RNA BIOLOGY
2018, VOL. 15, NO. 10, 1309–1318
https://doi.org/10.1080/15476286.2018.1526561

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

RNA-Seq employing a novel rRNA depletion strategy reveals a rich repertoire of snoRNAs in *Euglena gracilis* including box C/D and Ψ-guide RNAs targeting the modification of rRNA extremities

Ashley N. Moore*, David C. McWatters*, Andrew J. Hudson, and Anthony G. Russell

Department of Biological Sciences, and Alberta RNA Research and Training Institute, University of Lethbridge, Lethbridge, AB, Canada

ABSTRACT

Previous mRNA transcriptome studies of *Euglena gracilis* have shown that this organism possesses a large and diverse complement of protein coding genes; however, the study of non-coding RNA classes has been limited. The natural extensive fragmentation of the *E. gracilis* large subunit ribosomal RNA presents additional barriers to the identification of non-coding RNAs as size-selected small RNA libraries will be dominated by rRNA sequences. In this study we have developed a strategy to significantly reduce rRNA amplification prior to RNA-Seq analysis thereby producing a ncRNA library allowing for the identification of many new *E. gracilis* small RNAs. Library analysis reveals 113 unique new small nucleolar (sno) RNAs and a large collection of snoRNA isoforms, as well as the first significant collection of nuclear tRNAs in this organism. A 3′ end AGAUGN consensus motif and conserved structural features can now be defined for *E. gracilis* pseudouridine guide RNAs. snoRNAs of both classes were identified that target modification of the 3′ extremities of rRNAs utilizing predicted base-pairing interactions with internally transcribed spacers (ITS), providing insight into the timing of steps in rRNA maturation. Cumulatively, this represents the most comprehensive analysis of small ncRNAs in *Euglena gracilis* to date.

ARTICLE HISTORY
Received 6 July 2018
Revised 21 August 2018
Accepted 16 September 2018

KEYWORDS
Euglena gracilis; non-coding RNA; small nucleolar RNA; RNA modification; RNA-Seq methods; box C/D RNA; pseudouridine-guide RNA; RNA evolution; ribosome biogenesis

Introduction

Non-coding RNAs (ncRNAs) have essential roles in an array of gene expression mechanisms in all organisms [1–6]. Comprehensive strategies to identify ncRNAs have been developed utilizing deep sequencing technologies in the form of RNA 'sequencing' (RNA-Seq). During this procedure it is advantageous to fractionate and enrich the ncRNA population of interest from total cellular RNA prior to cDNA library creation to remove very abundant cellular RNAs that would dominate the sequence reads. This allows for the more efficient and cost-effective identification of less abundant and novel non-coding RNA species. Commercial kits have been developed to remove rRNA during library preparation; however, they are only available (or work efficiently) for a limited number of model organisms. They also increase the number of sample handling steps which can further increase the likelihood of generating unnatural RNA degradation products. Such kits are not yet available for most protists, a collection of primarily single-celled eukaryotic organisms that includes *Euglena gracilis*, the organism investigated in this study.

E. gracilis is a particularly interesting organism in which to characterize ncRNAs because of the many unusual features of its cellular biology and gene expression strategies that suggest it may contain a large collection of ncRNAs [7,8]. Recently, mRNA transcriptome studies have been performed [9,10] that indicate that *E. gracilis* has extensive protein-coding potential and it has been suggested that expression of nuclear protein-coding genes is extensively controlled at the post-transcriptional level. This organism contains a large subunit (LSU) rRNA that is naturally fragmented into 14 discrete pieces (compared to 2 in most other eukaryotes) by post-transcriptional processing events [11,12]. *E. gracilis* rRNA is also the most extensively modified of any examined organism to date, containing 211 2′-*O*-methylations (Nm) and 116 pseudouridine (Ψ) modifications [13]. The LSU rRNA is more extensively modified than the non-fragmented small subunit (SSU) rRNA, which instead has a similar amount of modification as its human counterpart. These extra modifications are predicted to help stabilize the highly fragmented LSU during ribosome assembly [13].

Small nucleolar RNAs (snoRNAs) are one class of ncRNA, most commonly used as part of ribonucleoprotein complexes (RNPs) to guide site-specific rRNA nucleotide modification, namely 2′-*O*- methylation and Ψ formation, and in the targeting of pre-rRNA cleavage sites. Since *E. gracilis* has so many rRNA modifications and processing sites it is predicted to also contain a large collection of targeting snoRNAs. Previously we had identified snoRNAs that guide 47% of the experimentally mapped rRNA 2′-*O*-methylation sites but only 11% of the Ψ sites. Two of the *E. gracilis* rRNA nucleotide modifications are located very close to the 3′ ends of two different rRNA species. We had not yet uncovered any snoRNA species capable of

CONTACT Anthony G. Russell tony.russell@uleth.ca Department of Biological Sciences, University of Lethbridge, Lethbridge, AB, Canada

*These authors contributed equally to this manuscript.

Supplemental data for this article can be accessed here.

targeting these sites and thus the mechanism of modification was still unclear.

The small number of *E. gracilis* Ψ-guide RNAs identified previously structurally differ from the H/ACA box snoRNAs first characterized in other eukaryotes. The prototypical structure consists of two extended stems, either or both of which are interrupted by single-stranded regions that base-pair to rRNA to form pseudouridylation guide pockets. A single-stranded linker region containing the H box sequence (ANANNA) separates the two stems, and an 'ACA' consensus sequence box element follows the second stem and is usually located 3 nt from the 3′ end of the RNA [14]. In contrast, the small number of *Euglena* Ψ-guide RNAs identified previously possess only a single-stem structure (no H box) and possess an AGA rather than an ACA (box) sequence motif at their 3′ ends [15,16]. AGA box Ψ-guide RNAs have also been identified in trypanosome species [17,18]. Whether or not this is a structurally common form for these RNAs in Euglenozoa, the evolutionarily-diverse phylum containing the euglenids (including *Euglena* species), kinetoplastids (including trypanosomes) and many other classes of protist organisms, requires a more comprehensive characterization of Ψ-guide RNAs in *E. gracilis* and its relatives [15,16,19].

In this study, we used a newly developed RNA library preparation strategy for RNA-Seq experiments to identify and characterize a large collection of small ncRNAs in *E. gracilis*. These ncRNAs shed new light on the events of rRNA maturation, the evolution of Ψ-guide RNAs in *E. gracilis*, and provide new information on the structural and sequence characteristics of small nucleolar RNA classes.

Methods

Library construction

E. gracilis total RNA (~ 112 μg) was resolved on a 15% denaturing polyacrylamide gel and RNA fragments less than 400 nt in size were excised and isolated [20]. A poly-G tail was added to the 3′ ends of the size-selected RNA [21]. The tailing reaction contained size-selected or TMG cap-enriched *Euglena* RNA, 1X Poly(A) Polymerase (PAP) buffer (USB), 0.5 mM GTP, 60 U of yeast PAP (USB) and 20 U of RNase Inhibitor (NEB) incubated at 37°C for 60 min. The reaction was extracted once with phenol:chloroform (1:1), then twice with chloroform and the aqueous phase was ethanol precipitated with added acrylamide carrier. The RNA was then treated with 10 U of Tobacco Acid Pyrophosphatase (TAP) (epicentre®) in a 10 μL reaction containing 1X TAP buffer (epicentre®) and 20 U of RNase Inhibitor (NEB) at 37°C for 60 min and the RNA was extracted and precipitated (as above).

An RNA oligonucleotide linker was ligated to the 5′ termini of the TAP-treated RNA [21]. The RNA was first mixed with 200 pmol of linker and incubated at 65°C for 5 min. The ligation reaction containing 10 U of T4 RNA ligase (NEB), 1 mM ATP, 1X T4 RNA ligase buffer (NEB), and 20 U of RNase Inhibitor (NEB) was then performed at 4°C overnight (16 hrs), after which another 10 U of T4 RNA ligase was added and the reaction further incubated at 37°C for 30 min. The RNA was then extracted and precipitated.

