CHARACTERIZATION OF SMALL NUCLEOLAR RNAS IN THE PROTIST ORGANISM *EUGLENA GRACILIS*

ASHLEY NICOLE MOORE
B.Sc., University of Lethbridge, 2009

A Thesis
Submitted to the School of Graduate Studies
of the University of Lethbridge
in Partial Fulfilment of the
Requirements for the Degree

DOCTOR OF PHILOSOPHY

IN

BIOMOLECULAR SCIENCE

Biological Sciences
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

PREPARATION OF THESIS

ASHLEY NICOLE MOORE

Date of Defence: June 30, 2015

Dr. Anthony (Tony) Russell Supervisor	Assistant Professor	Ph.D.
Dr. Elizabeth Schultz Thesis Examination Committee Member	Associate Professor	Ph.D.
Dr. James Thomas Thesis Examination Committee Member	Professor	Ph.D.
Dr. Stacey Wetmore Internal Examiner	Professor	Ph.D.
Dr. George Owttrim External Examiner University of Alberta Edmonton, Alberta	Professor	Ph.D.
Dr. Theresa Burg Chair, Thesis Examination Committee	Associate Professor	Ph.D.

Dedication

This thesis is dedicated to my father Will Marchuk – the first scientist to inspire me.

Abstract

Non-coding RNAs (ncRNAs) have diverse cellular roles in all three domains of life. One class of ncRNAs termed small nucleolar RNAs (snoRNAs) play a role in RNA modification and processing in archaea and eukaryotic organisms. ncRNAs (including snoRNAs) remain largely unexplored in a group of unicellular eukaryotes known as protists. The focus of this study was to characterize snoRNAs in the protist organism *Euglena gracilis* in terms of their genomic arrangement, expression and evolution using experimental and computational methods. Numerous novel snoRNAs were characterized, many of which reside in tandemly repeated clusters that can be transcribed polycistronically. Some snoRNA gene clusters are surprisingly large, possibly the largest characterized to date. A mechanism of snoRNA evolution to support the unusually large number of snoRNAs and clustered modification sites in *E. gracilis* was characterized. These findings exemplify snoRNA diversity and highlight the importance of ncRNA characterization in a broad range of organisms.

Acknowledgements

First, I would like to acknowledge my supervisor Dr. Tony Russell for providing me with the opportunity to work in his lab. From the beginning when I was an undergraduate summer student and the lab was first taking off, to where we are today, the learning experiences you provided me with have been invaluable. Thank you for your encouragement, advice, and guidance over the past 7 years.

Next I would like to thank my Ph.D. committee members Drs. Elizabeth Schultz, James Thomas and H.J. Wieden for always providing me with valuable feedback and pushing me to succeed. Special thanks to my external thesis examination members Drs. George Owttrim and Stacey Wetmore and my defence chair Dr. Theresa Burg for taking the time to be a part of my defence committee.

To my lab mates Andy Hudson and David McWatters – I honestly never would have made it through grad school without you guys. Thank you for celebrating the successes with me and encouraging me through those seemingly impossible days (we all know I had my share of them…). This experience would not have been the same without your puns, our game nights, and all of our super-nerdy conversations. Your support and guidance helped me get to this point and I appreciate you both more than you know. Thanks to all of the past undergraduate students who were a part of the Russell Lab for your assistance. I would like to especially thank Ashlee Matkin and Kenzie Visser who helped progress my research project. Thanks also to the members of ARRTI for your willingness to provide assistance in anyway you could.

Next, I want to acknowledge my family and friends, without whom none of my success would be possible. To my husband Dustin – you are my rock and without your support, love, and positivity I never would have had the strength to pursue a Ph.D. I love

you. To my parents Bonnie and Will, you have always encouraged me to follow my dreams. Your confidence in me helped me through ALL these years of school and I can't thank you enough for being the amazing parents you are. To Kelsie, Kim and Tina – you three are the meaning of true, lifelong friends and I can't imagine my life without you in it. To the rest of my family and friends, thank you all so much for your love and support, you mean the world to me.

Lastly, I would like to acknowledge the National Sciences and Engineering Research Council of Canada (NSERC) for providing me with financial support through the duration of my graduate career.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Ψ	Pseudouridine
ATP	Adenosine 5′ triphosphate
BAC	Bacterial artificial chromosome
bp	Base pair
BLAST	Basic logical alignment search tool
BLASTN	Nucleotide BLAST
BSA	Bovine serum albumin
C3	3 hydrocarbon oligonucleotide modification
cDNA	Complementary DNA
cfu	Colony forming unit
Ci	Curie
dATP	2′-deoxyadenosine 5′-triphosphate
dCTP	2'-deoxycytidine 5'-triphosphate
DEPC	Diethylpyrocarbonate
dGTP	2'-deoxyguanosine 5'-triphosphate
dH_2O	distilled water
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dNTP	Deoxynucleotide 5′ triphosphate
dsRNA	Double-stranded RNA
dTTP	2'-deoxythymidine 5'-triphosphate
E-value	Expect value
EDTA	Ethylenediaminetetraacetic acid
Eg-m	*Euglena gracilis* methylation-guide RNA
Eg-p	*Euglena gracilis* pseudouridine-guide RNA
EM	Electron microscopy
Gb	Gigabase
GTP	Guanosine 5′ triphosphate
HF	High fidelity
HOTAIR	HOX antisense intergenic RNA
k-turn	Kink turn
Kb	Kilobase
LB	Luria Bertani broth
LECA	Last Eukaryotic Common Ancestor
LINE	Long interspersed element
LNA	Locked nucleic acid
lncRNA	Long non-coding RNA
LSU	Large subunit rRNA
miRNA	MicroRNA
mRNA	Messenger RNA
MRP	Mitochondrial RNA processing
ncRNA	Non-coding RNA
NEB	New England Biolabs

