Euglena Transcript Processing

8

David C. McWatters and Anthony G. Russell

Abstract

RNA transcript processing is an important stage in the gene expression pathway of all organisms and is subject to various mechanisms of control that influence the final levels of gene products. RNA processing involves events such as nuclease-mediated cleavage, removal of intervening sequences referred to as introns and modifications to RNA structure (nucleoside modification and editing). In *Euglena*, RNA transcript processing was initially examined in chloroplasts because of historical interest in the secondary endosymbiotic origin of this organelle in this organism. More recent efforts to examine mitochondrial genome structure and RNA maturation have been stimulated by the discovery of unusual processing pathways in other Euglenozoans such as kinetoplastids and diplonemids. Eukaryotes containing large genomes are now known to typically contain large collections of introns and regulatory RNAs involved in RNA processing events, and *Euglena gracilis* in particular has a relatively large genome for a protist. Studies examining the structure of nuclear genes and the mechanisms involved in nuclear RNA processing have revealed that indeed *Euglena* contains large numbers of introns in the limited set of genes so far examined and also possesses large numbers of specific classes of regulatory and processing RNAs, such as small nucleolar RNAs (snoRNAs). Most interestingly, these studies have also revealed that *Euglena* possesses novel processing pathways generating highly fragmented cytosolic ribosomal RNAs and subunits and non-conventional intron classes removed by unknown splicing mechanisms. This unexpected

D.C. McWatters • A.G. Russell, Ph.D. (✉)
Department of Biological Sciences, University of Lethbridge, 4401 University Dr W, Lethbridge, AB, Canada, T1K 6T5

Alberta RNA Research and Training Institute, University of Lethbridge, Lethbridge, AB, Canada
e-mail: david.mcwatters@uleth.ca; tony.russell@uleth.ca

S.D. Schwartzbach, S. Shigeoka (eds.), *Euglena: Biochemistry, Cell and Molecular Biology*, Advances in Experimental Medicine and Biology 979, DOI 10.1007/978-3-319-54910-1_8

diversity in RNA processing pathways emphasizes the importance of identifying the components involved in these processing mechanisms and their evolutionary emergence in *Euglena* species.

Keywords
Euglena • RNA processing • Transcript • Small nucleolar RNA • Mitochondrial RNA • Chloroplast RNA • Intron • Gene expression • RNA modification • Spliced leader RNA

Abbreviations

gRNA	Guide RNA
IGS	Intergenic spacer
indel	Insertion/deletion
ITS	Internal transcribed spacer
kbp	Kilo base-pairs
LSU	Large subunit
MITE	Miniature Inverted Repeat Transposable Element
mRNA	Messenger RNA
mtDNA	Mitochondrial DNA
Nm	2′-O-methylation of the ribose sugar
nt	Nucleotide
ORF	Open-reading frame
PCR	Polymerase chain reaction
PRORP	Protein-only ribonuclease P
rDNA	Ribosomal DNA
rprotein	Ribosomal protein
rRNA	Ribosomal RNA
SL RNA	Spliced leader RNA
snoRNA	Small nucleolar RNA
snRNA	Small nuclear RNA
SSU	Small subunit
tRNA	Transfer RNA
Ψ	Pseudouridine modification of RNA

8.1 Nuclear-Encoded Introns

Many nuclear-encoded protein-coding genes in *Euglena* contain introns which possess variable properties resulting in their classification into at least two distinct categories: conventional spliceosomal introns that are predicted to be removed from precursor mRNAs by the characterized *Euglena* spliceosome components and so-called "non-conventional" (non-canonical) introns that are excised by unknown cellular components. From the limited set of *Euglena* genes whose sequences have been determined and compared to their expressed mature mRNA sequences, it appears that having multiple introns and possessing both intron types in an individual gene is relatively common.

The non-conventional introns are defined as containing extensive secondary structural potential via base-pairing of intron 5′ and 3′ end proximal sequences, but little overall intron sequence conservation (Tessier et al. 1991; Canaday et al. 2001; Russell et al. 2005; Milanowski et al. 2014, 2016; Muchhal and Schwartzbach 1992, 1994) (Fig. 8.1b). They also frequently contain direct repeat sequences, of variable length, at the intron termini creating uncertainty in the accurate prediction of splice donor and acceptor sites for some of these introns. The lack of strict conservation of the direct repeats and their sequence variability indicates that they are unlikely to have a role in the splicing mechanism but may instead be remnants of intron sequence insertion and mobility events. Milanowski et al. have noted that these features are reminiscent of MITE-like transposon elements (Milanowski et al. 2014); therefore, if the non-conventional introns have been derived from such elements then perhaps *trans*-acting factors that associate with them may have been co-opted to be involved in the splicing mechanism. While there is no apparent conservation of extended sequence elements in these introns, intron sequence comparisons have revealed a preference for intron 5′ end proximal

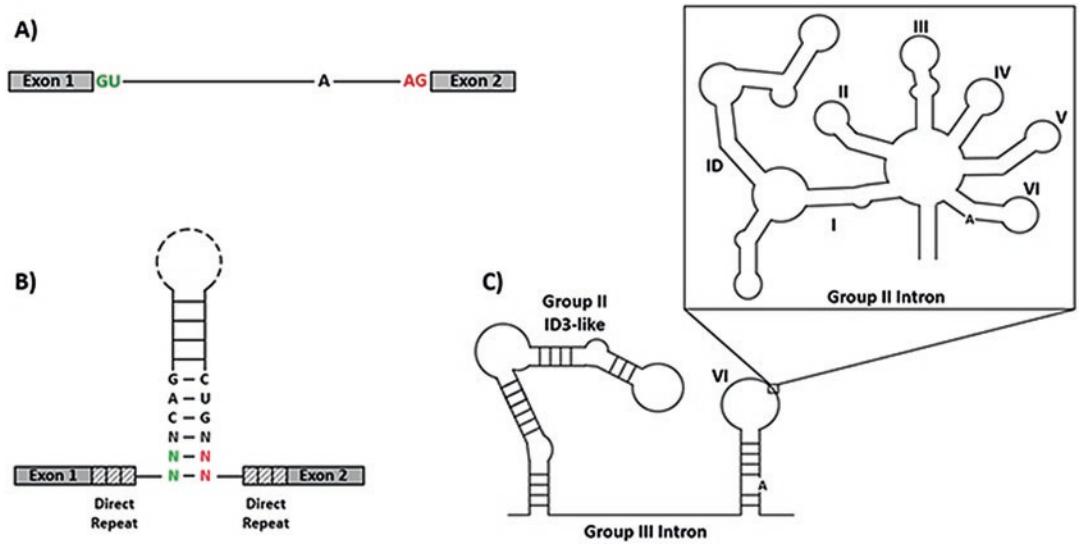

Fig. 8.1 Structural features of the different *E. gracilis* intron classes. (**a**) Canonical structure for a U2-type spliceosomal intron with intronic 5′ splice site boundary nucleotides in green and 3′ nucleotides in red, and the branch point A in black. (**b**) Secondary structure of an *E. gracilis* non-conventional intron. For those introns lacking direct repeats, nucleotides that sometimes show adherence to 5′ and 3′ splice site boundaries of conventional introns are in green and red respectively. Conserved +4 to +6 nucleotides and base paired nucleotides are shown in black. There are variable numbers of nucleotides in the base-paired stems and the dashed line represents the variable length intronic region. (**c**) General secondary structure for a chloroplast twintron arrangement, in this case composed of a group II intron inserted within a group III intron. Conserved structural domains for both introns are labelled. Position of insertion of the group II intron in domain VI of the group III intron is indicated by the small box. Relative positions of branch point A nucleotides for both introns are indicated

nucleotide positions +4 to +6 to be 'CAG' and the complementary sequence (CTG) starting 6 nucleotides from the intron 3′ end (Milanowski et al. 2014, 2016) (Fig. 8.1b). The conservation of these short sequences and their ability to base pair may be required for the splicing mechanism and accurate determination of splice site boundaries.

Some of the *Euglena* non-conventional introns contain intron terminal nucleotides (5′GT or AG3′) identical to those of most conventional spliceosomal introns (Canaday et al. 2001; Milanowski et al. 2014) (Fig. 8.1a). Additionally, some of the predicted spliceosomal introns show extended base-pairing potential in intron locations similar to regions of secondary structure observed in the non-conventional introns. Such observations have raised questions about whether interconversion between intron classes may occur during intron sequence evolution in *Euglena* and whether introns demonstrating mixed features of both classes should be classified as a distinct type called "intermediate" introns (Canaday et al. 2001; Russell et al. 2005). These could potentially be excised using components of both the spliceosome and *trans*-acting non-conventional intron splicing factors.