An antisense primer containing an adaptor sequence and poly-C stretch was designed to anneal to the 3′ poly-G tail. This primer (100 pmole) was incubated with 10 μL of prepared RNA from the previous step and dNTPs (500 μM) at 65°C for 5 min and then immediately chilled on ice. Superscript II RT (Invitrogen) was used to synthesize cDNA at 47°C for 60 min following the manufacturer's protocol. The cDNA was then used as template for PCR amplification with Phusion Taq Polymerase (Thermo Scientific) using oligonucleotides designed to anneal to the 3′ poly-G tail and the 5′ linker sequence with or without the addition of blocking primers (also see below and **Table S1** for oligonucleotide sequences; **Table S6** for PCR conditions). When assessing relative levels of rRNA (and not employing the blocking primers), PCR products were purified by gel-extraction and cloned into the pJET1.2/blunt vector following the manufacturer's protocol. Transformed *E. coli* cells were then used for colony PCR screening, using primers that anneal upstream and downstream of the cloning site. Automated DNA sequencing of these PCR product clones was performed by Macrogen Corp USA.

Preventing amplification of large subunit rRNA fragments

Blocking primer sets were designed with a C3 spacer (3 hydrocarbon) modification at their 3′ end and each of these modified primers anneals both to the 3′ end of the added 5′ linker sequence and to the 5′ end of a specifically-targeted individual LSU rRNA fragment (see **Table S2** for blocker oligonucleotide sequences). At the PCR amplification step of library preparation, in addition to the general amplification primers, each blocking primer was also added to the reaction to a concentration of 5 pmole/μL to prevent amplification of the unwanted rRNA species. The final resulting PCR-generated cDNA library was purified using the E.Z.N.A ® Cycle-Pure Kit (Omega) and sent to Genome Québec for high-throughput sequencing using the Illumina MiSeq 250 platform.

Bioinformatic analysis

The Illumina MiSeq sequence reads, for both size-selected and TMG-enriched libraries, were first sorted based on the presence of the 5′ linker sequence using the FASTQ Barcode Splitter tool from the FASTX-Toolkit (Hannon Lab website). The 5′ and 3′ adaptor sequences were then removed (allowing 2 mismatches) using the Trim Ends tool in Geneious v8.0.4 software and cutadapt software package. Typically, the sequence quality was very poor following the 3′ poly-G tract and therefore the 3′ ends were trimmed downstream of a poly-G tract ≥ 12 nt long. The two most highly abundant sequences were also removed from the collection. The UCLUST algorithm, a component of the USearch [22] software package, was used to cluster related sequences together based on pair-wise alignments, using an identity threshold of 0.8. To remove previously characterized *Euglena* RNAs from the newly formed sequence clusters, databases of *E. gracilis* snoRNAs, rRNA, snRNAs and tRNAs were created. First, the UBlast algorithm [22] was used to find matches between the database and the RNA-Seq library sequences using an E-value of 1e-9. Then to ensure removal of as many sequences as possible, searches using the USearch global algorithm were performed with an id value of 1. Matches to these databases were subsequently removed prior to library analysis.

Two approaches were used to identify new snoRNAs. First, trimmed sequences between 50 – 80 nt in length were extracted (using Geneious) and then scanned for *E. gracilis* snoRNA features using the pattern matching program 'Scan for Matches'[23]. A consensus pattern was created based on all previously identified *Euglena* snoRNAs including size, sequence box elements, and secondary structure potential. To identify additional box C/D snoRNAs, trimmed sequences between 55 – 90 nt were analyzed using the Snoscan webserver [24] with *E. gracilis* rRNA sequence, including internal transcribed spacer sequences, as potential modification targets.

Positive hits from both approaches were then further manually inspected as previously described [15]. Sequences that strictly maintained conserved features of snoRNAs but did not display significant base-pairing potential to any mapped modified rRNA site were sorted into the orphan 'snoRNA' category. Reads per million (RPM) for snoRNA species were calculated using quality filtered single end reads from size-selected and TMG-capped libraries. Individual RNAs were quantified using USearch algorithm searches with an id value of 0.95, then normalized for library size.

For tRNA identification, the USearch algorithm was used to BLAST characterized *Trypanosoma brucei* tRNAs against our trimmed and dereplicated library. Potential candidates from the library were then further analyzed using the ARAGORN webserver [25] to look for conserved sequence and structural elements indicative of tRNAs.

Results and discussion

Preventing amplification of unwanted RNAs during RNA-Seq library construction

In *E. gracilis* the large subunit rRNA is naturally fragmented into 14 discrete pieces most of which fall within the size range of many other small ncRNA species. This complicates the generation of small RNA libraries in *E. gracilis* and other Euglenozoa since high levels of rRNA dominate the sequence reads from even a carefully size-selected library. To address this issue we have developed a strategy adapted from a technique previously described for eliminating unwanted DNA sequences from environmental samples [26]. This strategy utilizes blocking oligonucleotides which contain a hydrocarbon chain modification at their 3′ end during the PCR amplification step of cDNA library construction to prevent amplification of targeted sequences.

A set of blocker oligonucleotides was designed to anneal to the end of the added 5′ linker sequence + 5′ end of each of the 14 LSU rRNA fragments (**Fig. S1a**). To be effective, the blocker oligonucleotides must anneal efficiently to the unwanted target cDNAs while not significantly affecting the amplification of all other cDNAs. We found that 13 nt of complementarity with 5′ linker sequence + 13 nt of complementarity with the 5′ end of an LSU rRNA fragment worked effectively. We also found that having a 10 fold excess of each blocker oligonucleotide relative to adaptor-specific general amplification oligonucleotide greatly diminished the amplification of specific LSU sequences (Fig. 1 and **Fig. S1b**). Without employing blocker oligonucleotides, LSU fragments that are prevalent in a size-selected (< 400 nt) *E. gracilis* RNA fraction are efficiently amplified. This can be detected by including specific LSU reverse primers with the adapter-specific amplification primers during the PCR step of the procedure (Fig. 1, lanes 1 and 4). When an LSU fragment-specific blocker oligonucleotide is also included during the PCR step, the targeted LSU sequence is either greatly reduced or undetectable following PCR amplification (Fig. 1, lanes 2, 3, 5 and 6, and **Fig. S1b**). Importantly, this was also observed when multiple blocker oligonucleotides were used in the same PCR amplification (data not shown). Using this approach, the relative number of rRNA sequence reads in RNA-Seq data is greatly reduced.

The efficiency of blocking was further assessed, prior to library deep-sequencing, by shotgun cloning cDNA library products and sequencing 178 clones by Sanger sequencing of PCR products obtained via bacterial colony PCR. When no blocker oligonucleotides were used, 30% of the total unfiltered reads were rRNA (Fig. 2A). After removing reads that were poor quality (29%) and also those less than 20 nucleotides in length, the remaining sequence reads we termed 'informative sequences'. Of these informative sequences, 75% were rRNA fragments. Initially, blocker oligos had only been designed to the 10 smallest LSU species, whose mature sizes fall in the range of small ncRNAs. Following amplification including these blockers, rRNA still dominated the informative reads; however, most of the rRNA sequences were now fragments of the 4 largest LSU species (i.e. the blocking of smaller LSU species was

