or primer extension. Co-transcription was demonstrated only by reverse-transcriptase PCR analysis (Stevenson and Hallick 1994). Further experiments addressing the question whether *trnR* is co-transcribed with *psbK* are discussed below.

PCR amplification of the *psbK* operon from nine diverse Euglenoids

To determine whether the gene order and intron content is evolutionarily conserved, *psbK* operons were analyzed from 11 additional Euglenoid species. Primers specific to highly conserved regions of the *trnR* (C1) and *trnT* genes (P1), were used to amplify the putative *psbK* operons by PCR. If all amplification products have the same gene and intron content as *E. gracilis*, the resultant PCR products would be approximately 1100 nt in length. A single PCR product was obtained from each species, except *P. accuminata*, which yielded two major bands, including a non-specific band around 1400 nt long and a specific band of about 900 nt, as well as some additional minor bands. PCR product sizes for the 11 species ranged from 700–1400 nt, which is suggestive of

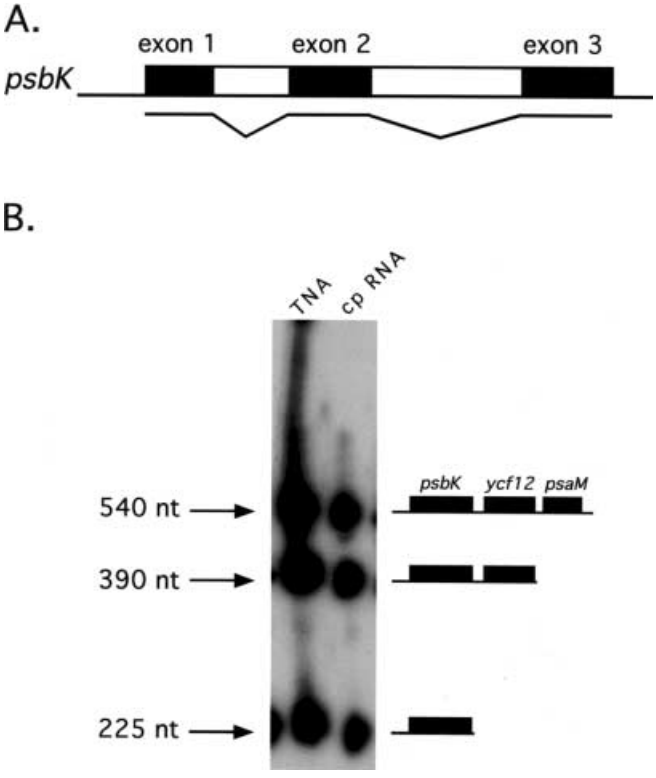

Fig. 1A, B Northern hybridization analysis of the *E. gracilis psbK* operon. **A** A riboprobe was generated from the clone pEZC 1073, which contains the fully spliced *psbK*. The diagram depicts the intron-exon structure of *psbK*. Exon sequences are depicted as *black boxes* and introns as *white boxes*. **B** Results of Northern analysis with the exon probe. Each lane contains 5 µg of either total nucleic acid (TNA, lane 1) or chloroplast (cp) RNA (lane 2). The size given for each RNA transcript detected, shown on the *left*, is approximate since the exact processing sites are unknown, but agrees with the sizes expected for the transcripts schematically shown on the *right*

conserved genes with variable intron content. Each amplified DNA, except those of *E. pisciformis* and *E. spirogyra*, was cloned and sequenced (GenBank Accession Nos. AF241276–84). The *psbK* operons are 730, 769, 869, 875, 951, 998, 1086, 1166, 1263, and 1418 nt in length for *E. mutabilis*, *L. beutschlii*, *E. myxocylindracea*, *P. accuminata*, *E. viridis*, *E. anabaena*, *E. gracilis*, *E. deses*, *E. sanguinea*, and *E. stellata*, respectively. ORFs homologous to the *psbK*, *ycf12*, *psaM*, and *trnR* genes were present in each DNA. Gene order is conserved relative to that of *E. gracilis*. Differences in the length of the operon correlate with the number and size of introns. All of the *psbK* and *ycf12* genes contain introns, except *ycf12* from *E. mutabilis*. None of the 10 *psaM* genes has an intron.

Identification of introns in *psbK* operons

Group III introns are present in *E. deses* (118 nt), *E. stellata* (99 nt), and *E. myxocylindracea* (80 nt) at the same site as the first *E. gracilis psbK* group III intron. In each species *psbK* intron 1 is located in the precursor polypeptide between the 6th and 7th codons proximal to the putative pre-sequence cleavage site. In *P. accuminata*, a 76-nt intron is located between the 7th and 8th codons upstream of the pre-sequence. The remaining five species lack an intron at this site.