Milanowski et al. have recently examined the conservation of intron position and class in conserved nuclear genes in different *Euglena* species to shed light on such questions (Milanowski et al. 2014, 2016). These studies have further refined the limited conserved sequence and structural features of non-conventional introns (as described above) and revealed that non-conventional intron gain/loss appears to occur much more frequently than observed for euglenid spliceosomal introns. There is also much greater intron length variation in different species at conserved non-conventional intron positions than is the case for the conserved spliceosomal introns. A preference for a 5′ purine

nucleotide in non-conventional introns has also been observed (Milanowski et al. 2016) that perhaps affects splicing efficiency, thus explaining the frequent observation of non-conventional introns starting with the sequence 5′GT/C (i.e. spliceosomal-like) but also containing all other typical features of non-canonical introns. Such introns had previously been categorized as intermediate type; however, many of these introns have only poor base-pairing potential to the characterized *Euglena* spliceosomal U1 snRNA sequence (Breckenridge et al. 1999) making it unclear whether these introns are in fact in a transition state between intron classes and utilize any spliceosome components.

Identification of many instances of U12-type (minor-type) spliceosomal introns residing in identical gene positions to U2-type (major-type) introns in distantly-related species has provided evidence of evolutionary conversion between spliceosomal intron classes (Burge et al. 1998; Basu et al. 2008). To date, no instance of a conserved intron position being a conventional spliceosomal intron in one *Euglena* species and non-conventional in another species has been identified. Milanowski et al. did however recently discover the first case of a non-conventional intron containing 5′GC and AC3′ intron terminal sequences, the best candidate so far for an intermediate intron since both splice sites match those of conventional spliceosomal introns (Milanowski et al. 2016). They also identified a recently acquired non-conventional intron in the *gapC* gene in *Euglena agilis* that contains significantly longer extended intron boundary direct repeats than had previously been observed, leading them to propose that DNA double-strand break repair processes may be involved in intron emergence/acquisition in *Euglena*.

Only a very limited set of genes and small number of introns have been characterized in detail in *Euglena*. Recent extensive mRNA transcriptome studies under different physiological stress conditions (O'Neill et al. 2015; Yoshida et al. 2016; Ferreira et al. 2007) and future determination of more complete genome sequences from different *Euglena* species should permit a much more extensive analysis of intron evolution in euglenids and the detection of intron class conversion, if it occurs. Also important will be the identification of the cellular factors required for the removal of non-conventional introns, the experimental determination of critical intron structure and sequence requirements for splicing reactions, and the further identification of conventional spliceosomal components in *Euglena.* snRNAs have been identified but no experimental analysis of spliceosomal proteins or snRNP complexes has yet been performed.

8.2 Nuclear-Encoded Cytosolic rRNA

Expression and maturation of cytoplasmic ribosomal RNA in *Euglena gracilis* differs dramatically from what occurs in almost all other examined eukaryotes. The most striking feature is the cytoplasmic large subunit (LSU) rRNA, which in its mature form is fragmented into 14 discrete pieces, including the 5.8S rRNA (also called LSU1) (Schnare and Gray 1990). All 14 LSU fragment species along with the encoding sequence for the intact mature 19S rRNA of the small subunit (SSU) are encoded on an 11,056 base pair extrachromosomal DNA circle that is transcribed as a contiguous large RNA (read-around transcription) by RNA polymerase I (Greenwood et al. 2001; Schnare et al. 1990) . These rDNA circles number between 800 and 4000 copies per cell (Cook and Roxby 1985; Revel-Chapuis et al. 1985; Greenwood et al. 2001) and possess a single origin of DNA replication (Ravel-Chapuis 1988). The 19S, 5.8S (LSU1), and other 13 LSU rRNA fragments are separated by internal transcribed spacer (ITS) sequences ranging in size from 10 to 1188 base pairs in length, while LSU14 and the 19S SSU rRNA sequence are separated by an intergenic spacer (IGS) of 1743 base pairs (Greenwood et al. 2001). The spacer regions are removed post-transcriptionally producing a number of processing intermediates (Schnare et al. 1990; Greenwood and Gray 1998). Despite detection of these intermediate processing steps, very little is known about the mechanisms and components

responsible for processing and maturation of the initial single transcript into each final rRNA species. Even the nearly universally conserved rRNA processing RNase MRP complex still remains uncharacterized (or detected) in *Euglena* (López et al. 2009). Ribosome assembly in *E. gracilis* is almost certainly highly complex and likely requires a number of novel processing components. A better understanding of *E. gracilis* ribosome assembly may shed light on how evolutionary processes have shaped the development of such a fragmented ribosome structure and perhaps even reveal insights about steps in more canonical eukaryotic ribosome assembly pathways. The only RNA species of the cytoplasmic ribosome not found on the rDNA circle is the 5S rRNA (Schnare et al. 1990), which is instead typically genomically-encoded within 600 base pair long tandem repeats with spliced-leader (SL) RNAs, at an estimated copy number of 300 repeated units per haploid genome (Keller et al. 1992). Evidence also suggests single copy 5S and SL genes are present, however these appear to be less conserved.

8.3 *Euglena* snoRNAs and Their Expression

The *E. gracilis* rRNA has the largest number of modified nucleotide positions of any rRNA examined to date. The SSU and LSU rRNA subunits contain 88 and 262 identified modifications respectively (Schnare and Gray 2011). Therefore, there is a significant increase in the density of modifications in the fragmented large subunit (LSU) rRNA species in *E. gracilis* relative to the non-fragmented SSU rRNA suggesting that the additional modifications may have an important structural stabilizing role and/or function in the more complicated ribosome biogenesis pathway in this organism. The majority of these modifications are 2′-O-methylations (Nm) (209) and pseudouridines (Ψ) (119) contradicting the usual trend of multicellular organismal rRNA being more heavily modified than that of simpler organisms. In addition to having conserved modifications at many positions also modified in other eukaryotes, *E. gracilis* also appears to contain a large number of species-specific and euglenozoan-specific modifications (Schnare and Gray 2011; Eliaz et al. 2015).

In eukaryotes, the two most prevalent modifications in rRNA are isomerization of uridine to Ψ and 2′-O-methylation (Li et al. 2016; Sharma and Lafontaine 2015). Most of these modifications are targeted by small guide RNAs called small nucleolar (sno) RNAs. SnoRNAs targeting Nm sites are called C/D box snoRNAs while those that target sites of Ψ formation are called H/ACA box snoRNAs, with both classes defined by conserved sequence and structural features (Bratkovič and Rogelj 2014; Lui and Lowe 2013). Since *E. gracilis* has so many modifications, the initial prediction was that it would also require a large collection of snoRNAs to specify all these modified sites. Identification of *E. gracilis* snoRNAs through biochemical, genomic amplification (PCR) strategies and bioinformatic analysis has revealed that this is indeed the case (Moore and Russell 2012; Russell et al. 2004, 2006). Not only are there a large number of different snoRNA species but also a very large collection of sequence-related isoforms of each species, the full extent of which has yet to be determined.

Elucidation of the organization of snoRNA genes in *E. gracilis* has revealed that these genes are usually tandemly repeated in the genome with genes for the two classes of snoRNAs interspersed (Moore and Russell 2012). This organization pattern is similar to what has been observed in several trypanosome species and various plant species (Barneche et al. 2001; Brown et al. 2003; Liang et al. 2005). The modified sites in *E. gracilis* rRNA are not evenly dispersed along the lengths of the rRNAs, but rather typically clustered and sometimes densely clustered, such as a region in LSU species 6 where in a stretch of 22 nucleotides nearly half are Nm (2′-O-methylated) (Schnare and Gray 2011). This modification pattern is related to the organization of snoRNA genes. We have identified several instances where adjacent or nearby genes encode snoRNA species that target adjacent rRNA modification sites (Moore and Russell 2012). How did such a situation

arise? Many *Euglena* snoRNAs are encoded by tandemly repeated genes and when sequence divergence occurs in a paralogous gene copy that alters the guide region of a snoRNA, new base-pairing potential emerges to target a new modification site; that is, a new snoRNA species has been created. We have documented several cases where small insertion/deletions have occurred in nearby snoRNA gene copies that allows targeting of adjacent rRNA modification sites (Moore and Russell 2012). It seems that the apparent sequence repetitiveness in the *E. gracilis* genome, and the unexplained propensity to create gene copies, has been a driving factor in the creation of the large collection of snoRNA species and modification sites in this organism. However, what is not so clear is why this is selectively affecting modification of the various LSU rRNA species more than the SSU rRNA. Perhaps *E. gracilis* rapidly gains and then loses new snoRNA species through this genomic amplification mechanism but there is stronger selective pressure to retain snoRNAs targeting LSU fragment species as this is more beneficial for ribosome function in this organism. Also intriguing to consider is whether initially the fragmented nature of the *E. gracilis* rRNA necessitated a mechanism to rapidly create new snoRNA isoforms (snoRNAs targeting the same site) and species (those targeting different sites) or vice versa; fragmentation emerged as it could be tolerated in a cellular environment containing an unusually large number of snoRNAs with largely redundant functions.