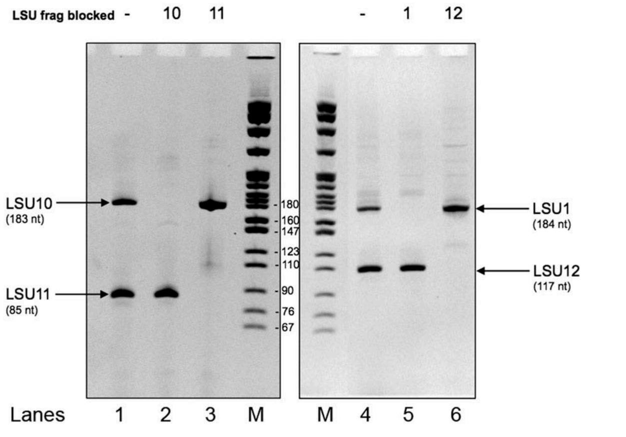
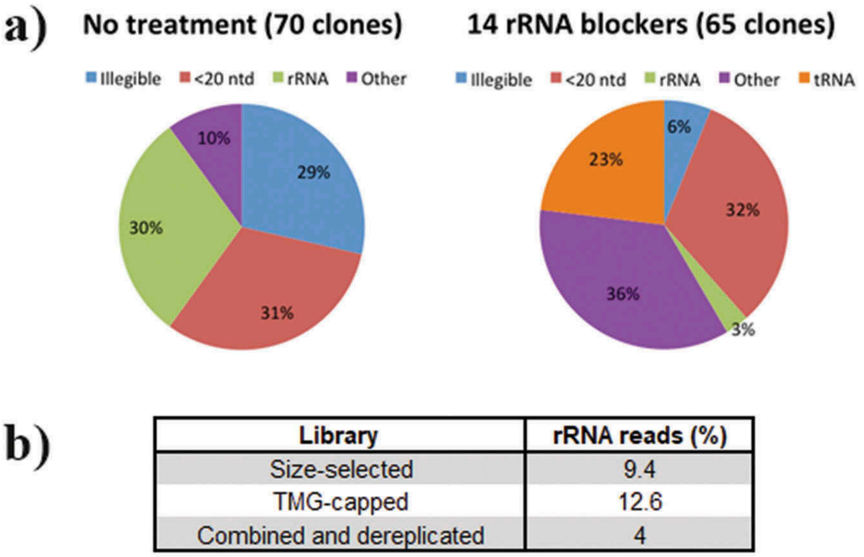


Figure 1. Primer blocking strategy to prevent rRNA amplification during small RNA library preparation. a) Forward oligonucleotides that anneal to LSU fragments and reverse oligonucleotides that anneal to the linker were used to amplify specific LSU fragments (LSU 1, 10, 11, and 12) from the library (lanes 1 and 4). Excess of blocker oligonucleotide specific to each fragment was added to assess blocking efficiency (lanes 2, 3, 5, and 6). PCR products were resolved on a 6% native polyacrylamide gel. M = pBR322 MspI digest.



Library	rRNA reads (%)
Size-selected	9.4
TMG-capped	12.6
Combined and dereplicated	4

Figure 2. Sequencing results of a *Euglena gracilis* small RNA library before and after use of primer blocking to prevent amplification of rRNA fragments. **a)** PCR-amplified library products were cloned and sequenced using Sanger sequencing to assess the efficacy of the primer blocking strategy. Sequence reads that were legible and longer than 20 nt were considered informative sequences. **b)** Proportion of reads attributed to rRNA for the size-selected, TMG-capped, and combined and dereplicated small RNA libraries generated using RNA-Seq.

successful). Additional blocker oligonucleotides were then designed such that all 14 LSU species were targeted for depletion during library construction to also reduce amplification of these LSU rRNA degradation productions. This was very effective, resulting in only 3% of all reads matching rRNA sequence and a significant enrichment of informative ‘other’ sequences (36%); that is, sequences that consist of ncRNAs other than rRNA and tRNA were now evident (Fig. 2A). This indicates that adding blocker oligonucleotides targeting all 14 *E. gracilis* LSU fragments at the PCR amplification step of library synthesis is very effective in reducing the number of rRNA sequences in the final library, especially when considering that in studies in other organisms, RNA-Seq data from total RNA samples with no rRNA depletion step typically contain > 90% rRNA reads [27–30].

Two different LSU rRNA-depleted *E. gracilis* small RNA libraries were created, a size-selected (< 400 nt) library and a TMG cap pull-down non-size-selected library, and both were individually sequenced using paired-end 250 bp sequencing on an Illumina MiSeq platform (Genome Québec). In total, following quality control filtering, there were 3,080,604 high-quality reads when combining the reads from the size-selected and TMG-cap pull-down libraries. In the size-selected library reads, 9.4% of reads were rRNA sequence, and 12.6% of the TMG-capped library reads were rRNA (Fig. 2B). Following dereplication of the combined library reads, there were 727,447 unique reads and searching for previously annotated *E. gracilis* RNAs revealed that approximately 4% of these reads were rRNA, 19% were snRNAs, 1.4% were known tRNAs (only one nuclear-encoded but a complete set of chloroplast tRNAs had been previously characterized in this organism) and < 1% were previously characterized snoRNA sequences. Cumulatively, this indicated that our rRNA depletion strategy was very successful and worked similarly when employed on the two independently processed library fractions (size-selected and TMG cap pull-down). It is also the first indication of the apparent diversity of ncRNAs in this organism.

The strategy described here is very useful for RNA-Seq experiments in any organism for which no commercial rRNA depletion kit is available and is also adaptable because theoretically any unwanted (or previously characterized) RNA species that would otherwise dominate the library reads can be depleted at this stage. This can serve as a useful tool for RNomics in less-studied species (such as protists) where there is currently a lack of information regarding the abundance and diversity of ncRNAs.

Identification of new snoRNAs

Previously identified modification-guide snoRNAs in *E. gracilis* displayed a relatively uniform size distribution, between 50 and 90 nt, so we focused on examining sequence reads in that size range. First we identified candidates by scanning for conserved sequence and structural features [15,16,19], and then requiring that candidates be able to base-pair to corresponding rRNA target modification sites [13]. This approach identified 82 new box C/D snoRNAs, 31 box AGA RNAs (**Figs. S2, S3, and S4**) and numerous isoforms of both types – we define isoforms as those sequence-related RNAs predicted to target the same modification site. Cumulatively, including all biochemically, genomically (PCR-mediated), and now RNA-Seq identified RNAs, we have characterized snoRNAs that guide modifications of approximately 88% of the 2′-*O*-methylated sites and 45% of pseudouridylated sites in *E. gracilis* rRNA [15,16,19], 227 unique snoRNA species in total.

In order to examine snoRNA representation in our data set we used single end reads from both the size-selected and TMG-capped libraries to calculate reads per million (RPM) for each newly identified snoRNA and all previously identified snoRNAs. We found that RPM for the newly identified snoRNAs were consistent with the range of RPM values found for previously identified snoRNAs in both our libraries (**Table S3**). Additionally, when comparing reads found in the two libraries we observed that RNAs from the size-selected library consisted of a variety of both mature and precursor forms while the TMG-capped reads were generally more uniform in size and had a higher proportion of reads representing mature RNAs. All but 4 methylation guide and 2 pseudouridylation guide RNAs were detected in the size-selected library while only 78% and 51% of each type respectively were found in the TMG-capped library. When considering the very large observed relative enrichment of the U3 snoRNA and U2 snRNA in the TMG-capped library (**Table S3**), two ncRNAs anticipated to have hypermethylated caps, the lack of significant enrichment (or even complete absence) in this library of most *E. gracilis* modification guide snoRNAs may suggest that a large fraction of these RNAs do not possess hypermethylated caps.

All of the 82 new C/D box snoRNAs identified by RNA-Seq appear to be single-guide RNAs and predominantly utilize the region upstream of the D′ box to target modification, similar properties to the previously characterized *E. gracilis* box C/D RNAs [15,16,19]. Double-guide RNAs are exceedingly rare in *E. gracilis* and of the 182 different box C/D RNA species now identified, only 2 appear to be double-guides and in both cases, each species utilizes its two guide regions to target nearby 2′-*O*-methyl sites. This is noticeably different from what is observed in other eukaryotes and even more so, in archaeal organisms where double-guide box C/D RNAs constitute a much higher fraction of the total modification-guide RNA repertoire. The *Euglena* snoRNAs identified so far are also more uniform and smaller in size than those in other characterized eukaryotes and show closer resemblance to their archaeal counterparts.