All species characterized also had either a group III intron or a putative group III twintron between the 1st and 2nd nucleotides of the 15th codon distal to the predicted precursor cleavage sites. The *E. gracilis* intron in this position is shown below to be a 203-nt group III twintron. Group III introns are present in *E. myxocylindracea* (86 nt), *E. mutabilis* (97), *L. beutschlii* (90), and *P. accuminata* (85). The sizes of the introns in *E. stellata* (201), *E. anabaena* (205), *E. sanguinea* (191), *E. deses* (209), and *E. viridis* (183) were all within the size range expected for group III twintrons. Within each putative twintron, 5′-NUNNG splice sites of an internal group III intron occur approximately 50 nt in from the 5′ end of the twintron. Group III domains VI occur about 100 nt downstream of the 5′ consensus (data not shown). RT-PCR was used to confirm 2/5 twintron predictions (described below). The group III external intron of the twintron at codon 15 of the mature peptide was the most highly conserved of all the *psbK* operon introns, occurring in every species surveyed.

A group III intron also occurs between the 1st and 2nd nucleotides of the 4th codon of *ycf12* in all species except for *E. mutabilis* (Fig. 2). (*P. accuminata* apparently lacks two codons upstream of the intron, but it is located at the identical site with reference to the C-terminal portion of the protein. The exact lengths of the *ycf12* group III introns are as follows: *E. gracilis* 107, *E. stellata* 99, *E. deses* 105, *E. viridis* 100, *E. anabaena* 106, *E. sanguinea* 95, *L. beutschlii* 78, *P. accuminata* 83, and *E. myxocylindracea* 92 nt. In addition, the *ycf12* coding region in both *E. stellata* and *E. sanguinea* was interrupted in different codons by intervening sequences of 367 and 326 nt, respectively. The *E. stellata* 367-nt intron is located between the 2nd and 3rd nucleotides of the 25th codon, and the *E. sanguinea* 326-nt intron is between the 22nd and 23rd codons. These sequences contain the characteristic domain V and the remainder of the six domains consistent with their being *Euglena* chloroplast group II introns.

RT-PCR analysis

To confirm intron and twintron loci and predicted intron-exon boundaries, partially and fully spliced mRNAs were characterized by sequencing of RT-PCR products. Three species, *E. gracilis*, *E. stellata*, and *E. myxocylindracea*, were chosen for detailed analysis (Fig. 2). To determine the intron-exon boundaries of *E. gracilis psbK*, total nucleic acid extract was subjected to reverse transcriptase with primer C3, which is specific for *psbK* exon 3. The cDNA was amplified with primers P2 and C3, which anneal to exons 1 and 3, respectively (data not shown). Processing intermediates were obtained that contained either the partially (internal intron excised) or fully spliced twintron, with or without the first intron. The isolation of partially spliced intermediates containing the external intron of *psbK* intron 2 (*psbK* intron2a), with the internal intron (*psbK* intron 2b) excised, confirm that *psbK* intron 2 is a group III twintron. Not only were cDNAs representing unspliced or partially spliced twintron obtained with a spliced first intron, but cDNA clones containing the unspliced first intron with the fully spliced twintron were also isolated. No partially spliced cDNAs lacking intron 1 were detected. Therefore, splicing may occur by an unordered pathway, with intron 1 splicing being the final reaction.

To characterize the *E. gracilis ycf12* group III intron (*ycf12*i1) splicing reaction, reverse transcription was carried out on *E. gracilis* total nucleic acid with primer C1, located in the *trnR* gene. PCR amplification of the C1 cDNAs was carried out with primer C1 and primer P3, which anneals to the 2nd exon of *psbK*. The cDNAs from this reaction correspond in size to partially spliced and fully spliced intermediates (Fig. 2B). The smallest product is fully spliced. This cDNA spans the splice sites of *ycf12* intron 1 and the *psbK* twintron, confirming the boundary sequences of each group III intron. The fact that cDNA clones were obtained which contain both the *psbK-ycf12-psaM* transcript and the *trnR* transcript also demonstrates that the *trnR* gene is co-transcribed from the *psbK* promoter. Therefore, although RNA blotting failed to detect transcripts from the *psbK* promoter containing the *trnR* coding sequence, the *E. gracilis psbK* operon includes *psbK*, *ycf12*, *psaM*, and *trnR*.