Most of the *E. gracilis* snoRNA genes are expressed initially as polycistronic precursor transcripts of unknown lengths (we have detected transcripts upwards of 800 nts), containing several individual snoRNA sequences that are then processed into individual snoRNA species (Moore and Russell 2012). They are assembled with conserved core protein binding partners by an undefined processing and assembly mechanism in *Euglena*. Polycistronic transcripts containing both snoRNA classes have been detected. Transcription initiation and termination elements for expression of these genomic snoRNA clusters have yet to be determined; however, some of the spacer regions between mature snoRNA sequences display significant structural potential that may play a role in the expression mechanism (Moore and Russell, unpublished results). Not all *E. gracilis* snoRNAs are expressed polycistronically as the U3 snoRNA, a snoRNA that functions in pre-rRNA processing steps (i.e. specifying rRNA cleavage sites instead of targeting modification sites) appears to be expressed monocistronically (Greenwood et al. 1996; Charette and Gray 2009). Although the U3 snoRNA genes are multi-copy and frequently found associated with either U5 snRNA or tRNA genes, the U3 genes are in the opposite transcriptional orientation to the nearby U5 or tRNA genes (Charette and Gray 2009). Unlike U3, two other predicted *E. gracilis* processing snoRNAs, U14 and the Eg-h1 H/ACA-like RNA, are instead encoded by closely-spaced tandemly repeated genes like the modification-guide snoRNAs and are likely polycistronically expressed (Moore and Russell 2012). Therefore, there is no simple relationship between snoRNA function and expression mode in *E. gracilis*.

Currently, it is not definitively known which RNA polymerases are being used to express different snoRNA species in *E. gracilis.* In trypanosomatids and plants, U3 snoRNA genes are transcribed by RNA polymerase III (Fantoni et al. 1994; Kiss et al. 1991; Marshallsay et al. 1992), and the close linkage of some *E. gracilis* U3 genes with tRNAs suggests that at least these gene copies may be transcribed by this RNA polymerase. However, in trimethylguanosine cap pull-down RNA libraries we have found an abundance of *E. gracilis* U3 sequences consistent with these U3 species being transcribed by RNA polymerase II (Moore and Russell, unpublished results). Since not all *E. gracilis* U3 genes are linked with tRNA genes, it is possible that both RNA polymerases may be involved in U3 snoRNA expression depending on genomic context of individual U3 genes. The frequent expression of *E. gracilis* modification guide snoRNAs as polycistronic transcripts and relative transcript size is more consistent with RNA polymerase II transcriptional properties.

8.4 *Euglena* Chloroplast RNAs and Processing

Most recently, much of what has been deduced about *Euglena* choloroplast genome RNA-coding capacity has been through the determination of complete chloroplast genome structures from a collection of representative species from the Euglenaceae (Hrdá et al. 2012; Wiegert et al. 2012; Dabbagh and Preisfeld 2017; Bennett and Triemer 2015) and comparison to the much earlier determined choloroplast genome structure of *Euglena gracilis* Strain Z (Hallick et al. 1993). An examination of transcription patterns of the 96 genes contained on the *E. gracilis* plastid genome under different physiological states and stress conditions has also been performed (Geimer et al. 2009). Chloroplast RNA processing information has been derived primarily from Richard Hallick's group. They identified and then examined splicing patterns of a large collection of chloroplast introns and investigated expression modes for rRNA and tRNA, and the chloroplast RNA polymerase activities required for their expression. Identification of any other chloroplast non-coding RNAs, and protein or ribonucleoprotein complexes involved in chloroplast RNA maturation will require future biochemical studies and other types of analyses.

8.4.1 Chloroplast rRNA and tRNA

In the two examined strains of *E. gracilis*, chloroplast rRNA is encoded in operons approximately 6000 nt in length. The operon codes for 16S, 23S, and 5S rRNA genes separated by internal transcribed spacers some of which contain tRNA genes or pseudogenes, an overall arrangement similar to many bacterial rRNA operons. The operon structure is tandemly repeated three times, with a fourth partial repeat containing only a complete 16S rRNA sequence and additional open reading frame (ORF) found in Strain Z but was not confirmed in var. bacillaris (Hallick et al. 1993; Bennett and Triemer 2015). These operons make up 13.7% of the length of the genome.

There are a total of 27 tRNAs (not including the pseudogenes) found in Strain Z which are actively expressed (Hallick et al. 1993). An additional 9 pseudo-tRNAs which do not appear to be transcribed are found in regions within the rRNA operon repeats. The bacillaris strain possesses 31 actively transcribed tRNA genes, with only 4 pseudogenes (Bennett and Triemer 2015). trnI-trnA genes are co-transcribed with the rRNA operons and are the only chloroplast tRNAs that are multicopy. Most of the tRNA genes reside in clusters with short spacers, sometimes closely-linked with protein-coding genes.

There are at least two different RNA polymerase activities in *E. gracilis* chloroplasts that can be biochemically separated and are active when used in *in vitro* transcription assays (Greenberg et al. 1984). They display differences in enzymatic properties including salt concentration tolerance, optimum Mg^{2+} concentrations and temperature activity profiles. The RNA polymerase activity that remains tightly associated with chloroplast genomic DNA has been shown to selectively transcribe the rRNA operons (Greenberg et al. 1984). The soluble RNA polymerase activity transcribes most of the chloroplast tRNAs excluding those that are contained within the rRNA operons. Specificity of these RNA polymerase activities for transcribing the various protein-coding genes has not been extensively examined.

Polycistronic transcription and subsequent processing of these extended transcripts appears to be a prevalent mode of gene expression in *E. gracilis* chloroplasts for transcripts produced by either RNA polymerase activity (Christopher and Hallick 1990; Greenberg and Hallick 1986). Greenberg and Hallick were first able to isolate *E. gracilis* soluble chloroplast extracts that were capable of transcribing polycistronic transcripts containing multiple tRNA species that also accurately processed these primary transcripts to generate mature tRNA 5′ and 3′ termini (accurate CCA 3′ end addition was not verified in this study) (Greenberg and Hallick 1986). Either chloroplast DNA or cloned tRNA genes served as appropriate transcription and subsequent processing substrates

for the soluble extracts. Christopher and Hallick then demonstrated that polycistronic transcription also occurs for chloroplast ribosomal protein genes where one transcription unit was characterized that contains 11 rprotein genes, an isoleucine tRNA gene, and an ORF of unknown function (Christopher and Hallick 1990). This transcription unit is also predicted to contain at least 15 introns making it a large polycistronic transcription unit and complex gene expression pathway. It appears that the tRNA is processed and matured from this large transcript, as opposed to alternative individual transcription of the tRNA as a nested transcription unit, since the spacers flanking the mature sequence are short and do not appear to contain obvious promoter or termination elements. The authors noticed that the codon that would be deciphered by this particular isoleucine tRNA isoacceptor is enriched in mRNAs coding for constitutively expressed proteins (such as ribosomal proteins) relative to the codon's frequency in mRNAs for light-induced proteins. They speculate this may be the reason for this tRNA residing in this particular polycistronic unit. Through detection of RNA processing intermediates and products via nucleic acid hybridization experiments, it appears that RNA endonucleases are utilized for liberating individual RNA species from the polycistronic transcript and also for other transcription units containing tRNA species. A prediction would be the key involvement of the tRNA 5′ end maturation endonuclease RNase P in the various polycistronic transcript processing pathways.

8.4.2 Chloroplast Introns

An unusual feature of the *Euglena* chloroplast genome structure is the very large number of introns. Surveys of *Euglenaceae* chloroplast genome sequences have revealed a high degree of variability in intron content (Bennett and Triemer 2015; Pombert et al. 2012). The two sequenced *E. gracilis* chloroplast genomes possess the greatest number of introns in this taxa with the strain Z chloroplast containing 155 introns and var. bacillaris containing 134. This results in 66.7 and 68.3% of protein coding genes containing at least 1 intron in the two strains, respectively (Thompson et al. 1995; Bennett and Triemer 2015; Hallick et al. 1993). Curiously, despite this high intron content, none of the *Euglena* chloroplast tRNA genes contain introns. This differs markedly from what is found in green algae where over 50% of tRNA genes contain introns.

Chloroplast introns in *E. gracilis* include members of both group II (self-splicing) introns and a unique related class designated group III introns (Copertino and Hallick 1993). The *E. gracilis* group II introns contain most of the conserved features of this class of introns including structural domains I-VI (Fig. 8.1c), EBS-IBS pairings, and predicted ε-ε' and γ-γ'interactions (Copertino and Hallick 1993). These introns are however A-U rich (striking scarcity of G-C base-pairs in some cases) and show some structural "looseness" and variability relative to those introns found in more distantly related organisms. The group III introns appear to be degenerate or minimalized group II introns that contain only domain VI (predicted catalytic and branch point 'A' containing) and domain I; although even this later domain can be very minimalized in some predicted group III intron structures (Fig. 8.1c). Since *in vitro* splicing assays have not been performed with any of these *Euglena* introns, it is not known which of them are in fact self-splicing. It seems probable that the group III introns (at least) may have degenerated to the point where they are now completely dependent on *trans-acting* protein and/or RNA splicing factors for either or both of the two transesterification reactions, assuming they use such a splicing pathway.