Importance of rRNA spacer regions and timing of snoRNA-guided rRNA modification

Two of the new snoRNAs identified by RNA-Seq have the required base-pairing potential to target the modification sites found at the 3′ extremities of rRNA species. The first, Eg-m121 guides 2′-*O*-methylation at position LSU3906, near the 3′ end of LSU species 12. Base-pairing interactions between the snoRNA and rRNA are typical of the length most commonly observed between box C/D RNAs and rRNA in *E. gracilis* (10 bp) and most interestingly, this base-pairing interaction extends into the (intergenic) spacer region that separates LSU species 12 and 13 on the primary rRNA transcript (Fig. 3A). The second, Eg-p38 RNA guides Ψ formation at position SSU2305, the penultimate nucleotide at the 3′ end of the SSU rRNA and the base-pairing interaction between the snoRNA and rRNA extends into ITS 1, the spacer between SSU rRNA and LSU species 1 (Fig. 3B). In fact, the entire interaction between the 5′ half of the snoRNA bi-partite base-pairing interaction to form the pseudouridine pocket with the rRNA target site occurs using only the ITS region. To our knowledge, these are the first examples of modification guide snoRNAs predicted to employ base-pairing interactions to mature rRNA-spacer sequence boundaries.

The way in which rRNA modifications coordinate with other maturation steps such as pre-rRNA cleavage is not well understood. Co-transcriptional modification has been observed in yeast prior to pre-rRNA cleavage [31], but it is unclear if this is a common feature among different eukaryotes. This is particularly interesting in the case of *E. gracilis* as the fragmentation of the LSU, along with the high degree of rRNA modification, results in significant enrichment of rRNA modifications near cleavage sites compared to other eukaryotes. For the two snoRNAs described above, the interaction between snoRNA and rRNA must occur before pre-rRNA cleavage that removes these particular spacer regions during the biogenesis pathway that generates the mature 3′ ends. It is commonly suggested that snoRNA-rRNA base-pairing interactions occur very early in the ribosome biogenesis pathway [14]. In *E. gracilis*, the LSU is highly modified and naturally fragmented into 14 pieces indicating a high degree of complexity in pre-rRNA processing and ribosome assembly. It is interesting to consider that these rRNA modifications and associated modification complexes may also play some role in removal of these spacer sequences to generate mature LSU rRNA fragments.

Identification and characterization of novel ψ-guide snoRNAs

In previous studies, only 12 *E. gracilis* Ψ-guide RNA species had been identified, all of which contained a conserved ‘AGA’ sequence 3 nt from their 3′ ends (i.e. AGA box RNAs) [15,16]. The small number identified was due to the inefficient immunoprecipitation of these RNA-containing complexes when using antibodies targeted at the protein Cbf5p, the more challenging nature of identifying encoding regions for these structurally more complex RNAs compared to identifying box C/D RNAs within PCR-amplified *Euglena* genomic snoRNA cluster regions, and the fact that genomic amplification primers were primarily based on identified box C/D RNA sequences that presumably favor amplification of clusters encoding primarily box C/D RNAs. Bioinformatic analysis of the small ncRNA library has now significantly increased those identified to 45 Ψ-guide snoRNA species with predicted rRNA target sites. Even though we allowed for much greater size variation when searching the library, the size range of these RNAs is 60 – 72 nt, with an average size of 66 nt, a remarkably uniform size distribution compared to Ψ-guide RNAs characterized in other eukaryotes. It has previously been observed that the interaction between the target rRNA and the snoRNA is responsible for positioning the target substrate uridine (and Ψ pocket) 13 – 16 nt from either an ACA or H box sequence for the typical two-stem Ψ-guide RNA structure characterized in other eukaryotes [32,33]. The *Euglena* Ψ-guide RNAs share this property (distance from their AGA box) with the exceptions of Eg-p7, Eg-p30 and Eg-p37 (17–18 nt, **Table S4**). It is currently uncertain whether structural differences in *Euglena* Ψ-guide snoRNP structure may allow for such variation while still efficiently targeting a

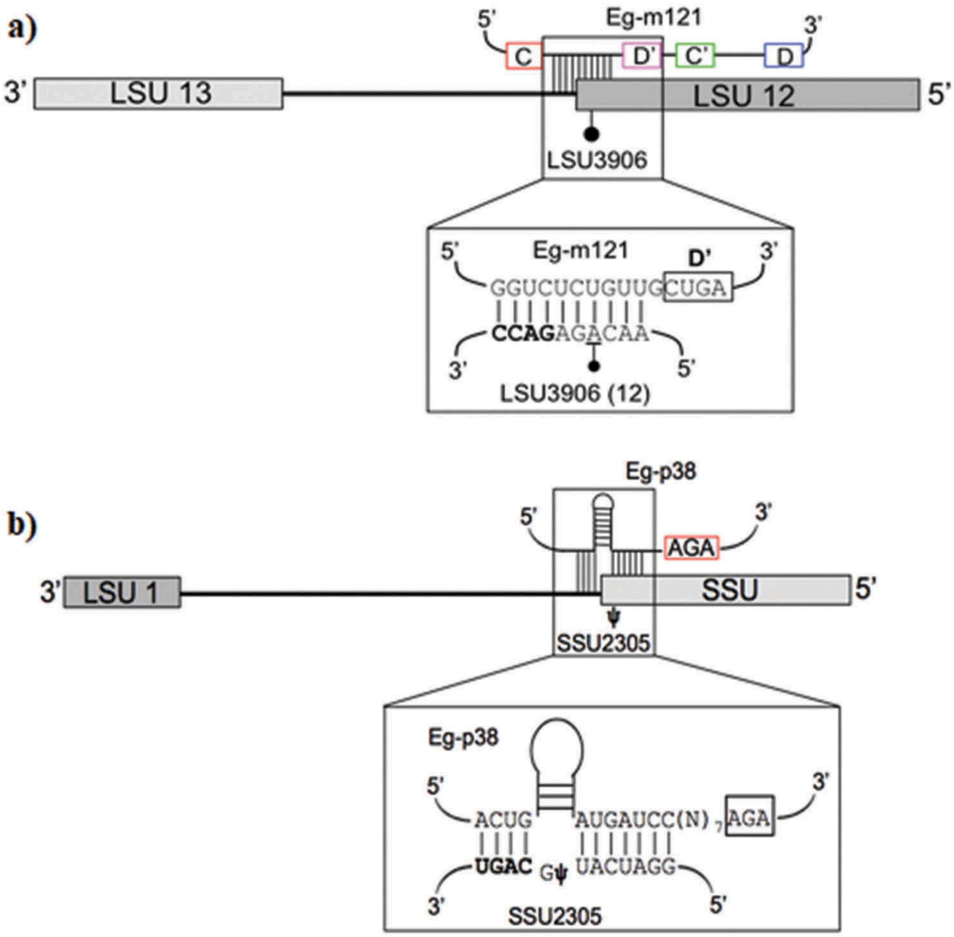


Figure 3. **Identified *E. gracilis* snoRNAs whose guide regions base-pair with pre-rRNA intergenic sequence**. Two snoRNAs were identified that each guide a modification site found at 3′ extremities of rRNA subunit species with their guide regions base-pairing to intergenic sequence. **a)** Eg-m121 snoRNA guides a 2′-*O*-methylation at position A3906 which is located 3 nucleotides from the mature 3′ end of LSU fragment 12. The snoRNA guide region pairs with 4 nucleotides of the spacer region (inset, bolded nucleotides) between LSU fragments 12 and 13. **b)** Eg-p38 snoRNA guides a Ψ modification at position U2305 which is located 2 nucleotides from the mature 3′ end of the SSU rRNA. The entire 5′ portion of the pseudouridine pocket (relative to the snoRNA) is formed by base-pairing to the internal transcribed spacer 1 region (bolded nucleotides), the pre-rRNA region located between the 3′ end of the SSU and the 5′ end of the 5.8S rRNA (LSU 1).

modification site since we cannot definitively rule out the possibility that other isoforms of these RNAs may exist with more optimal distances that weren't amplified or detected in our library.