NET	NaCl-EDTA-Tris
Nm	2′-O-methylation
nt	Nucleotide
OD	Optical density
OEB	Oligonucleotide elution buffer
PAGE	Polyacrylamide gel electrophoresis
PAP	Poly(A) polymerase
PCR	Polymerase chain reaction
piRNA	Piwi-interacting RNA
PNK	Polynucleotide kinase
Pol	Polymerase
pre-RNA	Precursor RNA
RACE	Random amplification of cDNA ends
RB	Resuspension buffer
RISC	RNA induced silencing complex
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
RNAi	RNA interference
RNase	Ribonuclease
RNP	Ribonucleoprotein particle
rRNA	Ribosomal RNA
RT	Reverse transcriptase
RTL-P	Reverse transcription at low dNTP-PCR
scaRNA	Small Cajal body RNA
sdRNA	snoRNA-derived RNA
SDS	Sodium dodecyl sulphate
siRNA	Small interfering RNA
snoMEN	snoRNA modulator of gene expression
snoRNA	Small nucleolar RNA
snoRNP	Small nucleolar ribonucleoprotein complex
snoRT	Small nucleolar RNA retroposon
snRNA	Small nuclear RNA
snRNP	Small nuclear ribonucleoprotein complex
sRNA	snoRNA-like RNAs in archaea
sRNP	Small RNA ribonucleoprotein complex
SRP	Signal recognition particle
SSC	Saline-sodium citrate
SSPE	Saline-sodium phosphate EDTA
SSU	Small subunit rRNA
TAP	Tobacco acid pyrophosphatase
TE	Tris-EDTA
TEN	Tris-EDTA-NaCl
TMG	2,2,7-trimethyl guanosine
Tris	2-Amino-2-hydroxymethyl-propane-1,3-diol
tRNA	Transfer RNA
UTR	Untranslated region
Xist	X inactivation specific transcript

Chapter 1: Introduction

1.1 Non-coding RNA

Cellular life is based on the expression of genetic information that is stored in the form of deoxyribonucleic acid (DNA). Historically, ribonucleic acid (RNA) was largely considered just a temporary intermediate between DNA (the source of genetic information) and proteins (the functional embodiment of genetic information) (Crick, 1958). This flow of genetic information from DNA to RNA to protein was coined 'the central dogma of molecular biology' (Crick, 1970) and for decades, it was thought that the major role of RNA was as a messenger of genetic information. In the 1980s, the discovery of catalytically active RNAs called ribozymes (Kruger et al., 1982; Guerrier-Takada et al., 1983) and development of the 'RNA World' hypothesis (Gilbert, 1986) challenged the view of RNA as merely an intermediate product. The RNA World hypothesis proposes that life originated with RNA, a molecule that can both store genetic information and carry out catalytic functions. The importance of RNA in conserved cellular processes has been well known for decades. It serves as a primer during DNA replication, is a messenger of genetic information (mRNA), an adaptor during protein synthesis (tRNA) and an essential component of the ribosome. In fact, the catalytic center of the ribosome (the peptidyltransferase center) is composed of RNA (Ban et al., 2000) and peptide bond formation is the result of ribozyme activity (Nissen et al., 2000). Undeniably, RNA is involved in the most fundamental and conserved processes in all cells. However, uncovering the coding capacity of genomes for other functional RNAs and discovering the diverse roles RNA molecules play in the cell, is a challenging ongoing process.

Since the human genome was sequenced more than a decade ago (Lander et al., 2001), much work has been done to annotate and decode the information. Surprisingly, protein coding genes only make up less than 2% of the genome (International Human Genome Sequencing Consortium, 2004), while up to 90% is predicted to be transcribed (Clark et al., 2011); however, this value has been disputed (Pertea, 2012) (Figure 1). These transcribed (but not translated) regions are termed 'non-coding' and include introns, untranslated regions (UTRs) and genomic regions that are not annotated as protein coding (Frith et al. 2005). Large tiling arrays of 10 human chromosomes (Cheng et al., 2005) and massively parallel signature sequencing (Jongeneel et al., 2005) support a large ratio of transcribed non-coding regions to translated regions (coding regions). A remarkable number of functional non-coding RNAs (ncRNAs) constitute a large portion of the untranslated regions of the human transcriptome (Mattick and Makunin, 2006).

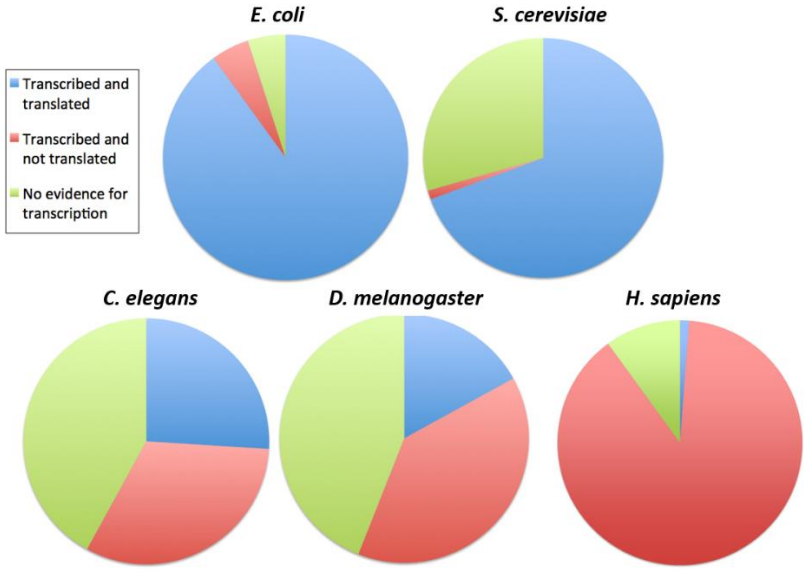


Figure 1. The relative proportion of various organisms' genomes that are transcribed into RNA, translated into protein or not expressed. As organismal complexity increases, the proportion of the genome dedicated to protein-coding genes (translated regions) decreases, as the proportion that is transcribed but not translated increases. Presumably, these transcribed (but not translated regions) encode numerous non-coding RNAs with diverse functions. Adapted from Huttenhofer et al. (2005).