Euglena chloroplast group II and group III introns can be found individually or as so-called twintrons: introns interrupting introns (Hallick et al. 1993; Bennett and Triemer 2015). Twintrons have been identified containing pairs of group II or group III introns, group II interrupting group III (and vice-versa), and even arrangements containing larger numbers of nested introns than just two. Hong and Hallick (1994) identified a case of a twintron arrangement in the *E. gracilis ycf8* gene where the outer intron can be a group II intron interrupted by two spaced group II introns;

that is, two introns each inserted at different locations within the outer intron or alternatively this outer intron can be classified as a group III intron interrupted by a group II intron. Alternative splicing dictates which combination of introns are removed and if the group II + III intron combination is removed, this pathway prevents removal of the outer group II intron by truncating several key structural regions.

The strict definition of a twintron, as defined for example by Hafez and Hausner (2015), is an embedded arrangement where the inner intron must be removed first to allow formation of the correct structure that catalyzes removal of the outer intron. In many of the *Euglena* twintron arrangements the insertion site of the inner intron is in domain V or VI of a group II intron, insertion positions that would be predicted to disrupt the tertiary structure required for outer intron removal in other well-studied group II introns. However, these *Euglena* group II introns already show some structural differences and flexibility relative to those studied in other organisms and together with the existence of the structurally minimalized group III introns, it may be premature to assume strict adherence to an ordered splicing pathway for all *Euglena* twintron arrangements. The frequency of twintrons in *Euglena* chloroplast genomes and the overall large number of introns suggests that intron mobility and insertion into new genomic sites is a relatively common occurrence in *E. gracilis* and more prevalent than is seen in other euglenids (Thompson et al. 1997; Pombert et al. 2012)—many of these introns appear to be unique to *E. gracilis.* Through recent determination of chloroplast intron structure and location in *Monomorphina aenigmatica*, a species occupying an intermediate branching position in euglenids, Pombert et al. (2012) have provided further evidence that group II/III intron abundance in *Euglena gracilis* appears to have resulted from more "recent" proliferation events, including the establishment of twintron arrangements (Hrdá et al. 2012; Wiegert et al. 2012). They found cases of intermediate stages of intron evolution in which *M. aenigmatica* contains a single group II intron (i.e. no twintron arrangement) inserted at the same gene position as the outer intron of a twintron arrangement in *Euglena gracilis*. The maintenance of twintron arrangements is the strongest argument so far for ordered splicing pathways; that is, insertion into a site that disrupts splicing of the outer intron requires first removing the inner intron to prevent gene function inactivation that would otherwise be the result of the insertion event.

It is curious that both the *E. gracilis* nuclear and chloroplast genomes are so intron-rich and also contain intron classes not known to exist outside of euglenids. We may then speculate about whether there is an evolutionary relationship between the non-conventional nuclear introns and the chloroplast group III introns, both of which maintain few conserved intron structural features for their respective splicing mechanisms. Were the non-canonical introns the end result of a large scale invasion event of the nuclear genome by group III mobility elements derived from an ancestral euglenid chloroplast? A detailed understanding of the splicing mechanisms and components involved for removal of these different intron types, and a large- scale analysis of introns in *E. gracilis* and other euglenids may reveal new insights into intron evolution in eukaryotes and the importance of these various intron classes in regulating gene expression in these organisms.

Perhaps the most surprising feature of gene expression in *E. gracilis* chloroplasts is the fact that there appears to be little differential variation in RNA species level when cells are examined at different stages of development and/or subject to various stress-inducing agents (Geimer et al. 2009) This is somewhat unexpected considering the complexity of processing required to remove the large number of introns in precursor chloroplast transcripts and the unusual adaptability of this organism in general to adjust to a wide range of environmental fluctuations. It was observed however that there can be significant changes to global chloroplast RNA levels under these various tested conditions. Such observations may indicate that if differential changes are occurring at the proteome level in *E. gracilis* chloroplasts, the regulation may be occurring at the translational control level.

8.5 Mitochondrial Genome Structure, Expression, and RNA Editing

RNA processing in Euglenozoan mitochondria has been shown to be both mechanistically unique and amazingly diverse compared to other eukaryotic phyla. The three major groups within Euglenozoa: euglenids, kinetoplastids, and diplonemids show a broad range in mitochondrial chromosome structure, gene expression strategy, and RNA processing mechanisms. Comparatively little is currently known about euglenid mitochondria; in particular, until recently virtually nothing was reported about *E. gracilis* mtDNA structure. It now appears that there are significant differences in *E. gracilis* compared to mitochondria in the other Euglenozoan taxa. An understanding of these other Euglenozoans may then provide evolutionary insight into mitochondrial features in this phylum. Further analysis of mitochondrial DNA and RNA features in *E. gracilis* itself and other euglenids will be indispensable in understanding RNA maturation and genome structure in these species. Here, we put current knowledge of *E. gracilis* mitochondrial chromosome structure, RNA expression and processing, in the broader context of Euglenozoans collectively.

Diplonemid mitochondrial DNA is arranged into two classes of small circular chromosomes of different sizes, Class A (6 kbp) and Class B (7 kbp) (Marande et al. 2005). mRNAs in diplonemid mitochondria are not expressed as single contiguous transcripts but rather as short fragments (known as modules) of several hundred nucleotides (Kiethega et al. 2011; Vlcek et al. 2010; Marande and Burger 2007). Each module is encoded by a different chromosome that carries only that gene. Following expression the module transcripts require processing through endonucleolytic cleavage, polyadenylation of the 3′ module, and *trans*-splicing in order to form mature full length transcripts (Kiethega et al. 2013). The mechanism through which this *trans*-splicing occurs is not yet understood, though it has been proposed that small guide RNAs may help in facilitating this process (Kiethega et al. 2013; Moreira et al. 2016). Additional editing of modules may also occur, including addition of short uridine stretches (1–3 nucleotides) to module ends, as well as both C-to-U and A-to-I editing (Moreira et al. 2016). In the second major Euglenozoan group, the kinetoplastids, mitochondrial DNA (termed kinetoplast or kDNA) is also arranged into two classes of circular chromosomes. In contrast to diplonemids, kinetoplast chromosomes differ quite significantly in size and are classified as either large (maxicircles) or small (minicircles) (Riou and Delain 1969; Kleisen et al. 1976; Steinert and Van Assell 1975). Maxicircle copy number varies between species, from 25 to 50 copies per cell in examined species, while thousands of minicircles can be present. Kinetoplastid maxicircle chromosomes primarily carry the mitochondrial protein-coding and rRNA genes (Eperon et al. 1983; Westenberger et al. 2006; Simpson et al. 1987). Minicircles code for small guide RNAs (gRNA) (Pollard et al. 1990; Corell et al. 1993; Jasmer and Stuart 1986a, b; Deschamps et al. 2011) which form ribonucleoprotein complexes called editosomes that act in a unique form of uridine insertion/deletion (U indel) editing of mRNA. This form of U indel editing has made gene identification difficult as the gene sequence may have little resemblance to the mature edited mRNA, and up to 553 insertion and 89 deletion sites have been characterized for a single transcript (Koslowsky et al. 1990).

The *Euglena gracilis* mitochondrial genome is also atypical but appears to be quite different from those of other Euglenozoans. Rather than circular chromosomes as seen in the diplomenids and kinetoplastids, *E. gracilis* possesses a collection of heterogeneous linear chromosomes ranging in size from a distribution peak at 4 kbp, up to 8 kbp (Spencer and Gray 2011; Dobáková et al. 2015). Only seven protein coding genes (*cox1*, *cox2*, *cox3*, *cob*, *nad1*, *nad4*, and *nad5*) have been identified in the genome (Dobáková et al. 2015; Tessier et al. 1997; Yasuhira and Simpson 1997). This is predicted to be the full complement of protein-coding genes in the mtDNA, with the remaining proteins likely encoded in the nuclear genome. Comparison of the gene

sequence and corresponding mRNA for these genes shows no evidence that editing or splicing is required for the formation of mature transcripts (Dobáková et al. 2015; Spencer and Gray 2011). This is quite surprising as unique and extensive mRNA editing appears to be a core feature of RNA maturation in the mitochondria of many other Euglenozoans. A second surprising feature of *E. gracilis* mtDNA is that in addition to full-length versions of mitochondrial genes, there are also many small mRNA and rRNA gene fragments scattered throughout the genome (Spencer and Gray 2011). These fragments retain high sequence identity to segments of the full length genes, in some cases even being perfect matches, but do not appear to be expressed. These small fragments and the presence of many short direct repeats have been proposed as possible evolutionary predecessors to the minicircle-encoded gRNAs of kinetoplastids, possibly produced through recombination between flanking repeats to produce "guide-like recombination products" (Spencer and Gray 2011). Transcription of the complementary strand of the gene fragments could then result in anti-sense RNAs capable of base-pairing to mRNAs, potentially allowing sequence drift in protein coding regions that could be corrected by RNA editing.