The increased number of characterized Ψ-guide snoRNAs now allows for a more thorough examination of the common structural features of these RNAs (see **Table S4 and S5**). When considering all *E. gracilis* extended 'single-stem' Ψ-guide snoRNAs discovered to date, the basal stems vary from 4 – 9 bp in length (6 bp median) and the majority have predicted canonical base-pairing interactions in this region. There also appears to be a preference for higher G-C content in the basal stem (4 G-C bp median), with only a single RNA possessing a basal stem containing < 50% G-C content. There are only 7 instances where this stem appears to be interrupted by bulged nucleotides. The average size of the more variable apical stem is 12 bp (range of 7 – 16 bp) and the majority of the identified snoRNAs have at least one mismatch or bulged nucleotide in this region. When the nucleotide changes observed in the various identified isoforms of a Ψ-guide RNA species are mapped onto its predicted RNA secondary structure, sequence variation is common in the apical stem and predicted to affect secondary structure (Fig. 4 and S6). Much less frequently do sequence changes occur that alter the structure of the basal stem. This indicates that sequence and structural variability in the apical stem is likely accommodated in *Euglena* Ψ-guide snoRNP complexes. This is consistent with what has been observed in yeast, where basal stems are essential for snoRNA accumulation while the apical stems do not contribute as significantly to snoRNA stability [34,35]. Both stems are however essential for the pseudouridylation reaction [35]. Nucleotide substitutions between isoforms in *Euglena* are also very commonly found in the apical loop region (Fig. 4) seemingly indicating a lack of strict structural/sequence motifs for snoRNA stability and possibly functionality, in those regions.

In the collection of *Euglena* Ψ-guide snoRNA sequences, 50% contain a uridine immediately downstream of the AGA box ('AGA**U**NN') and 32% of the RNAs have the 3′ end sequence 'AGAUGN' (**Fig. S4**). We therefore define a new consensus motif for these RNAs and refer to them as AGAUGN box Ψ-guide RNAs. It is also noteworthy that this sequence resembles the internal portion of the consensus

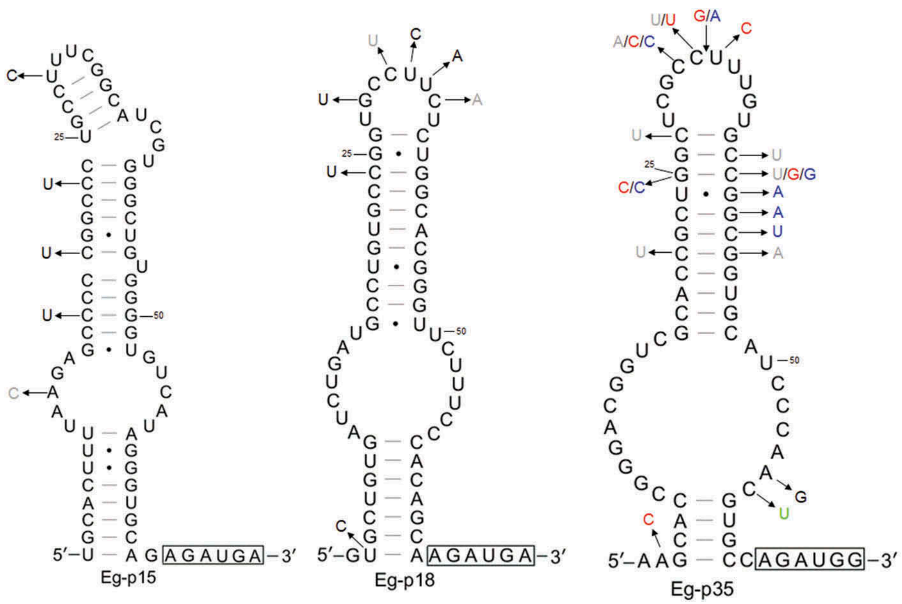


Figure 4. **Examples of predicted secondary structures of *E. gracilis* AGAUGN box snoRNA species and isoforms**. The boxed nucleotides highlight the AGAUGN box consensus sequence element and arrows indicate sequence variation between characterized isoforms. Nucleotide changes most frequently occur in the stem above the internal single-stranded loop regions that form the pseudouridylation pocket and/or in the apical loop region. Each different color represents the nucleotide changes present in a single isoform species. Black = Eg-p#.1; Grey = Eg-p#.2; Red = Eg-p#.3; Blue = Eg-p#.4; Green = Eg-p#.5. See **Fig. S6** for additional examples.

C and C′ box sequences (UGAUGA) of the methylation guide box C/D snoRNAs.

Additional programs such as snoSeeker [36], snoGPS [37], and Psiscan [38] have been developed to search for snoRNAs in sequence data libraries. However, they are either based on conserved yeast or mammalian snoRNA features, many of which are not present in *E. gracilis* snoRNAs, or depend on a computational pipeline based on trypanosome sequence analysis and comparison. Some of these programs were tested for their ability to successfully find already characterized *Euglena* snoRNA sequences in the library sequence data, but were largely unsuccessful. Therefore, we found that manually inspecting the initial hits provided through pattern search parameters (see methods) generated the best overall results. This strategy is based on the features of the previously identified snoRNAs in *E. gracilis* and hence it is possible that the remaining 'missing' snoRNAs may be significantly structurally different from those already characterized. Perhaps a set of *Euglena* Ψ-guide RNAs exist which contain multiple extended stem-loop structures. However, as no double-stem H/ACA-like snoRNAs were identified using the software listed above which are optimized to search for such structures, any *Euglena* RNAs of this variety would have to be significantly structurally divergent from other eukaryotic (yeast and human) H/ACA box snoRNAs on which the programs were trained.

Single stem Ψ-guide snoRNAs possessing conserved AGA sequences have also been identified in trypanosomes, a group of organisms within the Euglenozoans. Analysis of *Leishmania major* [18] and *Trypanosoma brucei* [17] Ψ-guide snoRNAs reveals significant structural similarity to the RNAs identified in our study. Some AGA box Ψ-guide RNAs in trypanosomes exceeding 100 nucleotides in length have recently been identified but these RNAs appear to still maintain the extended single stem secondary structure [18]. The genome-wide search for snoRNAs in these trypanosomes found rRNA modification sites with no apparent corresponding snoRNA guide. It has been proposed that these modifications might be carried out by protein-only enzymes [17,18]. This may also be the case in *E. gracilis* as there still remains a collection of rRNA modifications for which no snoRNA guide has been identified. The unusual rRNA processing pathway could require the cooperative action of both stand-alone enzymes and snoRNA-guided modification complexes.

Evolution of ψ-guide snoRNA species in *Euglena*

During the analysis of the newly identified AGAUGN box snoRNAs we found a few sequence isoform groups in which the isoform members displayed more sequence variation than is typical for the isoforms of a single Ψ-guide species. The first pair, Eg-p32 and Eg-p33 were initially clustered in our bioinformatics analysis as a single snoRNA species; however, closer inspection revealed that these two RNAs contain significant changes in their Ψ-guide pocket regions resulting in the targeting of two different Ψ modification sites (Fig. 5A,B). Similarly another pair, Eg-p36 and Eg-p42, possess a high degree of sequence similarity to each other but contain significant changes in their guide regions (Fig. 5A,B). The predicted structural differences that are evident when comparing pairs such as Eg-p32 and Eg-p33 (Fig. 5C) further illustrates the previously described properties of this class of *Euglena* ncRNA. This includes the lengthening of the basal stem in Eg-p33 without creating bulges or mismatches, apparent tolerance of a few different mismatches/bulges in the apical stem