Splicing of the *E. stellata psbK* operon introns was analyzed in three RT-PCR reactions. Reverse transcription of *E. stellata* total nucleic acid extracts with primer C4 (located in *psaM*), and PCR amplification with primers C4 and P3 (in *psbK* exon 2) confirmed the

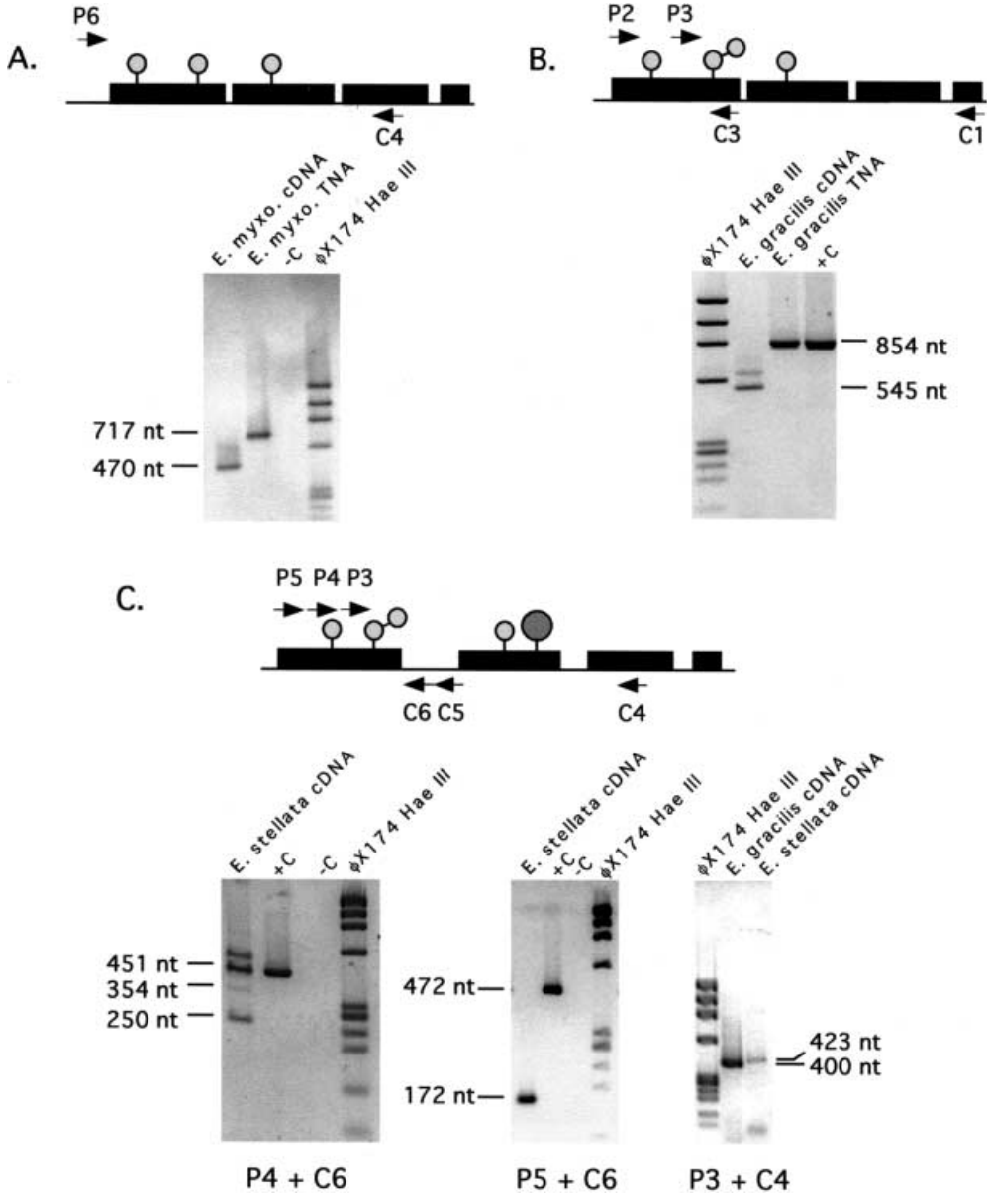

Fig. 2A–C RT-PCR analysis of RNA from selected species confirms the splicing of the introns in the *psbK* operon. The top portion of each panel shows a schematic diagram depicting primer and intron locations. **A** Total nucleic acid (TNA) from *E. myxocylindracea* was subjected to reverse transcription with primer C4; the resulting cDNA was then amplified by PCR with primers P6 and C4 (lane cDNA). As a positive control, TNA from *E. myxocylindracea* was also amplified without prior reverse transcription (lane TNA). **B** *E. gracilis* TNA was reverse transcribed with primer C1 and the cDNA amplified with primers P3 and C1 (lane cDNA). Positive controls included PCR amplification of *E. gracilis* TNA (lane TNA) and amplification of the full-length *psbK* operon clone pEZC 2040.2 (N. Doetsch, manuscript submitted) (lane +C). **C** For the gels shown in the *left* and *center* panels, *E. stellata* TNA was reverse transcribed with primer C5. The resultant cDNA was amplified with P5 and C6 (left) or P4 and C6 (center) (lanes cDNA). The products of reverse transcription of *E. stellata* and *E. gracilis* TNA with primer C4 and PCR amplification with primers P3 and C4 are shown in the panel on the *right*. In all reactions a negative control (–C) was carried out without addition of nucleic acid. A *Hae*III digest of ϕX174 DNA was used as a size marker