The mitochondrial genomes of two other euglenids, *Peranema trichophorum* and *Petalomonas cantuscygni* have been examined using electron microscopy (Roy et al. 2007). These results show that the *P. trichophorum* genome consists of many linear DNA molecules ranging from 1 to 75 kbp in size. In contrast, *P. cantuscygni* possesses linear 40 kbp molecules, with a small number of circular 40 kbp and much smaller 1–2.5 kbp molecules. More comprehensive examination of mitochondrial genome structure and content in other euglenids will indicate whether linear chromosomes are the predominant form and whether RNA editing is present in euglenids other than kinetoplastids.

The diversity found in structure and transcript processing in Euglenozoan mitochondria raises many questions about the evolutionary history that gave rise to these various states. Flegontov et al. have suggested that the genome of the Euglenozoans last common ancestor (ELCA) was likely circular and that the diversity found in this phylum may have arisen through constructive neutral evolution (Flegontov et al. 2011). It will be important to examine more representatives of all three major groups to determine the extent of possible genome types and novel mechanisms for RNA processing in these organelles.

8.5.1 Mitochondrial Ribosomal RNA

Ribosomal RNA structure and processing in Euglenozoan mitochondria is also highly variable. The mitochondrial SSU and LSU RNAs from *E. gracilis* have been identified and each appears to be expressed as two separate RNAs, termed SSU-R/SSU-L and LSU-R/LSU-L (Spencer and Gray 2011). Both SSU rRNA fragments have been sequenced and found to be chromosomally-unlinked independently transcribed genes, rather than products of cleavage of a single initial contiguous pre-SSU rRNA transcript. Extensive analysis failed to detect any full-length mature SSU RNAs providing strong evidence that these bipartite RNAs represent the mature fragmented functional form of this rRNA, not being further processed through a *trans*-splicing pathway to form a single contiguous SSU RNA. The 3′ end of SSU-R shows little heterogeneity. The SSU-L consists of three variants containing between 1 to 3 terminal A's at its 3′ end. Two LSU fragments have also been identified and found to have discrete 3′ ends; however, full length genomic encoding regions could not be located for either fragment. While it is likely that these represent the functional mitochondrial LSU RNAs, further analysis will have to be done to determine whether, like the SSU, each fragment is encoded individually and contiguously in the genome. Evidence has also been found that both the LSU and SSU contain modified nucleosides, including two tandem N^6,N^6-dimethyladenosines and an N^4-methylcytidine in the SSU and a Ψ in the LSU (Spencer and Gray 2011). Structural modeling has been performed for both the SSU and LSU fragments. The SSU fragments were found to form conserved long range base-pairing interactions

resulting in the formation of a secondary structure with similar features to the eubacterial 16S SSU rRNA. The first several hundred nucleotides of the 5′ end of the SSU show the greatest divergence in structure as a result of a high A + T content. LSU terminal regions also showed great similarity to the eubacterial 23S LSU RNA. In comparison, kinetoplast rRNA secondary structure has been found to be even more divergent from the eubacterial rRNA structure and in fact shows relatively little structural similarity to the *E. gracilis* rRNA.

Fragmented mt-rRNA has also been identified in the diplonemid *Diplonema papillatum*. Like *E. gracilis*, two LSU fragments (534 and 352 nt) are present, encoded on two Class B chromosomes (Valach et al. 2014). These RNAs are *trans*-spliced to produce a single LSU rRNA of approximately 900 nt and appear to go through other additional processing steps. The 3′ fragment contains a poly-A tail that is not present in the mature spliced transcript nor encoded in the gene sequence. The presence of this transient poly-A stretch raises questions about possible extended poly-A tail processing intermediates for the *E. gracilis* SSU-L, considering the observed variable 3′ ends (see above). In *Diplonema papillatum*, a 26 nucleotide poly-U stretch is found separating the 5′ and 3′ portions of the mature spliced LSU that is not encoded in the genes for either LSU fragment indicating that a process related to uridine insertion into mRNA modules can also occur to diplonemid rRNA. A short 366 nt RNA has been proposed as a potential mitochondrial SSU rRNA, but as of yet it is unclear if this represents the entire SSU rRNA or an individual fragment (Moreira et al. 2016). Small rRNAs are not unheard of in Euglenozoans. The kinetoplastid species *Trypanosoma brucei* (Sloof et al. 1985; Eperon et al. 1983), *Leishmania tarentolae* (de la Cruz et al. 1985a, b), and *Crithidia fasciculata* (Sloof et al. 1985) possess the smallest yet identified mitochondrial rRNAs, composed of a 9S SSU (approximately 611–640 nt) and 12S LSU (approximately 1141 and 1230 nt), each expressed as a contiguous transcript from a single gene. It will be important then to examine the SSU in *D. papillatum* and determine if other fragments are required or if the single SSU rRNA represents a potentially minimal rRNA.

In summary, while recent studies have begun to identify key features of the *E. gracilis* mitochondrial genome, our current knowledge about its structure and expression is still lagging somewhat behind what has been elucidated for other Euglenozoans. Continued efforts to characterize *Euglena* RNAs will be required to both further investigate the possibility of unique processing mechanisms and define the full complement of mtDNA encoded genes, including the LSU subunits.

8.6 Spliced-Leader RNA

Spliced-leader *trans*-splicing is a process through which a short RNA sequence (called the spliced-leader exon) is added to form the 5′ end of nuclear pre-mRNAs in a spliceosome-dependent manner. A small non-coding RNA termed the spliced-leader (SL) RNA acts as the donor of the short sequence. It is composed of two regions: the spliced-leader exon at its 5′ end followed by an extended sequence termed the spliced-leader intron (Fig. 8.2a), that is not included in the mature mRNA but is important for forming interactions with the target mRNA. SL RNAs fold into stem-loop secondary structures and contain an internal Sm-protein binding site, similar to what is observed in several of the small nuclear RNAs (snRNAs) of the spliceosome. The 5′ splice site required for the splicing reaction is part of the SL RNA, while the branch point adenosine, polypyrimidine tract and 3′ splice site are located at the 5′ end of the precursor mRNA collectively referred to as the "outron" (Fig. 8.2b, c). Together with the spliceosome components these elements form a substrate competent for splicing.

Spliced-leader *trans*-splicing was described in euglenids (Tessier et al. 1991) following initial discovery in trypanosomatids (Boothroyd and Cross 1982; Sutton and Boothroyd 1986; Milhausen et al. 1984) and nematodes (Krause and Hirsh 1987). Addition of the spliced-leader

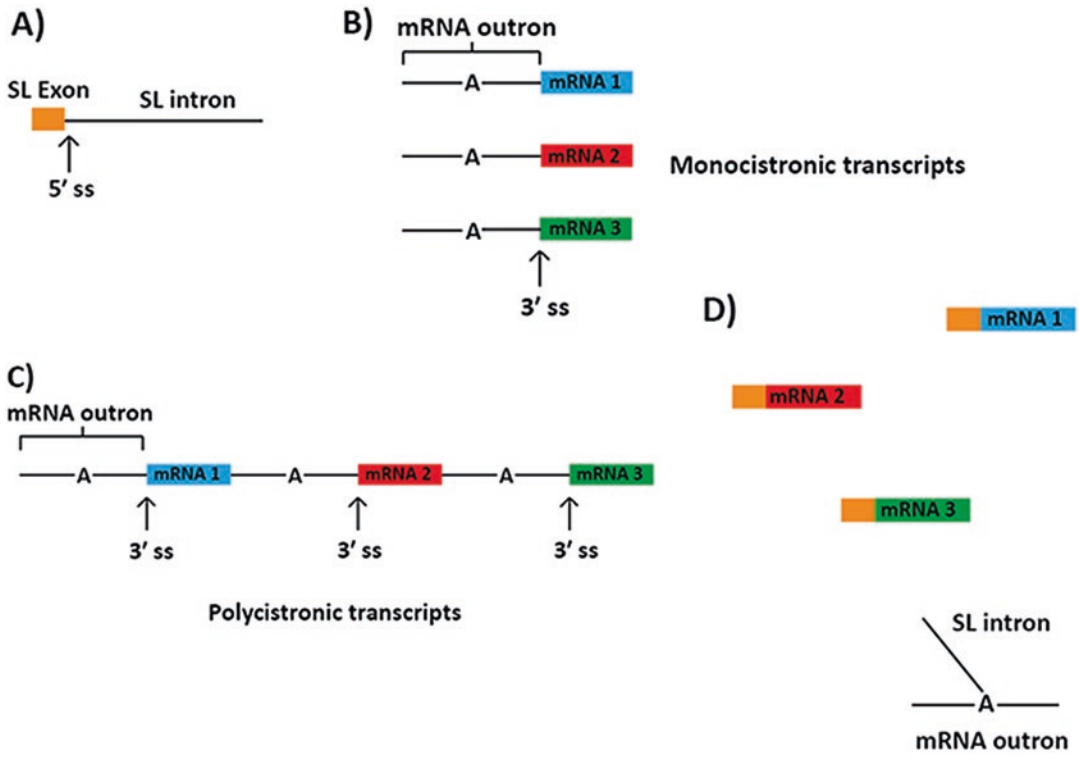

Fig. 8.2 (**a**) General structure of a spliced-leader RNA. Spliced leader *trans*-splicing can add the spliced-leader exon cap structure to (**b**) monocistronic transcripts or (**c**) to liberate individual protein-coding RNAs contained within a precursor polycistronic transcript. Both processing pathways result in (**d**) capped individual transcripts and removed Y shaped introns made up of the spliced-leader intron sequence attached to the mRNA outron. *ss* = splice site