The term ncRNA typically refers to any RNA molecule that is not translated into protein. They are often *trans*-acting molecules that have incredibly diverse cellular roles in all three domains of life (Huttenhofer et al., 2005; Mattick and Makunin, 2006; Matera et al., 2007). The use of both experimental and computational methods led to the surprising discovery that large portions of the genomes in a range of organisms are transcribed but not translated. More recently, this approach to characterizing transcribed RNAs has been termed 'experimental RNomics' (Huttenhofer et al., 2002; Huttenhofer et al., 2005). While prokaryotic genomes are typically dominated by protein-coding genes (80 – 95%) (Mattick, 2004), many ncRNAs with important regulatory roles have been discovered (Waters and Storz, 2009; Gottesman and Storz, 2011; Marchfelder et al., 2012). As organism complexity increases, the proportion of the genome encoding proteins decreases, with a concomitant increase in transcribed, non-coding sequences (Mattick, 2004; Huttenhofer et al., 2005) (Figure 1). Initially, many of these non-coding sequences were thought to be 'transcriptional noise' or 'junk' RNA that had no significant role in the cell. However, experimental evidence has revealed that many ncRNAs play important roles in an array of cellular processes. The functions of other more recently identified ncRNAs are less well understood and characterization of these RNAs is ongoing (Rother and Meister, 2011). In general, ncRNAs can be classified as 'small' (< 300 nt) or 'long' (> 300 nt) ncRNAs. Many classes of small ncRNAs have been characterized that are involved in RNA processing and modification, and the regulation of gene expression (Table 1). More recently, large-scale transcriptome analyses have also revealed an abundance of long ncRNAs (lncRNAs), which are often involved in chromatin modification (Goodrich and Kugel, 2009; Mercer et al., 2009; Faust et al., 2012) (Table 1).

Table 1. Examples of non-coding RNAs with diverse cellular functions

ncRNA	Function	Size	Found in*
	Translation		
tRNA	mRNA translation	70 - 90 nt	A, B, E
	Transcription regulation		
6S RNA	Transcription regulation through interaction with RNA polymerase	~184 nt	B
7SK RNA	Transcription regulation through association with a transcription elongation factor	~330 nt	Metazoans
Evf-2	Transcriptional regulation through interaction with activators	~3.8 kb	Mammals
	Gene expression regulation		
miRNA	Post-transcriptional gene silencing through translational repression or mRNA cleavage	21 - 25 nt	E
siRNA	Post-transcriptional gene silencing through mRNA cleavage	21 - 23 nt	E
piRNA	Germ line-specific transposon silencing	26 - 31 nt	Metazoans
Antisense RNA	Post-transcriptional gene silencing (variety of methods)	100 - >1 kb	B
	RNA processing & modification		
RNase MRP	rRNA processing	265 nt (*H. sapiens*)	E
RNase P	tRNA processing	340 nt (*H. sapiens*)	A, B, E
Spliceosomal RNA (snRNA)	Pre-mRNA intron removal	90 - 220 nt	E
Small nucleolar RNA (snoRNA)	Pre-rRNA processing and RNA modification (2′-O-methylation and pseudouridylation)	~70 - 330 nt	A, E
	Protein trafficking		
SRP RNA	Protein targeting to cellular membranes	110 - 520 nt	A, B, E
	Chromosome maintenance		
Telomerase RNA	Telomere synthesis	450 nt (*H. sapiens*)	E
	Chromatin modification		
HOTAIR	Transcriptional regulation through chromatin modification	2.2 kb	Mammals
Xist	X chromosome inactivation through chromatin modification	15 - 17 kb	Mammals

*Archaea (A), Bacteria (B), Eukaryotes (E)

Small ncRNAs are important components of many cellular processes. As previously mentioned, tRNAs are an essential component of protein translation in all domains of life. Other small RNAs are involved in transcription regulation either by interacting directly with the RNA polymerase complex (6S RNA in bacteria) (Wassarman, 2007) or through association with elongation factors (7SK in eukaryotes) (Peterlin et al., 2012). Some small ncRNAs regulate gene expression in eukaryotes post-transcriptionally in a process termed RNA interference (RNAi, see section 1.6) (Fire et al., 1998). Endogenous microRNAs (miRNAs) affect gene expression through complementary base-pairing to target mRNA species, which results in translational repression or mRNA cleavage (Huntzinger and Izaurralde, 2011). Small interfering RNAs (siRNAs) have a similar function, although they are often exogenous and introduced into cells to target specific genes for silencing through cleavage of the target mRNA (Rana, 2007). While there is no homologous RNAi pathway in bacteria, they can utilize antisense ncRNAs to regulate gene expression by base-pairing to target mRNA (Gottesman and Storz, 2011). In eukaryotes, splicing is another post-transcriptional process that ncRNAs are involved in. Pre-mRNA contains both coding regions (exons) and intervening sequences (introns), which are removed (spliced) prior to translation. Removal of introns is catalyzed by the spliceosome – a large ribonucleoprotein (RNP) complex composed of five small nuclear RNAs (snRNAs) and many associated proteins (Jurica and Moore, 2003; Matera et al., 2007). Small ncRNAs are also involved in RNA processing and modification. Small nucleolar RNAs (snoRNAs) play a role in the pre-ribosomal RNA (pre-rRNA) processing pathway and also guide RNA modification events (see 1.2) (Bachellerie et al., 2002). Additional ncRNAs involved in RNA processing include RNase P, which plays a role in pre-tRNA maturation and the evolutionarily-

related RNase MRP that is involved in pre-rRNA processing (Esakova and Krasilnikov, 2010).

Unique roles of large ncRNAs include protein translocation to cellular membranes (signal recognition particle, SRP RNA) (Rosenblad et al., 2009) and telomere synthesis (telomerase RNA) (Blackburn and Collins, 2011). In addition, multiple lncRNAs have been found that control transcription in eukaryotic cells by mediating changes in chromatin structure at specific genomic loci. HOTAIR is a human lncRNA that acts in *trans* to silence transcription across 40 kb of the HOXD locus by inducing heterochromatin formation (Rinn et al., 2007). Xist is a lncRNA that causes X chromosome inactivation (silencing of one of the X chromosomes) in mammals by facilitating the recruitment of chromatin silencing factors that promote heterochromatin formation (Penny et al., 1996; Zhao et al., 2008). Other lncRNAs regulate gene expression by modulating transcriptional activators. Such RNAs include Evf-2, which is transcribed from an enhancer element and subsequently recruits a transcriptional activator protein to this same enhancer element (Feng et al., 2006).

The collection of ncRNAs described here exemplify the diversity of mechanisms utilized by cellular RNAs to carry out an array of typically essential functions. There are many other identified ncRNAs whose functions remain to be elucidated (Costa, 2010).