splicing of the *ycf12* group III, *ycf12* group II, and the *psbK* group III twintron (Fig. 3). An additional RT-PCR reaction with RT primer C5, and PCR primers C6 and P4 (located in exons 1 and 3), yielded a single band corresponding in size to the fully spliced *E. stellata* product (*psbK* i1 and *psbK* i2 both spliced). To demonstrate that the ~200-nt intervening sequence at the *psbK* intron 2 site is a twintron, *E. stellata* total nucleic acid was reverse transcribed with C5 and PCR amplified with C6 and P5 (Fig. 2C). The P4 primer spans the intron-exon boundary of intron 1, and should therefore not generate spliced products containing a spliced intron 1. Three bands were obtained which, when sequenced, corresponded to fully unspliced (451 nt), partially spliced (internal intron spliced, 354 nt) and fully spliced transcripts (250 nt).

The predicted splice site of the first *psbK* intron of *E. myxocylindracea* results in an unexpectedly short

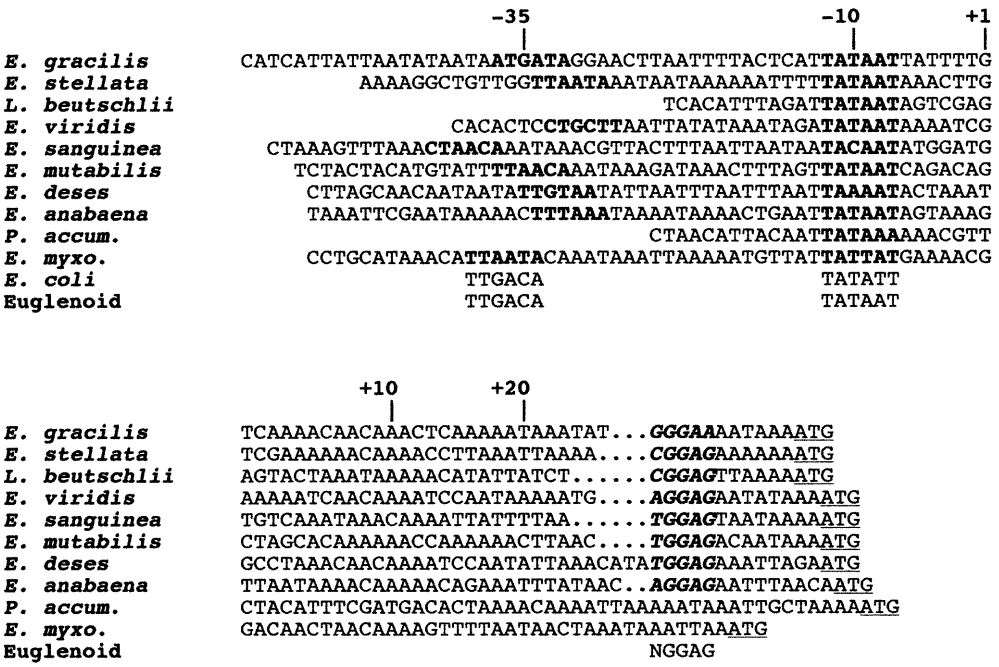

Fig. 3 Alignment of the *psbK* operon promoter regions. Conserved features include the –10 element (indicated in *bold*) and the ribosome binding site (*bold italics*). Putative –35 elements are also shown in *bold*. The *psbK* initiator codon is *underlined*. *E. coli* and Euglenoid consensus sequences are included

exon 1, and also in an intron length of 80 nt. This is much smaller than is customary for group III introns. In addition, the 2nd *psbK* intron is most likely to be a single intron in *E. myxocylindracea*, not a twintron. To test the predicted splice boundaries of the *E. myxocylindracea* introns, *E. myxocylindracea* total nucleic acid was reverse transcribed with primer C4. The resultant cDNA was amplified with primer C4 and primer P6, located in the *psbK* promoter region. This reaction resulted in a product of 470 nt in length, shown in Fig. 2A. The size of the product is consistent with a mature mRNA transcript after splicing of the group III introns *psbK*i1, *psbK*i2, and *ycf12*i1. The predicted splice sites were confirmed by sequencing.

Many of the group III introns identified in this study are among the smallest yet discovered. At least two have lengths in the 70–79 nt range, and many are within 80–89 nt long. According to the intron 5′-boundary sequences derived from amino acid comparison and confirmed by RT-PCR analysis, neither *E. stellata psbK* intron 2a (external) or *E. stellata ycf12* intron 2 (group II intron) contains an exact match to the group III consensus of 5′-NUNNG. The sequence for *E. stellata psbK* intron 2a is 5′-GTGA*A*, and that of *E. stellata ycf12* intron2 is 5′-G*A*GCG. All other introns identified in this study contain the canonical group III intron 5′-boundary motif (described below). Many examples now exist of Euglenoid group II and group III introns without a U in the 2nd position or a G in the 5th.