serves a number of purposes in different groups of organisms. In both *C. elegans* (Spieth et al. 1993) and trypanosomes (Muhich and Boothroyd 1988), addition of the spliced-leader exon acts as a mechanism for processing and capping of individual mRNAs contained within long polycistronic precursor transcripts, as well as for capping monocistronic transcripts (Johnson et al. 1987; Zorio et al. 1994) (Fig. 8.2b–d). Analysis of the *E. gracilis* transcriptome has estimated that approximately 56% of pre-mRNA transcripts undergo spliced-leader exon sequence addition (Yoshida et al. 2016). Little is currently known about whether *Euglena* mRNAs are transcribed mono- or polycistronically and therefore the various roles of SL splicing in *Euglena* remains to be determined.

The *E. gracilis* genome encodes at least six variants of a spliced-leader RNA. Each isoform is approximately 101 nucleotides in length, the first 26 nucleotides of which is the SL exon sequence that is added to the pre-mRNA transcript (Tessier et al. 1991). The specific type of cap structure (and extent of modification) of the *Euglena* spliced-leader RNA is unknown. In most organisms in which spliced-leader *trans*-splicing occurs, the SL RNA possess a 2, 2, 7 trimethylguanosine (TMG) cap. Trypanosome SL exons possess a unique type of cap structure termed ‘cap 4’ containing extensive modifications; including 7-methyl guanosine, 2′-O-methylation of the first four nucleotides, additional base methylations at the first and fourth nucleotides, and a Ψ at position 28 (Zamudio et al. 2009). The relatively close phylogenetic relationship between trypanosomes and *E. gracilis* suggests that a number of these modifications may also be present in *Euglena.* Information on *Euglena* SL cap structure may be lacking in part because recent studies indicate that there are an additional two

nucleotides at the 5′ end of the SL RNA exon from what was previously reported (our unpublished results). This is critical as these would be the nucleotides containing most of modified nucleotide positions for these RNAs. These additional nucleotides also prompt further questions about how SL RNA is expressed in *E. gracilis* and whether initial processing of pre-SL RNA may occur prior to capping and splicing. We are learning an increasing amount about SL *trans*-splicing in *E. gracilis* but the exact role and structure of this RNA requires further elucidation.

8.7 RNase P

Efficient and accurate processing of pre-tRNA molecules from both nuclear and organellar genomes is crucial for the production of functional tRNA molecules. RNase P is a key endonucleolytic enzymatic complex responsible for the maturation of the 5′ ends of tRNAs. Found in all three domains of life, RNase P most commonly functions as a ribonucleoprotein complex containing a single RNA (RNase P RNA) which is the catalytic component (Guerrier-Takada et al. 1983; Pannucci et al. 1999; Thomas et al. 2000; Kikovska et al. 2006), and a variable number of proteins depending on the species. A small number of protein-only RNase Ps (PRORP) have also been identified, primarily confined to the organelles of eukaryotes (Holzmann et al. 2008; Gobert et al. 2010). Interestingly however, several members of the phylum Euglenozoa appear to only possess protein-only versions of RNase P. A single predicted PRORP protein has been identified in *Euglena mutabilis* and many trypanosome species possess nuclear (PRORP1) and mitochondrial (PRORP2) protein-only enzymes that have been shown to accurately process 5′ tRNA ends *in vitro* in the absence of any additional protein or RNA factors (Lechner et al. 2015; Taschner et al. 2012). To date, no RNase P has been reported for the plastid or nuclear genomes of *Euglena gracilis* (Lechner et al. 2015). However, when we performed a blastp search using *Trypanosoma brucei* PRORP proteins it revealed a putative PRORP protein in the *E. gracilis* proteome database published by O'Neill et al. (2015) that contains both PRORP and PPR motif repeat domains like those found in other protein- only RNase Ps (our unpublished results). Whether this PRORP-like protein in *E. gracilis* possesses tRNA processing activity and whether or not *E. gracilis* also possesses an RNA dependent RNase P activity will need to be examined. The distribution of the apparent utilization of PRORP enzymes in Euglenozoa suggests that dependence on RNase P RNA may have been lost early in the evolution of this phylum.

8.8 Conclusions and Future Directions

RNA transcript processing in *Euglena* displays remarkable diversity when compared to similar processes in distantly-related eukaryotes but also compared to its most closely-related studied relatives, the kinetoplastids and diplonemids. While some information is now available about processing of select classes of *Euglena* RNA in nuclei, mitochondria and chloroplasts, our knowledge is still limited due to a lack of characterization of RNA processing factors and an incomplete understanding of nuclear and mitochondrial genome structure. These will be key research areas to investigate in the future that should be aided by advances in proteomics and high-throughput nucleic acid sequencing technologies. Also important will be the isolation and characterization of classes of non-coding RNAs through RNA-Seq approaches that will give a more complete picture of the abundance and diversity of non-coding RNA types in different *Euglena* species. So far, key elucidated features are that polycistronic transcription is a common gene expression strategy for several classes of *Euglena* nuclear and chloroplast RNA (unknown for mitochondrial transcripts at this stage), that both cytoplasmic and mitochondrial rRNA is unusually structurally fragmented (extensively so for cytosolic LSU rRNA), that novel introns are prevalent in both organelle and nuclear genes, and that non-coding RNAs and their sequence isoforms are apparently very abundant in *Euglena* which

seems to be related to the repetitiveness of its nuclear genome structure. What roles might these unusual RNA structural features and transcript processing mechanisms have on the environmental adaptability of *Euglena*? Additional surprising features and mechanisms will likely be discovered when we continue our efforts to study this fascinating genus that may provide new insights into the evolution of RNA and protein-RNA complexes in all organisms.

References

Barneche F, Gaspin C, Guyot R, Echeverría M (2001) Identification of 66 box C/D snoRNAs in *Arabidopsis thaliana*: extensive gene duplications generated multiple isoforms predicting new ribosomal RNA 2′-O-methylation sites. J Mol Biol 311:57–73

Basu MK, Rogozin IB, Koonin EV (2008) Primordial spliceosomal introns were probably U2-type. Trends Genet 24:525–528

Bennett MS, Triemer RE (2015) Chloroplast genome evolution in the Euglenaceae. J Eukaryot Microbiol 62:773–785

Boothroyd JC, Cross GAM (1982) Transcripts coding for variant surface glycoproteins of *Trypanosoma brucei* have a short, identical exon at their 5′ end. Gene 20:281–289

Bratkovič T, Rogelj B (2014) The many faces of small nucleolar RNAs. Biochim Biophys Acta 1839: 438–443

Breckenridge DG, Wantanabe Y, Greenwood SJ, Gray MW, Schnare MN (1999) U1 small nuclear RNA and spliceosomal introns in *Euglena gracilis*. Proc Natl Acad Sci 96:852–856

Brown JWS, Echeverria M, Qu L (2003) Plant snoRNAs: functional evolution and new modes of gene expression. Trends Plant Sci 8:42–49

Burge CB, Padgett RA, Sharp PA (1998) Evolutionary fates and origins of U12-type introns. Mol Cell 2:773–785

Canaday J, Tessier LH, Imbault P, Paulus F (2001) Analysis of *Euglena gracilis alpha-, beta-* and *gamma*-tubulin genes: introns and pre-mRNA maturation. Mol Gen Genomics 265:153–160

Charette JM, Gray MW (2009) U3 snoRNA genes are multi-copy and frequently linked to U5 snRNA genes in *Euglena gracilis*. BMC Genomics 10:528–546

Christopher DA, Hallick RB (1990) Complex RNA maturation pathway for a chloroplast ribosomal protein operon with an internal tRNA Cistron. Plant Cell 2:659–671

Cook JR, Roxby R (1985) Physical properties of a plasmid-like DNA from *Euglena gracilis*. Biochim Biophys Acta 824:80–83

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

Corell RA, Feagin JE, Riley GR, Strickland T, Guderian JA, Myler PJ, Stuart K (1993) *Trypanosoma brucei* minicircles enode multiple guide RNAs which can direct editing of extensively overlapping sequences. Nucleic Acids Res 21:4313–4320