1.2 Small nucleolar RNAs: structure and function

1.2.1 Ribosomal RNA modifications

Ribosomal RNA (rRNA) genes in eukaryotic organisms are transcribed as long, precursor transcripts that must undergo elaborate processing prior to becoming mature, functional rRNA molecules. The processing pathway, which occurs in a region of the

nucleus called the nucleolus, includes various nucleotide modifications and numerous successive endo- and exonucleolytic cleavage events, many of which occur co-transcriptionally (Sloan et al., 2014; Turowski and Tollervey, 2015). Two of the most abundant rRNA modifications are 2′-O-methyl groups (Nm) and pseudouridine residues (ψ) (Figure 2). 2′-O-methylation provides protection against alkaline hydrolysis and nuclease cleavage and can stabilize helices by favouring the 3′ endo ribose conformation (Williams et al., 2001), while the ψ modification can increase base stacking and enhance RNA stability by providing a more stable hydrogen bond donor (compared to uridine) (Ge and Yu, 2013). It is thought the main function of these modifications is to facilitate RNA folding and enhance stability (Helm, 2006). Many of these modifications cluster to highly conserved and functionally important regions of the ribosome, including the peptidyltransferase center, sites of tRNA binding, the peptide exit tunnel and the inter-subunit bridge (Decatur and Fournier, 2002). Many single rRNA modifications are dispensable, although systematic removal of several modifications in functionally important regions results in impaired translation (King et al., 2003; Baudin-Baillieu et al., 2009; Liang et al., 2009), slower growth, and increased sensitivity to ribosome-targeting antibiotics (Liang et al., 2007; Piekna-Przybylska et al., 2008). Conversely, global disruption of Nm or ψ modifications results in severe growth defects in yeast (Tollervey et al., 1993; Zebarjadian et al., 1999) and reconstituted *E. coli* ribosomes lacking rRNA modifications display serious catalytic defects (Green and Noller, 1996). The predominant hypothesis is that while individual rRNA modifications are non-essential, collectively they optimize rRNA structure for accurate and efficient ribosome synthesis and function.

The specific sites of nucleotide modification (and some cleavage events) in eukaryotes are predominately guided by a class of ncRNAs called small nucleolar RNAs (snoRNAs), which function in association with conserved protein components to form snoRNP complexes (Reichow et al., 2007; Watkins and Bohnsack, 2012). Homologs of snoRNAs and their associated proteins are found in archaea (Gaspin et al., 2000; Omer et al., 2000), where due to the lack of a nucleus, they are termed 'snoRNA-like' RNAs (sRNAs). The presence of such RNAs in these two domains of life indicates that RNA-guided rRNA modification is an evolutionarily-ancient mechanism (Omer et al., 2000) (see 1.4).

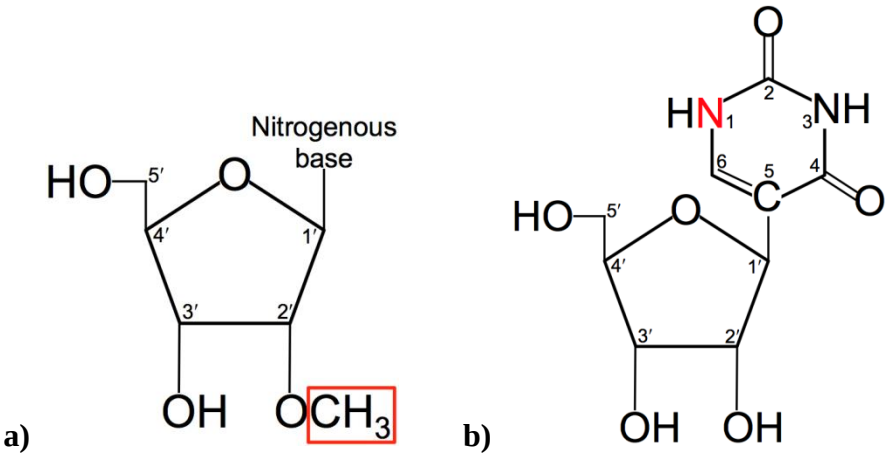


Figure 2. Common RNA modifications. a) The addition of a methyl group to the 2′ hydroxyl of the ribose sugar molecule results in a 2′-O-methylation (Nm) modification. **b)** Pseudouridine is formed through the isomerization of uridine. The glycosidic bond is broken, the base is rotated 180° and a new bond is formed between the C1′ position of the sugar and the C5 position of uracil. The sites indicated in red show the modified position.

1.2.2 snoRNA structure

There are two main classes of snoRNAs, box C/D and box H/ACA, which guide site-specific 2′-O-ribose methylation (Kiss-Laszlo et al., 1996; Nicoloso et al., 1996; Tycowski et al., 1996) and the isomerization of uridine to pseudouridine, respectively

(Ganot et al., 1997a; Ni et al., 1997). In addition, a subset of essential snoRNAs from both classes (U3, U14, U8 and U13 for box C/D, snR30/U17 and snR10 for box H/ACA) are instead involved in pre-rRNA cleavage events (see 1.2.3.2) (Henras et al., 2008). snoRNAs are *trans*-acting molecules that interact with their target rRNA site through base-pairing (Bachellerie et al., 1995) (Figure 3). Each class of snoRNA is defined by conserved sequence (or 'box') elements, as well as conserved secondary structural features.