Intron secondary structure analysis

Each of the 30 new group III introns identified in this study has a potential domain VI with a branch A at position –7 or –8 proximal to the 3′-splice site. Homologous domains VI were subjected to phylogenetic comparisons. As expected, the structure of domain VI was conserved between homologous introns. Domain VI consists of a small stem-loop generally about 8–9 nt in length with an unpaired A residue located in the middle of the 3′-side of the stem. The unpaired A is 7 or 8 nt upstream of the 3′-intron boundary. The branch A nucleotide was conserved in every intron identified in this study.

Sequences of *ycf12* intron 1 from which domain VI had been removed were also examined for conserved putative domain I secondary structures by RNAFOLD (Wisconsin GCG package) and manual folding analysis. All 10 intron sequences had the potential to fold into a single stem-loop structure. This finding is consistent with the structure previously proposed for domain I. Group II intron-like exon binding sequences could not be identified in the domain I loop regions.

Conservation of transcription, translation, and 3′-end processing elements

The region upstream of *psbK* in *E. gracilis* contains the promoter for the *psbK* operon (Stevenson 1994). To test for evolutionarily conserved promoter motifs among the 10 Euglenoid species, the regions 5′ to the start codon of *psbK*, excluding the *trnT* sequence, were aligned (Fig. 3). A region analogous to the prokaryotic –10 Pribnow box was present in each species. A potential transcription start site was identified 10 nt downstream of this conserved TATATT element. Based on the putative promoter, the starting nucleotide in the majority of the Euglenoid species would be G, with the exception of the *E. anabaena* and *P. accuminata* transcripts, which are predicted to start with a T. The consensus for the

Euglenoid –10 motif is TATAAT, which differs at 1/7 positions from that of *Escherichia coli* (TATATT). This region is very well conserved. Among 10 species, only three contain a single nucleotide difference. Regions similar to the –35 regions (TTGACA) might be present, but their distance from the –10 consensus sequence was highly variable (16–24 nt).

The optimal ribosome binding site (rbs) for *E. gracilis* chloroplast ribosomes is GGGAG (Steege et al. 1982). Each *psbK* 5′-UTR, except those of *P. accuminata* and *E. myxocylindracea*, has a close match to this motif centered 10–13 nt upstream from the *psbK* methionine initiation codon. The distance between the *psbK* rbs and the initiation codon is identical to that in the previously determined *atpH* promoter (Betts and Spremulli 1994). An alignment of the upstream *psbK* regions from all species is shown in Fig. 3. The Euglenoid *psbK* consensus is GGAG. The *ycf12* and *psaM* upstream regions do not contain similar ribosomal binding motifs. However these regions are highly AU-rich, which has been proposed to be necessary for translational initiation of *E. gracilis* genes lacking conventional Shine-Dalgarno sequences (Wang et al. 1989; Stern et al. 1997).

A feature common to many chloroplast coding sequences is an inverted repeat sequence within the 3′UTR (Stern et al. 1989). Inverted repeats are often found at the 3′ end of *E. gracilis* operons. Secondary structure at the 3′ end of *E. gracilis* operons has been correlated with mRNA processing events, as well as mRNA stability. While there is little or no primary sequence conservation in the spacer regions between the *psbK* operon genes, a large stable hairpin structure occurs between *psaM* and *trnR* in each species (not shown) . This structure may act as a 3′ mRNA stability element after the rapid cleavage of *trnR* from the tetracistronic transcript. A similar structure is present at the 3′ end of *E. gracilis psbA* (Stevenson and Hallick 1994).

Discussion

The *psbK* promoter directs transcription of a novel mixed tetracistronic operon

We have identified a new polycistronic mixed operon in *E. gracilis*. The genes *psbK*, *ycf12, psaM*, and *trnR* are co-transcribed from a single promoter upstream of *psbK*. Mature products are fully spliced, monocistronic RNAs derived from endonucleolytic intergenic cleavage and processing of two group III introns and a group III twintron. The *trnR* transcript is most likely to be cleaved from the primary transcript very rapidly, since a tetracistronic mRNA was detected by RT-PCR but not by Northern blotting. It is possible that tRNA genes which are co-transcribed in a mixed operon may be processed before the end of transcription (Stevenson and Hallick 1994).