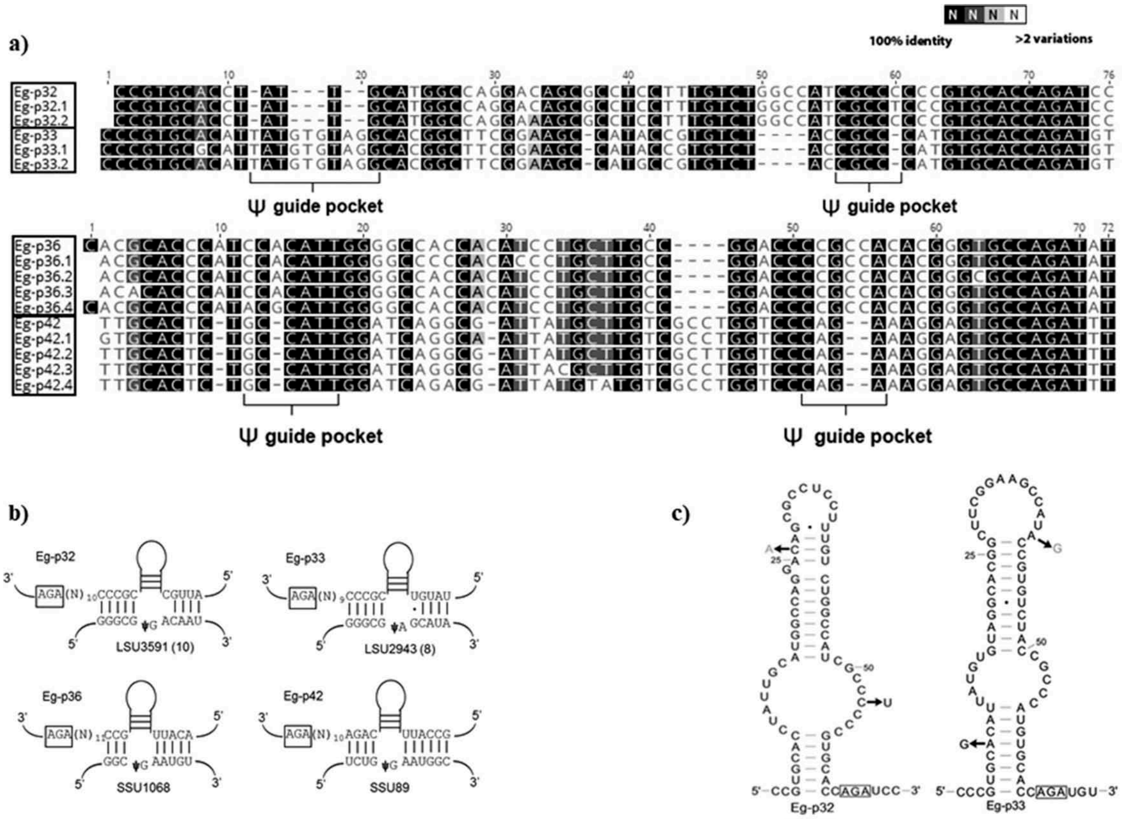

a)
N N N N
100% identity
>2 variations
Eg-p32 CCGTGCACCT-AT---T--GCATGGCCAGGACAGCGCCTCCTTTGTCTGGCCATCGCCCCCCGTGCACCAGATCC
Eg-p32.1 CCGTGCACCT-AT---T--GCATGGCCAGGACAGCGCCTCCTTTGTCTGGCCATCGCCTCCCGTGCACCAGATCC
Eg-p32.2 CCGTGCACCT-AT---T--GCATGGCCAGGAAAGCGCCTCCTTTGTCTGGCCATCGCCCCCCGTGCACCAGATCC
Eg-p33 CCCGTGCACATTATGTGTAGGCACGGCTTCGGAAGC-CATACCGTGTCT----ACCGCC-CATGTGCACCAGATGT
Eg-p33.1 CCCGTGCGCATTATGTGTAGGCACGGCTTCGGAAGC-CATACCGTGTCT----ACCGCC-CATGTGCACCAGATGT
Eg-p33.2 CCCGTGCACATTATGTGTAGGCACGGCTTCGGAAGC-CATGCCGTGTCT----ACCGCC-CATGTGCACCAGATGT
Ψ guide pocket
Ψ guide pocket
Eg-p36 CACGCACCCATCCACATTGGGGCCACCACATCCTGCTTGCC----GGACCCCGCCACACGGGTGCCAGATAT
Eg-p36.1 ACGCACCCATCCACATTGGGGCCCCCACACCCTGCTTGCC----GGACCCCGCCACACGGGTGCCAGATAT
Eg-p36.2 ACGCACCCATCCACATTGGGGCCACCACATCCTGCTTGCC----GGACCCCGCCACACGGGCGCCAGATAT
Eg-p36.3 ACACACCCATCCACATTGGGGCCACCACATCCTGCTTGCC----GGACCCCGCCACACGGGTGCCAGATAT
Eg-p36.4 CACGCACCCATACGCATTGGGGCCACCACATCCTGCTTGCC----GGACCCCGCCACACGGGTGCCAGATAT
Eg-p42 TTGCACTC-TGC-CATTGGATCAGGCG-ATTATGCTTGTCGCCTGGTCCCAG--AAAGGAGTGCCAGATTT
Eg-p42.1 GTGCACTC-TGC-CATTGGATCAGGCA-ATTATGCTTGTCGCCTGGTCCCAG--AAAGGAGTGCCAGATTT
Eg-p42.2 TTGCACTC-TGC-CATTGGATCAGGCG-ATTATGCTTGTCGCTTGGTCCCAG--AAAGGAGTGCCAGATTT
Eg-p42.3 TTGCACTC-TGC-CATTGGATCAGGCG-ATTACGCTTGTCGCCTGGTCCCAG--AAAGGAGTGCCAGATTT
Eg-p42.4 TTGCACTC-TGC-CATTGGATCAGACG-ATTATGTATGTCGCCTGGTCCCAG--AAAGGAGTGCCAGATTT
Ψ guide pocket
Ψ guide pocket
b)
Eg-p32
3' AGA (N) 10 CCCGC CGUUA 5'
5' GGGCG ΨG ACAAU 3'
LSU3591 (10)
Eg-p33
3' AGA (N) 9 CCCGC UGUAU 5'
5' GGGCG ΨA GCAUA 3'
LSU2943 (8)
Eg-p36
3' AGA (N) 12 CCG UUACA 5'
5' GGC ΨG AAUGU 3'
SSU1068
Eg-p42
3' AGA (N) 10 AGAC UUACCG 5'
5' UCUG ΨG AAUGGC 3'
SSU89
c)
25
50
5'-CCG CCAGAUCC-3'
Eg-p32
25
50
5'-CCCG CCAGAUGU-3'
Eg-p33

Figure 5. Evolution of *E. gracilis* Ψ-guide snoRNA species. a) Alignment of library cDNA sequences of isoforms of Ψ-guide snoRNAs. Two different sequence isoform clusters were found to have extensive sequence similarity (isoforms of Eg-p32 are similar to Eg-p33, Eg-p36 to Eg-p42) yet display sequence divergence in the regions that form the pseudouridylation pocket thus targeting different rRNA modification sites. Regions containing the nucleotides corresponding to the pseudouridylation guide pocket for each of the RNAs are labeled. Letters on a black background indicate identical nucleotides present at that position in 100% of isoforms, dark gray background indicate 1 isoform differs, light gray indicates 2 isoforms differ, and white background > 2 isoforms differ at that position **b)** Illustration of the predicted base-pairing interactions between the snoRNAs (top) and target rRNA pseudouridylated sites (bottom). Experimentally confirmed pseudouridine sites are indicated as 'Ψ'. Within the AGAUGN box elements the highly conserved AGA sequence is highlighted and the number of nucleotides (N) to the base-paired region is indicated. LSU = large subunit rRNA, SSU = small subunit rRNA and the *E. gracilis* LSU 'fragment' species where the modification site resides is indicated in parentheses. Full-length snoRNA sequences are shown in **Fig. S4. c)** Predicted secondary structures of the Eg-p32 and Eg-p33 pair with nucleotide changes mapped. Black nucleotides = Eg-p#.1; Grey nucleotides = Eg-p#.2.

of isoforms of Eg-p32 and significant sequence and length variation in the apical loops of the pairs. In both cases, these Ψ-guide snoRNA pairs are evolutionarily-related and the encoding genes have likely evolved by a mechanism similar to that previously characterized for box C/D snoRNA and rRNA target site evolution [15]. *Euglena* snoRNA genes are often found within tandemly repeated clusters containing multiple unique snoRNA isoforms of one or both classes [15,16]. snoRNA gene duplication followed by sequence divergence within modification guide regions allows the targeting and emergence of new modification sites [15]. This mechanism of snoRNA evolution has also been observed in plants [39,40], nematodes [41], and trypanosomes [18]. When comparing Ψ-guide RNA homologs between different trypanosome species, Eliaz *et al*. observed both snoRNA gene duplication and sequence divergence in the pseudouridylation pockets of the RNAs that allows the targeting of different rRNA nucleotides, similar to what we observe with the *Euglena* Ψ-guide snoRNA paralogs (see above).