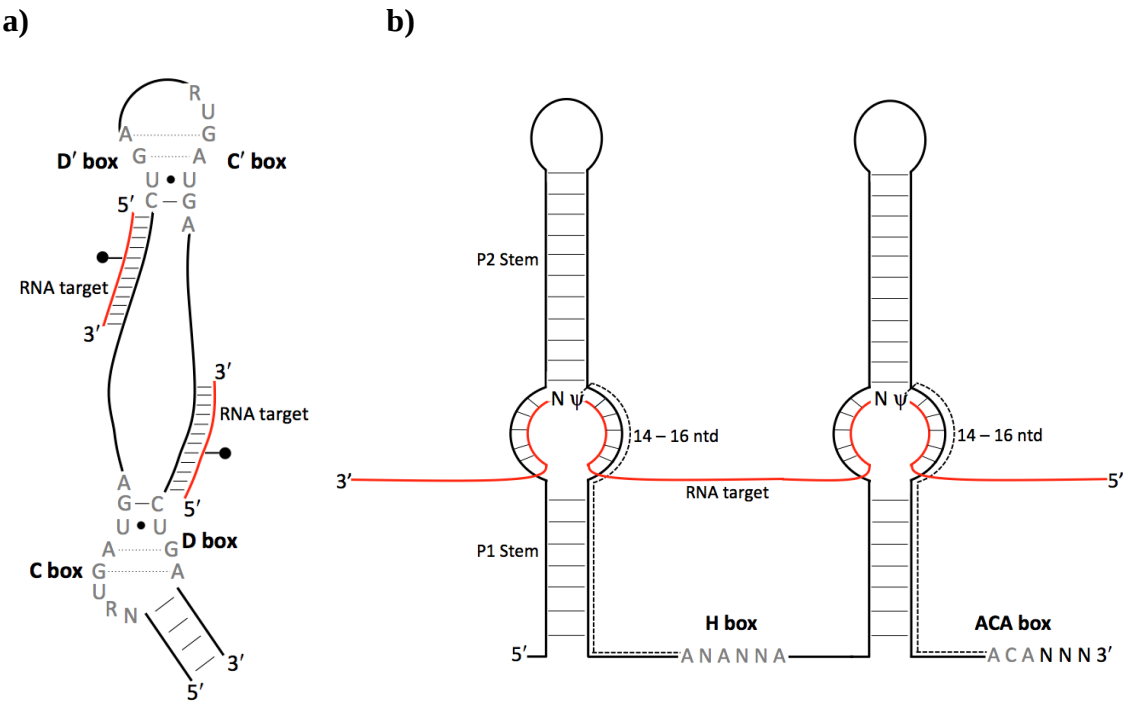


Figure 3. Typical eukaryotic small nucleolar RNAs interacting with target RNA molecule(s). a) Schematic of a representative box C/D snoRNA. The RNA site targeted for modification is paired exactly 5 nucleotides upstream of the D′ and/or D box element (shown with a filled circle). The C/D and C′/D′ elements interact to form conserved structural motifs (called the K-turn). **b)** Schematic of a box H/ACA snoRNA. The RNA target interacts with the snoRNA in the bulged 'pseudouridine pocket'. The nucleotide targeted for modification is an unpaired U that gets converted to a pseudouridine (ψ). In both schematics, the target RNA molecule(s) is shown in red and the conserved box elements are shown in grey.

1.2.2.1 Box C/D RNAs

Box C/D RNAs contain conserved sequence elements termed the 'C-box' (RUGUGA) and 'D-box' (CUGA), which are located at the 5′ and 3′ ends of the molecule, respectively (Figure 3a). There are also internal copies of the box elements (C′ and D′ box), which tend to be more sequence degenerate in eukaryotes. The sequence motifs are important for snoRNA processing, accumulation and localization (Huang et al., 1992; Balakin et al., 1996). The snoRNA 'guide region' (the region that interacts with the target RNA) is upstream of either the D or D′ box (or both, in which case it is referred to as a double-guide RNA). The specific nucleotide targeted for modification is base-paired to the snoRNA nucleotide positioned exactly 5 nucleotides upstream of the CUGA box element (the N+5 rule) (Kiss-Laszlo et al., 1996; Kiss-Laszlo et al., 1998) (Figure 3a). The length of complementary base-pairing between the target RNA and the box C/D RNA varies. In eukaryotes, this interaction is typically 10 – 21 bp (Kiss, 2002), while in archaea it is usually shorter due to the constrained distance between the box elements (Tran et al., 2005), which limits the possible maximum length of the guide regions.

Box C/D RNAs also adopt a conserved secondary structure. Typically, a terminal stem composed of 4-5 base pairs using nucleotides from each terminus brings the C and D box elements together (C/D motif) and together forms a conserved secondary structure called a kink-turn (K-turn) (Watkins et al., 2000), an important protein binding motif (Lilley, 2012). The more degenerate C′ and D′ boxes can also associate (C′/D′ motif), forming a K-turn like motif (the K-loop) (Reichow et al., 2007). These motifs provide two separate protein binding centers; the proteins that associate with box C/D RNAs are listed in Table 2. Recently, the crystal structure of an archaeal sRNP was resolved (Lin et

al., 2011) that contains one sRNA molecule and two copies of each protein. An earlier structure of a reconstituted box C/D sRNP generated by single-particle negative stain electron microscopy (EM) revealed a di-RNP with two sRNA molecules and four copies of each of the proteins (Bleichert et al., 2009). The mono-RNP is the more "conventional" structure and whether both structures may exist in the cell (perhaps transiently interchanging) still remains to be determined.

Table 2. Box C/D and box H/ACA RNP protein components in eukaryotes and archaea

Box C/D RNP		
Humans	**Yeast**	**Archaea**
15.5 kD	Snu13	L7Ae
Nop56	Nop56	Nop5
Nop58	Nop58	
Fibrillarin*	Nop1*	Fibrillarin*
Box H/ACA RNP		
Nhp2	Nhp2	L7Ae
Nop10	Nop10	Nop10
Gar1	Gar1	Gar1
Dyskerin*	Cbf5*	Cbf5*
* indicates the catalytic protein (methyltransferase for box C/D RNP or ψ synthase for box H/ACA RNP)		

While the general structure and sequence elements of box C/D RNAs are conserved amongst both archaeal and eukaryotic organisms, there are specific noteworthy structural differences. In archaea, box C/D RNAs are smaller and more uniformly sized than in most studied model eukaryotes; the median length of these RNAs in archaea is 57 nucleotides (with little variation) compared to eukaryotes where the median size is >70 nucleotides with more size variation (Dennis et al., 2001). Double-guide RNAs are more common in archaea, where they typically target nearby sites in the primary and/or tertiary rRNA structure (Dennis et al., 2001). This would allow for the formation of two guide

complexes on adjacent rRNA sites and could allow the hypothesized chaperone function of these RNAs (Bachellerie et al., 2002).