In the Euglenoid lineage, the gene content and order in the *psbK* operon gene is conserved, but this order differs from that in plant and algal chloroplast genomes (Fig. 4). Unicellular organisms contain a diversity of genes surrounding *psbK*. The gene order most similar to Euglenoids is found in *Chlorella vulgaris*, where the *psbK-ycf12-psaM* order is conserved, but flanking tRNAs vary (Wakasugi et al. 1997). Conservation of gene order between *Chlorella* and *Euglena* is consistent with the hypothesis that the original endosymbiont that gave rise to the *Euglena* chloroplast was derived from the chlorophyte lineage. Because the *psbK* operon is so variable, the *psbK* region could be a recombinational "hot spot". Genomic recombination may be facilitated by interactions between tRNAs (Howe et al. 1988; Hiratsuka et al. 1989; Reiter et al. 1989). Therefore, it is interesting to note that Euglenoid and higher-plant *psbK* operons are flanked by the genes for *trnR*/*trnT* and *trnQ*/*trnS*, respectively. Since chloroplast genes are generally transcribed in clusters, any novel genes placed downstream of the *psbK* gene could lead to their adoption by the *psbK* promoter. A high degree of recombination could possibly account for the mixed content of the Euglenoid *psbK* operon.

Comparative phylogenetic analysis of homologous group III introns

The insertion sites of *psbK* intron 1, the external intron of the *psbK* twintron, and *ycf12* intron 1 are conserved

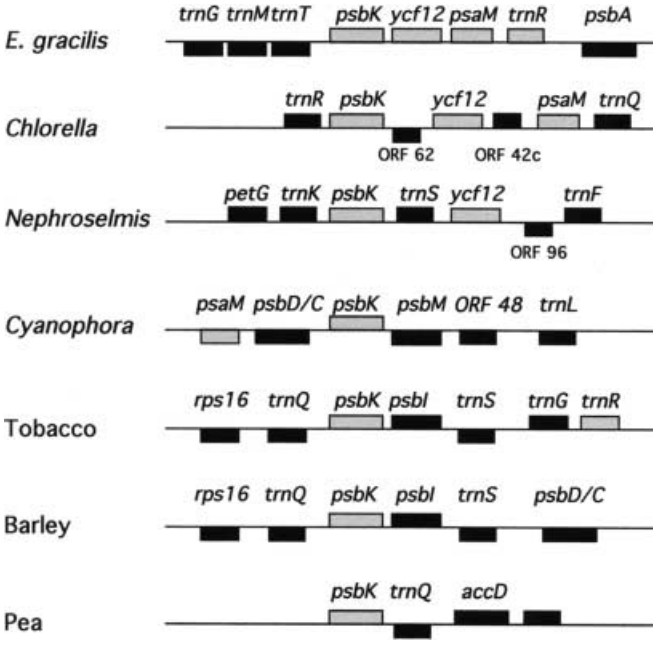

Fig. 4 Comparison of the organization of the *psbK* regions from selected organisms. Genes found in the *E. gracilis psbK* operon are shown as *grey boxes*. The *black boxes* represent flanking genes. The direction of transcription is indicated by the location of the gene *above or below* the connecting line

throughout the Euglenoid lineage. Introns located at a common insertion site in multiple species are most likely to be derived from a common ancestor. The first intron of *P. accuminata* is a possible exception because the insertion site is displaced by one codon. This could be a case of intron slippage or it could be a separate intron addition event. Without additional data on the evolutionary relationship between *P. accuminata* and the remainder of the Euglenoid species, the question whether the first intron is derived from a single insertion event cannot be answered definitively. In contrast, the group II introns of this operon, *E. stellata* and *E. sanguinea ycf12* intron 2, are probably not homologous.

As was observed for the group III twintron *psbC* intron 4, a group II intron-like domain VI was identified by comparative phylogenetic analysis of the secondary structures of *psbK* intron 2a and *ycf12* intron 1. Several lines of evidence now point to a common ancestor for group II and group III introns, from which a functional domain VI structure could have been acquired (Doetsch et al. 1998). The relationship of the putative group III domain ID3-like structure to its group II intron analog, however, is not as clear. While most of the introns could be folded into a single stem-loop structure in the domain I region, these structures were not similar enough in primary sequence or secondary motifs to consider them phylogenetically conserved. The domain I-type region may simply exist to bring together the ends of the intron so that splicing can be initiated. Alternatively, this region could contain as yet unidentified binding sites for *cis*-acting splicing factors which would not necessarily be the same for each intron.