Multiple isoforms are present for many of the newly identified *E. gracilis* snoRNA species of either class. Up to 9 isoforms were found for a single Ψ-guide RNA species (**Fig. S4**). This further highlights the high frequency of snoRNA gene duplication and abundance of this class of RNA in *Euglena*. The apparent frequency of snoRNA duplication events is likely the mechanism that allows for the extensive RNA modification in this organism and may be an adaptation to allow for (or even cause) the complex ribosomal biogenesis pathway; both extensive fragmentation and modification.

Orphan snoRNAs in Euglena gracilis

In addition to snoRNAs predicted to target rRNA, numerous potential orphan 'snoRNAs' were identified (**Fig. S5**). These RNAs are similar in size to rRNA targeting snoRNAs, have canonical sequence box elements and in the case of AGAUGN box RNAs, contain the conserved secondary structural features. However, no mapped modified nucleotide is present in the rRNA in regions that show any limited base-pairing potential to the appropriate regions of these RNAs. Of the newly identified AGAUGN box RNAs found in our library, 8 (20%) were orphans. For box C/D class snoRNAs, 17 (17%) of the newly identified RNAs were considered orphans. In total, orphan 'snoRNAs' represent 11% of the combined total of all snoRNA species identified so far in *Euglena*. These orphan

RNAs also do not appear to be involved in modifying snRNAs based on examining any base-pairing potential to the mapped *Euglena* snRNA modification sites or other snRNA regions not yet experimentally examined for modifications. Previous analysis of trypanosome snoRNAs also failed to identify any guides for snRNA modifications, with the exception of the spliced-leader RNA [18]. The apparent scarcity of snRNA modification-guide RNAs in Euglenozoa may indicate that these modifications are performed by stand-alone protein enzymes or guided by structurally novel RNAs.

While we do not know the function of these orphan *Euglena* RNAs, the discovery of pseudouridylation of mRNA species in yeast and humans [42–44] raises the possibility that these orphans may be used to target modification sites in mRNA or other cellular RNA species. Alternatively, they could be involved in *Euglena*'s unique rRNA processing/ assembly pathway, as is suggested as a possible function in some trypanosomes, or be processed into other types of ncRNA [18]. Our method for identification of box C/D snoRNAs relied primarily on the presence of the consensus sequence box elements at expected positions relative to RNA 5′ and 3′ ends and searches for significant base-pairing interactions to modified sites in *E. gracilis* rRNA. Comprehensive identification of all potential box C/D RNAs in the library sequences is particularly challenging for orphans (without rRNA base-pairing) because of the short length of C and D box elements, the even more sequence degenerate C′ and D′ boxes, and the absence of any obvious extended secondary structural conservation in this class of *Euglena* snoRNA. Consequently, we are likely underestimating the abundance of 'snoRNAs' that target modification or perform other functions on non-rRNA species. The conserved 'AGAUGN' box and secondary structural features makes it somewhat easier to identify orphan RNAs within this other 'snoRNA' class.

tRNA identification

As a result of sparse genomic sequence information very few *E. gracilis* nuclear tRNAs have been identified to date. Using the tRNA identification program ARAGORN [25] to detect library sequences with requisite conserved primary and secondary structural potential, we have identified 14 tRNAs in our library which do not map onto the complete *E. gracilis* chloroplast genome sequence [45]. Analysis of the *E. gracilis* mitochondrial genome failed to identify any tRNA coding genes, suggesting that these 14 tRNAs are encoded in the nuclear genome. We cannot rule out that some of these tRNAs may function in the mitochondria [46]. These tRNAs include single tRNA isoacceptors for Met (CAU), His (GUG), Ser (AGA), Cys (GCA), Leu (CAG), and multiple isoacceptors for Glu (CUC and UUC), Gln (UUG and CUG), Pro (UGG and CGG), and Ala (AGC, CGC, and UGC) (**Fig. S7**). While some of these identified sequences possess mature CCA 3′ ends, a large collection of the tRNA reads are precursors which contain additional sequence at their 5′ and 3′ ends. An abundance of nucleoside modifications in *E. gracilis* tRNAs likely explains the higher representation of precursors and the incomplete representation of all tRNA species in the library as many tRNA modifications block reverse transcriptases from extending beyond the modified site, requiring additional enzymatic treatments of the RNA prior to RT in order to obtain full-length cDNAs [47]. The limitation of sequence size for Illumina sequencing may also explain the absence of tRNA precursors containing introns from being identified in the library. Correct identification of the new *Euglena* tRNAs was further confirmed through alignment to the tRNA isoacceptors in *Trypanosoma* and *Leshmania* species. This showed close sequence similarity of tRNAs between all these species.

Disclosure of Potential Conflicts of Interest

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Discovery Grant funding to AGR from the Natural Sciences and Engineering Research Council of Canada (NSERC), Grant 355757-2013-RGPIN. ANM and AJH were supported by NSERC Canada Graduate Scholarship-Doctoral, and DCM was partly supported by a U of L. Tuition Award and Natural Sciences and Engineering Research Council of Canada[355757-2013-RGPIN];

ORCID

Andrew J. Hudson http://orcid.org/0000-0001-8097-2325

References

[1] Huttenhofer A, Schattner P, Polacek N. Non-coding RNAs: hope or hype? Trends Genet. 2005 May;21(5):289–297. [pii]. PubMed PMID: 15851066; eng.

[2] Mattick JS, Makunin IV. Non-coding RNA. Hum Mol Genet. 2006 Apr 15;15 Spec No 1:R17–29. [pii]. PubMed PMID: 16651366; eng.

[3] Matera AG, Terns RM, Terns MP. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nat Rev Mol Cell Biol. 2007 Mar;8(3):209–220. [pii]. PubMed PMID: 17318225; eng.

[4] Goodrich JA, Kugel JF. From bacteria to humans, chromatin to elongation, and activation to repression: the expanding roles of noncoding RNAs in regulating transcription. Crit Rev Biochem Mol Biol. 2009 Jan-Feb;44(1):3–15. [pii]. PubMed PMID: 19107624; eng.

[5] Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet. 2009 Mar;10(3):155–159. [pii]. PubMed PMID: 19188922; eng.

[6] Faust T, Frankel A, D'Orso I. Transcription control by long non-coding RNAs. Transcription. 2012 Mar-Apr;3(2):78–86. [pii]. PubMed PMID: 22414755; eng.

[7] McWatters DC, Russell AG. *Euglena* transcript processing. In: Schwartzback SD, Shigeoka S, editors. Euglena: Biochemistry, Cell and Molecular Biology. Vol. 979, Advances in Experimental Medicine and Biology. Berlin: Springer International Publishing; 2017. p. 141–158.

[8] Ebenezer TE, Carrington M, Lebert M, et al. *Euglena gracilis* genome and transcriptome: organelles, nuclear genome assembly strategies and initial features. In: Schwartzback SD, Shigeoka S, editors. Euglena: Biochemistry, Cell and Molecular Biology. Vol. 979, Advances in Experimental Medicine and Biology. Berlin: Springer International Publishing; 2017. p. 159–182.