1.2.2.2 Box H/ACA RNAs

Box H/ACA RNAs adopt a secondary structure of two extended stem-loop domains connected by a single-stranded hinge region; they also contain conserved sequence elements, namely the 'H-box' (ANANNA) in the hinge region between stems and an 'ACA-box' invariably located 3 nucleotides from the 3′ end of the molecule (Figure 3b) (Ganot et al., 1997b). Both the box elements and stem loop structures are important for snoRNA accumulation and processing *in vivo* (Ganot et al., 1997b). The guide regions of box H/ACA RNAs are the single-stranded bulge regions that interrupt the stem loops; the RNA target binds to this region in a bipartite fashion (Figure 3b). The length of this bipartite base-pairing interaction between the guide RNA and the target RNA is typically 9 – 13 bp. The nucleotide targeted for modification is a 5′ unpaired uridine in this so-called 'pseudouridylation pocket' and is typically positioned 14 – 16 nucleotides from either an H or ACA box element (Ganot et al., 1997a; Bortolin et al., 1999). Both stem loops can be used to guide the isomerization of uridine to pseudouridine (double-guide RNA) or just one may be used. In yeast and mammals, both hairpins are required for *in vivo* pseudouridylation activity, even if only one is used to guide the modification (Bortolin et al., 1999) and both stems are essential for RNP formation *in vitro* (Dragon et al., 2000). However, in some single-celled eukaryotes such as trypanosomes and *Euglena gracilis*, identified pseudouridine guide RNAs contain only a single stem (see 1.5) (Russell et al., 2004; Liang et al., 2005; Moore and Russell, 2012),

which is also the typical structure of these RNAs in archaeal organisms (two and three stem box H/ACA RNAs are also found in archaea) (Omer et al., 2003).

The evolutionarily conserved proteins that associate with box H/ACA RNAs are listed in Table 2. The crystal structure of the archaeal H/ACA sRNP shows one copy of each protein bound to a single-stem RNA (Li and Ye, 2006). Recently, numerous high resolution structures of the H/ACA complex with different combinations of RNP components have been generated (Hamma and Ferre-D'Amare, 2010), providing insight into the molecular structure and intricate protein-RNA interactions of the complex. The apical (P2) stem of archaeal H/ACA RNAs (refer to Figure 3b) contains a K-turn motif; interestingly, the archaeal L7Ae protein binds the K-turn motif in both the C/D and H/ACA RNAs (Rozhdestvensky et al., 2003). Eukaryotic H/ACA RNAs do not contain a K-turn/K-loop motif. It is proposed that the general architecture of the eukaryotic snoRNP is similar to the archaeal sRNP, given the similarities between the protein components and RNA structure (Reichow et al., 2007). However, because typical eukaryotic box H/ACA RNAs contain two stems, two sets of proteins would presumably assemble on one RNA molecule (Watkins et al., 1998; Kiss et al., 2010).

1.2.3 snoRNA function

1.2.3.1 Other RNA targets for modification

Guiding ribosomal RNA modifications is the main and best characterized function of snoRNAs (see 1.2.1). RNA targets for modification-guide snoRNAs are not limited to rRNA, however. In eukaryotes, a group of snoRNA-like molecules localize to Cajal bodies – subnuclear foci involved in snRNP and snoRNP biogenesis (Machyna et al., 2013). These RNAs are called small Cajal body RNAs (scaRNAs) and they guide the modifications of snRNAs (Tycowski et al., 1998; Darzacq et al., 2002). In archaea and

possibly in some eukaryotes, snoRNAs can also guide tRNA modifications (Dennis et al., 2001; Zemann et al., 2006; Chakrabarti et al., 2007; Liu et al., 2009b). It was previously demonstrated that the *Trypanosome* spliced leader RNA has a snoRNA-guided pseudouridine site (Liang et al., 2002b; Zamudio et al., 2009). Recent studies reveal that mRNA from humans and yeast are pseudouridylated and the modifications may depend on both stand-alone pseudouridine synthases as well as snoRNA-guided pseudouridine synthases (Carlile et al., 2014; Lovejoy et al., 2014; Schwartz et al., 2014). In these studies, it was also reported that numerous snoRNA species themselves contain pseudouridine modifications (Schwartz et al., 2014). The modifications appear to be regulated in response to environmental conditions and are proposed to enhance transcript stability. The isomerization of uridine to pseudouridine on mRNA templates can also facilitate non-canonical base-pairing in the ribosome decoding centre, leading to alterations in the genetic code (Karijolich and Yu, 2011; Fernandez et al., 2013). How widespread mRNA pseudouridylation is amongst different organisms and the role snoRNAs may play in targeting these sites is currently unknown.

1.2.3.2 Pre-rRNA processing

In addition to guiding RNA modification, a subset of snoRNAs are involved in eukaryotic rRNA processing (i.e. RNA cleavage and/or folding). Pre-rRNA transcripts are synthesized as long polycistronic precursors by RNA polymerase I (RNA pol I), which undergo a series of cleavage events to produce mature rRNA species. U3, an evolutionarily conserved box C/D snoRNA, is a component of the small subunit (SSU) processome – a large RNP complex essential for 18S rRNA processing (Phipps et al., 2011). U3 base-pairs with multiple pre-rRNA sites and it is predicted that these interactions facilitate rRNA folding and/or potentially guide nucleases to cleavage sites

(Hughes, 1996; Dutca, et al. 2011). U14, another box C/D snoRNA, is a dual-function snoRNA in many eukaryotes. Utilizing two separate domains, U14 binds to two distinct regions of pre-rRNA where one domain directs nucleotide methylation and the other plays an essential role in 18S rRNA processing (Liang and Fournier, 1995) (see 3.3). Other box C/D snoRNAs shown to be involved in pre-rRNA processing include U8 and U13. These RNAs are only found in higher eukaryotes and are predicted to regulate RNA folding (Watkins and Bohnsack, 2012). In *Xenopus*, U8 plays an essential role in 5.8S and 28S rRNA maturation (Peculis and Steitz, 1993; Peculis, 1997). While the functional importance of U13 has yet to be conclusively determined, it may play a role in 18S rRNA 3′ end formation (Cavaille et al., 1996). There are also H/ACA box snoRNAs involved in eukaryotic pre-rRNA processing. snR30 (U17 in humans) interacts with a eukaryotic specific internal region of the 18S rRNA using evolutionarily conserved sequence elements and is essential for 18S rRNA processing (Morrissey and Tollervey, 1993; Atzorn et al., 2004; Fayet-Lebaron et al., 2009). In yeast, snR10 acts a dual-function snoRNA using the 3′ domain to guide a pseudouridylation modification, and rRNA processing requires the 5′ domain (Liang et al., 2010). Since pre-rRNA processing at many sites occurs co-transcriptionally, it is predicted that these snoRNAs also function co-transcriptionally and interact with rRNA prior to some cleavage events (Turowski and Tollervey, 2015).