Euglenoid and group III intron evolution

The *psbK* operon contains a higher diversity of intron content than any region of the Euglenoid genome studied to date. When *psbK* operon intron content is compared to the available *rbcL* phylogeny, some observations concerning group III intron evolution can be made (Fig. 5) (Thompson et al. 1995). *E. stellata*, a fairly deep branching species, contains the whole suite of possible *psbK* operon introns. *E. myxocylindracea*, *E. viridis*, and *E. anabaena* are more derived than *E. stellata*, but contain fewer introns. Because the *psbK* operon introns are homologous, the most parsimonious explanation is to assume at least two independent intron losses. Apparently the internal intron was lost from *E. myxocylindracea*, and the *E. viridis*-*E. anabaena* clade has lost *psbK* intron 1. Both of these introns are located in regions subject to low coding sequence restraints (an external intron and the *psbK* pre-sequence). Interestingly, the introns located in functionally constrained regions – the *psbK* external intron and *ycf12* intron 1 – are present in every species. Whether the loss of either *psbK* intron 1 or *psbK* intron 2b occurred only once or multiple times is impossible to judge from this phylogenetic tree.

Several conclusions can be drawn from this comparison of *psbK* intron content and phylogeny. The first is that group III introns are deeply rooted and were apparently present in the common ancestor of all surveyed extant Euglenoid species. In agreement with the *psbC* intron 4 survey and unpublished data, group III introns continue to be identified in relatively large numbers from basally branching Euglenoid species. Since group III twintrons are observed in the deepest branching species, these introns may have been very mobile. High insertion rates may have been correlated with high excision rates, since several intron losses are now postulated.

The evolutionary history of group II introns may differ from that of group III introns. To date, a much smaller number of group II introns than group III introns have been found in basally branching species. A higher percentage of those introns are not homologous, but are secondary insertions, such as *ycf12* intron 2.

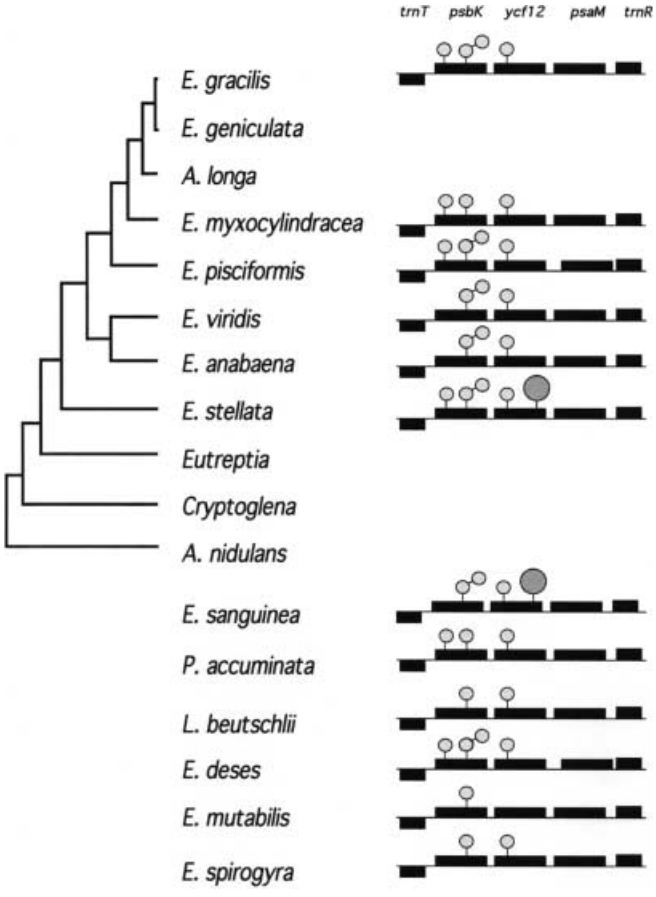

Fig. 5 The intron content of the *psbK* operon is depicted in relation to Euglenoid phylogeny. The tree shown is a consensus of the three most parsimonious trees determined by PAUP analysis of the nucleotide sequences of the *rbcL* gene (adapted from Thompson et al. 1995). The intron distribution in the *psbK* operons is indicated on the *right*. Group III introns are depicted as *small shaded lollipops* and group II introns as *large shaded lollipops*. The introns in *E. pisciformis* are predicted by PCR analysis, but have not been confirmed by sequence analysis. Species which were not used in the original phylogenetic analysis are shown *below* the tree to illustrate their intron contents only

Based on these observations, it now seems likely that group II introns are a later addition to the Euglenoid lineage than group III introns, and that group III introns pre-date their group II intron counterparts in the evolutionary history of the Euglenoid plastid genome.

Acknowledgements This work was supported by NIH grant GM35665.