Dabbagh N, Preisfeld A (2017) The chloroplast genome of *Euglena mutabilis*—cluster arrangement, intron analysis, and Intrageneric trends. J Eukaryot Microbiol 64(1):31–44

de la Cruz VF, Lake JA, Simpson AM, Simpson L (1985a) A minimal ribosomal RNA: sequence and secondary structure of the 9S kinetoplast ribosomal RNA from *Leishmania tarentolae*. Proc Natl Acad Sci 82: 1401–1405

de la Cruz VF, Simpson AM, Lake JA, Simpson L (1985b) Primary sequence and partial secondary structure of the 12S kinetoplast (mitochondrial) ribosomal RNA from *Leishmania tarentolae*: conservation of peptidyltransferase structural elements. Nucleic Acids Res 13:2337–2356

Deschamps P, Lara E, Marande W, López-García P, Ekelund F, Moreira D (2011) Phylogenomic analysis of kinetoplastids supports that trypanosomatids arose from within Bodonids. Mol Biol Evol 28:53–58

Dobáková E, Flegontov P, Skalický T, Lukeš J (2015) Unexpectedly streamlined mitochondrial genome of the euglenozoan *Euglena gracilis*. Genome Biol Evol 7:3358–3367

Eliaz D, Doniger T, Tkacz ID, Biswas VK, Gupta S, Kolev NG, Unger R, Ullu E, Tschudi C, Michaeli S (2015) Genome-wide analysis of small nucleolar RNAs of Leishmania major reveals a rich repertoire of RNAs involved in modification and processing of rRNA. RNA Biol 12:1222–1255

Eperon IC, Janssen JWG, Hoeijmakers JHJ, Borst P (1983) The major transcripts of the kinetoplast DNA of *Trypanosoma brucei* are very small ribosomal RNAs. Nucleic Acids Res 11:105–125

Fantoni A, Dare AO, Tschudi C (1994) RNA polymerase III-mediated transcription of the trypanosome U2 small nuclear RNA Gene is controlled by both intragenic and Extragenic regulatory elements. Mol Cell Biol 14:2021–2028

Ferreira VS, Rocchetta I, Conforti V, Bench S, Feldman R, Levin MJ (2007) Gene expression patterns in *Euglena gracilis*: insights into the cellular response to environmental stress. Gene 389:136–145

Flegontov P, Gray MW, Burger G, Lukeš J (2011) Gene fragmentations: a key to mitochondrial genome evolution in Euglenozoa? Curr Genet 57:225–232

Geimer S, Belicová A, Legen J, Sláviková S, Herrmann RG, Krajčovič J (2009) Transcriptome analysis of the *Euglena gracilis* plastid chromosome. Curr Genet 55:425–438

Gobert A, Gutmann B, Taschner A, Gößringer M, Holzmann J, Hartmann RK, Rossmanith W, Giegé P

(2010) A single *Arabidopsis* organellar protein has RNase P activity. Nat Struct Mol Biol 17:740–744
Greenberg BM, Hallick RB (1986) Accurate transcription and processing of 19 *Euglena* chloroplast tRNAs in a *Euglena* soluble extract. Plant Mol Biol 6:89–100
Greenberg BM, Narita JO, Deluca-Flaherty C, Gruissem W, Rushlow KA, Hallick RB (1984) Evidence for two RNA polymerase activities in *Euglena gracilis* chloroplasts. J Biol Chem 259:14880–14887
Greenwood SJ, Gray MW (1998) Processing of precursor rRNA in *Euglena gracilis*: identification of intermediates in the pathway to a highly fragmented large subunit rRNA. Biochim Biophys Acta 1443:128–138
Greenwood SJ, Schnare MN, Cook JR, Gray MW (2001) Analysis of intergenic spacer transcripts suggests 'read-around' transcription of the extrachromosomal circular rDNA in *Euglena gracilis*. Nucleic Acids Res 29:2191–2198
Greenwood SJ, Schnare MN, Gray MW (1996) Molecular characterization of U3 small nucleolar RNA from the early diverging protist, *Euglena gracilis*. Curr Genet 30:338–346
Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S (1983) The RNA moiety of Ribonuclease P is the catalytic subunit of the enzyme. Cell 35:849–857
Hafez M, Hausner G (2015) Convergent evolution of twintron-like configurations: one is never enough. RNA Biol 12:1275–1288
Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, Spielmann A, Stutz E (1993) Complete sequence of *Euglena gracilis* chloroplast DNA. Nucleic Acids Res 21:3537–3544
Holzmann J, Frank P, Löffler E, Bennett KL, Gemer C, Rossmanith W (2008) RNase P without RNA: identification and functional reconstitution of the human mitochondrial tRNA processing enzyme. Cell 135:462–474
Hong L, Hallick RB (1994) A group III intron is formed from domains of two individual group II introns. Genes Dev 8:1589–1599
Hrdá Š, Fousek J, Szabova J, Hampl VV, Vlček Č (2012) The plastid genome of *Eutreptiella* provides a window into the process of secondary endosymbiosis of plastid in euglenids. PLoS One 7:1–10
Jasmer DP, Stuart K (1986a) Conservation of kinetoplastid minicircle characteristics without nucleotide sequence conservation. Mol Biochem Parasitol 18: 257–269
Jasmer DP, Stuart K (1986b) Sequence Organization in African Trypanosome Minicircles is defined by 18 base pair inverted repeats. Mol Biochem Parasitol 18:321–331
Johnson PJ, Kooter JM, Borst P (1987) Inactivation of transcription by UV irradiation of *T. brucei* provides evidence for a multicistronic transcription unit including a VSG Gene. Cell 51:273–381
Keller M, Tessier LH, Chan RL, Weil JH, Imbault P (1992) In *Euglena*, spliced-leader RNA (SL-RNA) and 5S rRNA genes are tandemly repeated. Nucleic Acids Res 20:1711–1715
Kiethega GN, Turcotte M, Burger G (2011) Evolutionarily conserved cox1 trans-splicing without cis-motifs. Mol Biol Evol 28:2425–2428
Kiethega GN, Yan Y, Turcotte M, Burger G (2013) RNA-level unscrambling of fragmented genes in *Diplonema* mitochondria. RNA Biol 10:301–313
Kikovska E, Svärd SG, Kirsebom LA (2006) Eukaryotic RNase P RNA mediates cleavage in the absence of protein. Proc Natl Acad Sci 104:2062–2067
Kiss T, Marshallsay C, Filipowicz W (1991) Alteration of the RNA polymerase specificity of U3 snRNA genes during evolution and in vitro. Cell 65:517–526
Kleisen CM, Weislogel PO, Fonck K, Borst P (1976) The structure of kinetoplast DNA 2. Characterization of a novel component of high complexity present in the kinetoplast DNA network of *Crithidia luciliae*. Eur J Biochem 64:153–160
Koslowsky DJ, Bhat GJ, Perrollaz AL, Feagin JE, Stuart K (1990) The MURF3 Gene of *T. brucei* contains multiple domains of extensive editing and is homologous to a subunit of NADH dehydrogenase. Cell 62:901–911
Krause M, Hirsh D (1987) A trans-spliced leader sequence on actin mRNA in *C. elegans*. Cell 49:753–761
Lechner M, Rossmanith W, Hartmann RK, Thölken C, Gutmann B, Giegé P, Gobert A (2015) Distribution of Ribonucleoprotein and Protein-only RNase P in Eukarya. Mol Biol Evol 32:3189–3193
Li X, Ma S, Yi C (2016) Pseudouridine: the fifth RNA nucleotide with renewed interests. Curr Opin Chem Biol 33:108–116
Liang XH, Uliel S, Hury A, Barth S, Doniger T, Unger R, Michaeli S (2005) A genome-wide analysis of C/D and H/ACA-like small nucleolar RNAs in *Trypanosoma brucei* reveals a trypanosome-specific pattern of rRNA modification. RNA 11:619–645
López MD, Rosenblad MA, Samuelsson T (2009) Conserved and variable domains of RNase MRP RNA. RNA Biol 6:208–221
Lui L, Lowe T (2013) Small nucleolar RNAs and RNA-guided post-transcriptional modification. Essays Biochem 54:53–77
Marande W, Burger G (2007) Mitochondrial DNA as a genomic jigsaw puzzle. Science 318:415
Marande W, Lukeš J, Burger G (2005) Unique mitochondrial genome structure in diplonemids, the sister Group of Kinetoplastids. Eukaryot Cell 4:1137–1146
Marshallsay C, Connelly S, Filipowicz W (1992) Characterization of the U3 and U6 snRNA genes from wheat: U3 snRNA genes in monocot plants are transcribed by RNA polymerase III. Plant Mol Biol 19:973–983
Milanowski R, Gumińska N, Karnkowska A, Ishikawa T, Zakryś B (2016) Intermediate introns in nuclear genes of euglenids—are they a distinct type? BMC Evol Biol 16:1–11
Milanowski R, Karnkowska A, Ishikawa T, Zakryś B (2014) Distribution of conventional and nonconventional introns in tubulin (α and β) genes of euglenids. Mol Biol Evol 31:584–593