1.2.3.3 Novel functions

snoRNAs that show no apparent sequence complementarity to conventional substrates, such as rRNAs or snRNAs, are termed ‘orphan snoRNAs’ and are predicted to target other cellular RNA species. A brain-specific orphan box C/D snoRNA

(SNORD115, previously called HBII-52) shows complementarity to a region of the serotonin 2C receptor pre-mRNA. The role of SNORD115 in post-transcriptional regulation of this pre-mRNA has been studied in detail. This snoRNA (which is linked to the rare neurological disorder Pradar-Willi syndrome) affects both alternative splicing and editing of this mRNA (Vitali et al., 2005; Kishore and Stamm, 2006). SNORD115 is also processed into smaller RNAs and can influence alternative splicing of other pre-mRNAs (Kishore et al., 2010). A family of human snoRNAs were also identified that are partially complementary to several pre-mRNA species. While the effect (if any) these snoRNAs have on protein expression *in vivo* was not documented, the snoRNA sequences were modified to create snoRNA modulator of gene expression (snoMEN) vectors that could reduce gene expression by specifically silencing targeted genes (Ono et al., 2010).

Recently, numerous studies have reported that snoRNAs (both box C/D and H/ACA) are processed into shorter, stable RNA species (called snoRNA-derived RNAs, sdRNAs), many of which resemble miRNAs in that they depend on Dicer for processing and associate with argonaute proteins (Ender et al., 2008; Taft et al., 2009; Brameier et al., 2011; Ono et al., 2011). These studies suggest that sdRNAs are not merely snoRNA degradation products, as common processing patterns have been observed and post-transcriptional gene silencing activity was observed for a number of sdRNAs (Brameier et al., 2011). However, it is unclear if these sdRNAs are actually processed directly from functional snoRNPs or if they are the result of alternative processing of host gene introns (vertebrate snoRNAs are typically intron-encoded, see 1.3.1) (Watkins and Bohnsack, 2012).

In addition to potentially playing a role in regulating gene expression, snoRNAs have been implicated in other intriguing cellular processes. Studies have demonstrated

that snoRNAs play a role in metabolic stress pathways (Michel et al., 2011) and are involved in intracellular cholesterol trafficking (Brandis et al., 2013) in mammalian cells. In the first study, loss of three intronic box C/D snoRNAs (U32a, U33 and U35a) provided resistance to lipotoxicity and oxidative stress. The second study demonstrated that loss of a single box C/D snoRNA (U60) caused deficiency in cholesterol trafficking from the plasma membrane to the endoplasmic reticulum. In both cases, the loss of specific intronic snoRNAs, and not the spliced exonic transcripts, were responsible for the observed phenotypes. These non-canonical snoRNA cellular roles seem to be independent of their canonical role in rRNA modification; in both cases, sites of $O^{2'}$-ribose methylation predicted to be guided by these snoRNAs were unaffected. And finally, recent preliminary studies suggest that snoRNAs may play a role in cancer. Some snoRNAs have been shown to be differentially expressed in cancer cells and may affect oncogene expression (Mannoor et al., 2012; Williams and Farzaneh, 2012). Indeed it is becoming clear that snoRNAs are involved in a range of diverse cellular processes, many of which still require further characterization to fully understand the mechanistic role of the snoRNAs.

1.3 snoRNA gene organization and expression

Elements of snoRNA structure and function are highly conserved among eukaryotic organisms and between eukaryotes and archaea. The box elements, secondary structures, associated proteins and methods of targeting modification sites all show conservation. Conversely, the genomic organization and the modes of expression of snoRNA genes are highly variable. Hence, it has been said that snoRNAs represent a "paradigm for gene expression flexibility" (Dieci et al., 2009). There are two general

genomic arrangements of snoRNA genes: 1) as independent transcription units and 2) within introns of other genes. Variations of these general arrangements include: 1a) individual genes with their own promoters (independent-single); 1b) independent clusters of genes transcribed as polycistronic transcripts under a single promoter (independent-clustered); 2a) a single snoRNA gene per intron (intronic-single) and 2b) clusters of snoRNA genes within one intron (intronic-clustered) (Figure 4). In general, there tends to be one prevalent mode of organization in a particular taxon (Figure 5).

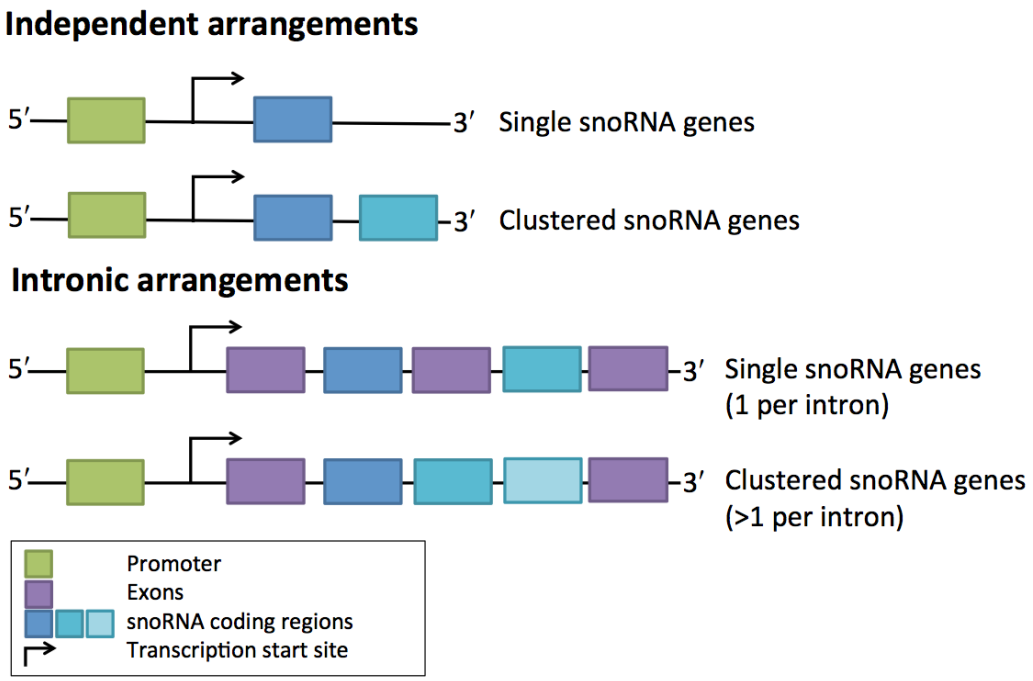


Figure 4. Modes of snoRNA gene organization. Eukaryotic snoRNA genes show a diverse range of genomic arrangements. Genes can be transcribed under their own promoters (independent arrangement), either as a single gene or as a polycistronic transcript. snoRNA genes also localize to the introns of other genes and are transcribed under the host gene's promoter (intronic arrangements). The snoRNA genes can be arranged as a single gene per intron or as clusters of genes in a single intron. Adapted from (Dieci et al., 2009).