[9] Yoshida Y, Tomiyama T, Maruta T, et al. *De nono* assembly and comparative transcriptome analysis of *Euglena gracilis* in response to anaerobic conditions. BMC Genomics. 2016;17(182):1–10.

[10] O'Neill EC, Trick M, Hill L, et al. The transcriptome of *Euglena gracilis* reveals unexpected metabolic capabilities for carbohydrate and natural product biochemistry. Mol Biosyst. 2015;11:2808–2820.

[11] Schnare MN, Cook JR, Gray MW. Fourteen internal transcribed spacers in the circular ribosomal DNA of Euglena gracilis. J Mol Biol. 1990 Sep 5;215(1):85–91. [pii]. PubMed PMID: 2118961; eng.

[12] Schnare MN, Gray MW. Sixteen discrete RNA components in the cytoplasmic ribosome of Euglena gracilis. J Mol Biol. 1990 Sep 5;215(1):73–83. [pii]. PubMed PMID: 2118960; eng.

[13] Schnare MN, Gray MW. 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. 2011 Oct 14;413 (1):66–83. [pii]. PubMed PMID: 21875598; eng.

[14] Watkins NJ, Bohnsack MT. The box C/D and H/ACA snoRNPs: key players in the modification, processing and the dynamic folding of ribosomal RNA. Wiley Interdiscip Rev RNA. 2012 May-Jun;3(3):397–414. PubMed PMID: 22065625; eng.

[15] Moore AN, Russell AG. Clustered organization, polycistronic transcription, and evolution of modification-guide snoRNA genes in Euglena gracilis. Mol Genet Genomics. 2012 Jan;287 (1):55–66. PubMed PMID: 22134850; eng.

[16] Russell AG, Schnare MN, Gray MW. Pseudouridine-guide RNAs and other Cbf5p-associated RNAs in Euglena gracilis. RNA. 2004 Jul;10(7):1034–1046. [pii]. PubMed PMID: 15208440; eng.

[17] Liang XH, Uliel S, Hury A, et al. 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. 2005 May;11(5):619–645. [pii]. PubMed PMID: 15840815; eng.

[18] Eliaz D, Doniger T, Tkacz ID, et al. 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. 2015 May 13;12:1222–1255. PubMed PMID: 25970223; Eng.

[19] Russell AG, Schnare MN, Gray MW. A large collection of compact box C/D snoRNAs and their isoforms in Euglena gracilis: structural, functional and evolutionary insights. J Mol Biol. 2006 Apr 14;357(5):1548–1565. [pii]. PubMed PMID: 16497322; eng.

[20] Sambrook J, Russell DW. Molecular cloning: A laboratory manual. 3rd ed. New York: Cold Spring Harbor Laboratory Press; 2001.

[21] Rederstorff M, Huttenhofer A. cDNA library generation from ribonucleoprotein particles. Nat Protoc. 2011 Feb;6(2):166–174. [pii]. PubMed PMID: 21293458; eng.

[22] Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010 Oct 1;26(19):2460–2461. [pii]. PubMed PMID: 20709691; eng.

[23] Dsouza M, Larsen N, Overbeek R. Searching for patterns in genomic data. Trends Genet. 1997 Dec;13(12):497–498. [pii]. PubMed PMID: 9433140; eng.

[24] Lowe TM, Eddy SR. A computational screen for methylation guide snoRNAs in yeast. Science. 1999 Feb 19;283(5405):1168–1171. PubMed PMID: 10024243; eng.

[25] Laslett D, Bjorn C. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 2004;32(1):11–16.

[26] Vestheim H, Jarman SN. Blocking primers to enhance PCR amplification of rare sequences in mixed samples - a case study on prey DNA in Antarctic krill stomachs. Front Zool. 2008;5:12. [pii]. PubMed PMID: 18638418; eng.

[27] He S, Wurtzel O, Singh K, et al. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. Nat Methods. 2010 Oct;7(10):807–812. [pii]. PubMed PMID: 20852648; eng.

[28] Chen Z, Duan X. Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. Methods Mol Biol. 2011;733:93–103. PubMed PMID: 21431765; eng.

[29] O'Neil D, Glowatz H, Schlumpberger M. Ribosomal RNA depletion for efficient use of RNA-seq capacity. Curr Protoc Mol Biol. 2013 Jul;Chapter 4:Unit4 19. PubMed PMID: 23821444; eng.

[30] Peano C, Pietrelli A, Consolandi C, et al. An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. Microb Inform Exp. 2013;3(1):1. [pii]. PubMed PMID: 23294941; eng.

[31] Kos M, Tollervey D. Yeast pre-rRNA processing and modification occur cotranscriptionally. Mol Cell. 2010 Mar 26;37(6):809–820. [pii]. PubMed PMID: 20347423; eng.

[32] Ganot P, Bortolin ML, Kiss T. Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. Cell. 1997 May 30;89(5):799–809. [pii]. PubMed PMID: 9182768; eng.

[33] Ni J, Tien AL, Fournier MJ. Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA. Cell. 1997 May 16;89(4):565–573. [pii]. PubMed PMID: 9160748; eng.

[34] Balakin AG, Smith L, Fournier MJ. The RNA world of the nucleolus: two major families of small RNAs defined by different box elements with related functions. Cell. 1996 Sep 6;86(5):823–834. [pii]. PubMed PMID: 8797828; eng.

[35] Bortolin ML, Ganot P, Kiss T. Elements essential for accumulation and function of small nucleolar RNAs directing site-specific pseudouridylation of ribosomal RNAs. EMBO J. 1999 Jan 15;18 (2):457–469. PubMed PMID: 9889201; eng.

[36] Yang JH, Zhang XC, Huang ZP, et al. snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. Nucleic Acids Res. 2006;34(18):5112–5123. [pii].

[37] Schattner P, Decatur WA, Davis CA, et al. Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the Saccharomyces cerevisiae genome. Nucleic Acids Res. 2004;32 (14):4281–4296. [pii].

[38] Myslyuk I, Doniger T, Horesh Y, et al. Psiscan: a computational approach to identify H/ACA-like and AGA-like non-coding RNA in trypanosomatid genomes. BMC Bioinformatics. 2008;9:471. [pii]. PubMed PMID: 18986541; eng.

[39] Barneche F, Gaspin C, Guyot R, et al. 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. 2001 Aug 3;311(1):57–73. [pii]. PubMed PMID: 11469857; eng.

[40] Brown JW, Clark GP, Leader DJ, et al. Multiple snoRNA gene clusters from Arabidopsis. RNA. 2001 Dec;7(12):1817–1832. PubMed PMID: 11780637; eng.

[41] Zemann A, op de Bekke A, Kiefmann M, et al. Evolution of small nucleolar RNAs in nematodes. Nucleic Acids Res. 2006;34 (9):2676–2685. [pii].

[42] Carlile TM, Rojas-Duran MF, Zinshteyn B, et al. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature. 2014 Nov 6;515(7525):143–146. [pii]. PubMed PMID: 25192136; eng.

[43] Lovejoy AF, Riordan DP, Brown PO. Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in S. cerevisiae. PLoS One. 2014;9(10):e110799. [pii].

[44] Schwartz S, Bernstein DA, Mumbach MR, et al. Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. Cell. 2014 Sep 25;159(1):148–162. [pii]. PubMed PMID: 25219674; eng.

[45] Hallick RB, Hong L, Drager RG, et al. Complete sequence of *Euglena gracilis* chloroplast DNA. Nucleic Acids Res. 1993;21 (15):3537–3544. PubMed Central PMCID: PMCPMC331456.

[46] Dobáková E, Flegontov P, Skalický T, et al. Unexpectedly streamlined mitochondrial genome of the Euglenozoan *Euglena gracilis*. Genome Biol Evol. 2015;7(12):3358–3367. PubMed Central PMCID: PMC26590215.

[47] Wilusz JE. Removing roadblocks to deep sequencing of modified RNAs. Nat Methods. 2015;12(9):821–822.