Milhausen M, Nelson RG, Sather S, Selkirk M, Agabian N (1984) Identification of a small RNA containing the trypanosome spliced leader: a donor of shared 5′ sequences of Trypanosomatid mRNAs? Cell 38: 721–729

Moore AN, Russell AG (2012) Clustered organization, polycistronic transcription, and evolution of modification-guide snoRNA genes in *Euglena gracilis*. Mol Gen Genomics 287:55–66

Moreira S, Valach M, Aoulad-Aissa M, Otto C, Burger G (2016) Novel modes of RNA editing in mitochondria. Nucleic Acids Res 44:4907–4919

Muchhal US, Schwartzbach SD (1992) Characterization of a *Euglena* gene encoding a polyprotein precursor to the light-harvesting chlorophyll *a/b*-binding protein of photosystem II. Plant Mol Biol 18:287–299

Muchhal US, Schwartzbach SD (1994) Characterization of the unique intron—exon junctions of *Euglena* gene(s) encoding the polyprotein precursor to the light-haresting chlorophyll a/b binding protein of photosystem II. Nucleic Acids Res 22:5737–5744

Muhich ML, Boothroyd JC (1988) Polycistronic transcripts in trypanosomes and their accumulation during heat shock: evidence for a precursor role in mRNA synthesis. Mol Cell Biol 8:3837–3846

O'Neill EC, Trick MH, Lionel RM, Dusi RG, Hamilton CJ, Zimba PV, Henrissat B, Field RA (2015) The transcriptome of *Euglena gracilis* reveals unexpected metabolic capabilities for carbohydrate and natural product biochemistry. Mol BioSyst 11:2808–2820

Pannucci JA, Haas ES, Hall TA, Harris JK, Brown JW (1999) RNase P RNAs from some Archaea are catalytically active. Proc Natl Acad Sci 96:7803–7808

Pollard VW, Rohrer SP, Michelotti EF, Hancock K, Hajduk SL (1990) Organization of Minicircle Genes for guide RNAs in Trypanosoma brucei. Cell 63: 783–790

Pombert J-F, James ER, Janouškovec J, Keeling PJ (2012) Evidence for transitional stages in the evolution of euglenid group II introns and twintrons in the *Monomorphina aenigmatica* plastid genome. PLoS One 7:1–8

Ravel-Chapuis P (1988) Nuclear rDNA in *Euglena gracilis*: paucity of chromosomal units and replication of extrachromosomal units. Nucleic Acids Res 16: 4801–4810

Revel-Chapuis P, Nicolas P, Nigon V, Neyret O, Freyssinet G (1985) Extrachromosomal circular nuclear rDNA in *Euglena gracilis*. Nucleic Acids Res 13:7529–7537

Riou G, Delain E (1969) Electron microscopy of the circular Kinetoplastic DNA from Trypanosoma cruzi: occurrence of catenated forms. Proc Natl Acad Sci 62:210–217

Roy J, Faktorová D, Lukeš J, Burger G (2007) Unusual mitochondrial genome structures throughout the Euglenozoa. Protist 158:385–396

Russell AG, Schnare MN, Gray MW (2004) Pseudouridine-guide RNAs and other Cbf5p-associated RNAs in *Euglena gracilis*. RNA 10:1034–1046

Russell AG, Schnare MN, Gray MW (2006) A large collection of compact box C/D snoRNAs and their isoforms in *Euglena gracilis*: structural functional and evolutionary insights. J Mol Biol 357:1545–1565

Russell AG, Wantanabe Y, Charette JM, Gray MW (2005) Unusual features of fibrillarin cDNA and gene structure in *Euglena gracilis*: evolutionary conservation of core proteins and structural predictions for methylation-guide box C/D snoRNPs throughout the domain Eucarya. Nucleic Acids Res 33:2781–2791

Schnare MN, Cook JR, Gray MW (1990) Fourteen internal transcribed spacers in the circular ribosomal DNA of *Euglena gracilis*. J Mol Biol 215:85–91

Schnare MN, Gray MW (1990) Sixteen discrete RNA components in the cytoplasmic ribosome of *Euglena gracilis*. J Mol Biol 215:73–83

Schnare MN, Gray MW (2011) Complete modification maps for the cytosolic small and large subunit rRNAs of *Euglena gracilis*: functional and evolutionary implications of contrasting patterns between the two rRNA components. J Mol Biol 413:66–83

Sharma S, Lafontaine DLJ (2015) 'View from a bridge': a new perspective on eukaryotic rRNA base modification. Trends Biochem Sci 40:560–575

Simpson L, Neckelmann N, de la Cruz VF, Simpson AM, Feagin JE, Jasmer DP, Stuart K (1987) Comparison of the maxicircle (mitochondrial) genomes of *Leishmania tarentolae* and *Trypanosoma brucei* at the level of nucleotide sequence. J Biol Chem 262:6182–6196

Sloof P, Van den Burg J, Voogd A, Benne R, Agostinelli M, Borst P, Gutell R, Noller H (1985) Further characterization of the extremely small mitochondrial ribosomal RNAs from trypanosomes: a detailed comparison of the 9S and 12S RNAs from *Crithidia fasciculata* and *Trypanosoma brucei* with rRNAs from other organisms. Nucleic Acids Res 13:4171–4190

Spencer DF, Gray MW (2011) Ribosomal RNA genes in *Euglena gracilis* mitochondrial DNA: fragmented genes in a seemingly fragmented genome. Mol Gen Genomics 285:19–31

Spieth J, Brooke G, Kuersten S, Lea K, Blumenthal T (1993) Operons in *C. elegans*: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. Cell 73:521–532

Steinert M, Van Assell S (1975) Large circular mitochondrial DNA in *Crithidia luciliae*. Exp Cell Res 96: 406–409

Sutton RE, Boothroyd JC (1986) Evidence for trans splicing in trypanosomes. Cell 47:527–535

Taschner A, Weber C, Buzet A, Hartmann RK, Hartig A, Rossmanith W (2012) Nuclear RNase P of *Trypanosoma brucei*: a single protein in place of the multicomponent RNA-protein complex. Cell Rep 2:19–25

Tessier LH, Keller M, Chan RL, Fournier R, Weil JH, Imbault P (1991) Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in *Euglena*. EMBO J 10:2621–2625

Tessier LH, van der Speck H, Gualberto JM, Grienenberger JM (1997) The cox1 gene from *Euglena gracilis*: a

protist mitochondrial gene without introns and genetic code modifications. Curr Genet 31:208–213

Thomas BC, Chamberlain J, Engelke DR, Gegenheimer P (2000) Evidence for an RNA-based catalytic mechanism in eukaryotic nuclear ribonuclease P. RNA 6:554–562

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 MD, Zhang L, Hong L, Hallick RB (1997) Two new group-II twintrons in the *Euglena gracilis* chloroplast are absent in basally branching *Euglena* species. Curr Genet 31:89–95

Valach M, Moreira S, Kiethega GN, Burger G (2014) Trans-splicing and RNA editing of LSU rRNA in *Diplonema* mitochondria. Nucleic Acids Res 42: 2660–2672

Vlcek C, Marande W, Teijeiro S, Lukeš J, Burger G (2010) Systematically fragmented genes in a multipartite mitochondrial genome. Nucleic Acids Res 39:979–988

Westenberger SJ, Cerqueira GC, El-Sayed NM, Zingales B, Campbell DA, Sturm NR (2006) Trypanosoma cruzi mitochondrial maxicircles display species- and strain-specific variation and a conserved element in the non-coding region. BMC Genomics 7:1–18

Wiegert KE, Bennett MS, Triemer RE (2012) Evolution of the chloroplast genome in photosynthetic Euglenoids: a comparison of *Eutreptia viridis* and *Euglena gracilis* (Euglenophyta). Protist 163: 832–843

Yasuhira S, Simpson L (1997) Phylogenetic affinity of mitochondria of *Euglena gracilis* and kinetoplastids using cytochrome oxidase I and hsp60. J Mol Evol 44:341–347

Yoshida Y, Tomiyama T, Maruta T, Tomita M, Ishikawa T, Arakawa K (2016) *De novo* assembly and comparative transcriptome analysis of *Euglena gracilis* in response to anaerobic conditions. BMC Genomics 17:1–10

Zamudio JR, Mittra B, Chattopadhyay A, Wohlschlegel JA, Sturm NR, Campbell DA (2009) *Trypanosoma brucei* spliced leader RNA maturation by the cap 1 2′-*O*-ribose Methyltransferase and SLA1 H/ACA snoRNA Pseudouridine synthase complex. Mol Cell Biol 29:1202–1211

Zorio DAR, Cheng NN, Blumenthal T, Spieth J (1994) Operons as a common form of chromosomal organization in *C. elegans*. Nature 372:270–272