References

Ausubel F, Brent R, Kingston R, Moore D, Seidman J, Smith J, Struhl K, Albright L, Coen D, Varki A, Janssen K (1995) Preparation and analysis of DNA: Electroblotting from a polyacrylamide gel to a nylon membrane. Current protocols in molecular biology. Wiley, New York, pp 2.2.9–2.9.15

Betts L, Spremulli L (1994) Analysis of the role of the Shine-Dalgarno sequence and mRNA secondary structure on the efficiency of translational initiation in the *Euglena gracilis* chloroplast *atpH* mRNA. J Biol Chem 269: 26456–26463

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

Copertino DW, Hallick RB (1993) Group II and group III introns of twintrons: potential relationships to nuclear pre-mRNA introns. Trends Biochem Sci 18: 467–471

Copertino DW, Van Hook FW, Hall ET, Jenkins KP, Hallick RB (1994) A group III twintron encoding a maturase-like gene excises through lariat intermediates. Nucleic Acids Res 22: 1029–1036

Doetsch N, Thompson M, Hallick R (1998) A maturase-encoding group III twintron is conserved in deeply rooted euglenoid species: are group III introns the chicken or the egg? Mol Biol Evol 15: 76–86

Hallick RB, Richards OC, Gray PW (1982) Isolation of intact, superhelical chloroplast DNA from *Euglena gracilis*. In: Edelman M (ed) Methods in chloroplast molecular biology. Elsevier, Amsterdam, pp 281–293

Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, Spielmann A, Stutz E (1993) Complete DNA sequence of *Euglena gracilis* chloroplast DNA. Nucleic Acids Res 21: 3537–3544

Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun CR, Meng B-Y, Li Y-Q, Kanno A, Nishizawa Y, Hirai A, Shinozaki K, Sugiura M (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 the cereals. Mol Gen Genet 217: 185–194

Holton TA, Graham MW (1991) A simple and efficient method for direct cloning of PCR products using ddT-tailed vectors. Nucleic Acids Res 19: 1156

Hong L, Stevenson JK, Roth WB, Hallick RB (1995) *Euglena gracilis* chloroplast *psbB*, *psbT*, *psbH* and *psbN* gene cluster: regulation of *psbB-psbT* pre-mRNA processing. Mol Gen Genet 247: 180–188

Howe CJ, Barker RF, Bowman CM, Dyer TA (1988) Common features of three inversions in wheat chloroplast DNA. Curr Genet 13: 343–349

Michel F, Ferat JL (1995) Structure and activities of group II introns. Annu Rev Biochem 64: 435–461

Reiter W-D, Palm P, Yeats S (1989) Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. Nucleic Acids Res 17: 1907–1914

Rott R, Drager RG, Stern DB, Schuster G (1996) The 3′ untranslated regions of chloroplast genes in *Chlamydomonas reinhardtii* do not serve as efficient transcriptional terminators. Mol Gen Genet 252: 676–683

Starr RC, Zeikus JA (1993) UTEX – The culture collection of algae at the University of Texas at Austin. J Phycol 29: 1–106

Steege D, Graves M, Spremulli LL (1982) *Euglena gracilis* chloroplast small subunit rRNA. J Biol Chem 257: 10430–10439

Stern DB, Gruissem W (1987) Control of plastid gene expression: 3′ inverted repeats act as mRNA processing and stabilizing elements, but do not terminate transcription. Cell 51: 1145–1157

Stern DB, Jones H, Gruissem W (1989) Function of plastid mRNA 3′ inverted repeats. RNA stabilization and gene-specific protein binding. J Biol Chem 264: 18742–18750

Stern D, Higgs D, Yang J (1997) Transcription and translation in chloroplasts. Trends Plant Sci 2: 308–315

Stevenson JK (1994) Transcription and intercistronic RNA processing of polycistronic operons of *Euglena gracilis* chloroplast. Ph.D. Dissertation. University of Arizona, Tucson, AZ

Stevenson JK, Hallick RB (1994) The *psaA* operon pre-mRNA of the *Euglena gracilis* chloroplast is processed into photosystem I and II mRNAs that accumulate differentially depending on the conditions of cell growth. Plant J 5: 247–260

Thompson MD, Copertino DW, Thompson E, Favreau MR, Hallick RB (1995) Evidence for the late origin of introns in chloroplast genes from an evolutionary analysis of the genus *Euglena*. Nucleic Acids Res 23: 4745–4752

Thompson M, Zhang L, Hong L, Hallick R (1997) Extensive structural conservation exists among several homologs of two *Euglena* chloroplast group II introns. Mol Gen Genet 257: 45–54

Wakasugi T, Nagai T, Kapoor M, Sugita M, Ito M, Ito S, Tsudzuki J, Nakashima K, Tsudzuki T, Suzuki Y, Hamada A, Ohta T, Inamura A, Yoshinaga K, Sugiura M (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 C, Roney W, Alston R, Spremulli L (1989) Initiation complex formation on *Euglena* chloroplast 30S subunits in the presence of natural mRNAs. Nucleic Acids Res 17: 9735–9747

Zhang L, Jenkins KP, Stutz E, Hallick RB (1995) The *Euglena gracilis* intron-encoded *mat2* locus is interrupted by three additional group II introns. RNA 1: 1079–1088