Nearly all vertebrate snoRNA genes are located in the introns of other genes, predominately in a "one-gene-per-intron" (intronic-single) arrangement (Figure 5, *H. sapiens*) (Dieci et al., 2009) and are transcribed by RNA pol II. These snoRNAs are

processed via a splicing-dependent pathway; intron removal followed by lariat debranching and exonucleolytic trimming (Brown et al., 2008). An exception is genes encoding essential snoRNAs that are involved in rRNA processing; these snoRNA genes localize to intergenic regions and are transcribed under their own promoters, typically by RNA pol II (Maxwell and Fournier, 1995). The majority of non-intronic human snoRNA genes are the result of retrotransposition events and may no longer produce functional snoRNAs (Weber, 2006; Luo and Li, 2007). Other exceptions include genes that encode two snoRNAs that guide the modification of snRNA species. The genes appear to be independently transcribed by RNA pol II, making them the first characterized modification-guide snoRNAs in vertebrates that are organized as independent genes (Tycowski et al., 2004). As independent-single gene organization is the prevalent mode of expression in yeast, some protist organisms, and archaea, this suggests that intronic localization of snoRNA genes may have evolved later in eukaryotic evolution.

The model invertebrate species in which snoRNA gene organization has been best characterized are *C. elegans* and *D. melanogaster*. The predominant mode of gene organization in invertebrates is intronic-single genes (Deng et al., 2006; Tycowski and Steitz, 2001), although deviations from this organization exist. In *C. elegans*, approximately 25% of snoRNA genes are independently transcribed from their own promoters (Figure 5) and a unique organization for animals, intronic-clustered, appears in *D. melanogaster* (Huang et al., 2004).

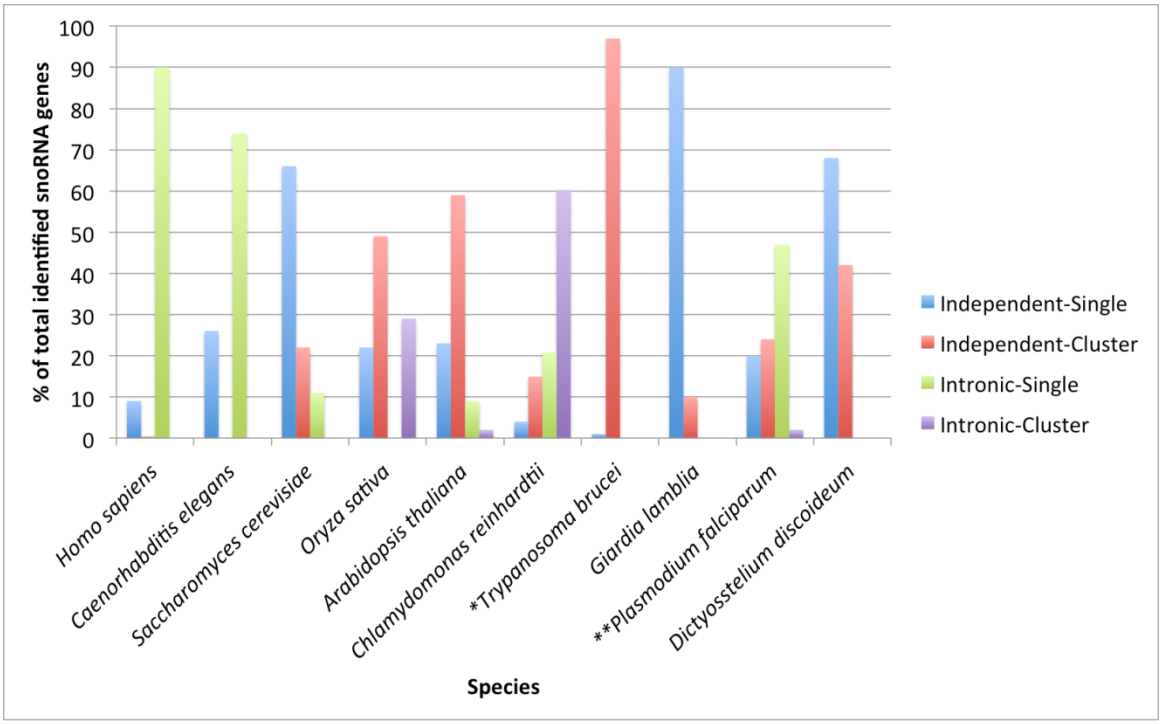


Figure 5. Genomic localization of snoRNA genes in a range of eukaryotic organisms. Typically there is one predominant snoRNA genomic arrangement in a particular species. snoRNA data for the organisms analyzed are based on the following references: *H. sapiens, C. elegans, S. cerevisiae, O. sativa* and *A. thaliana* (Dieci et al., 2009); *C. reinhardtii* (Chen et al., 2008); *T. brucei* (Liang et al., 2005; Uliel et al., 2004); *G. lamblia* (Chen et al., 2011; Hudson et al., 2012; Yang et al., 2005); *P. falciparum* (Chakrabarti et al., 2007; Mishra et al., 2009); *D. discoideum* (Aspegren et al., 2004).
* only two *cis*-spliced introns have been identified in *T. brucei*, so the possibility exists that snoRNA genes exist in uncharacterized introns. ** *P. falciparum* also has a snoRNA gene located in the 3′ UTR of a ribosomal protein gene. Adapted from (Dieci et al., 2009).

In yeast, the majority of snoRNA genes are independently arranged and expressed as monocistronic transcripts (Figure 5). Alternative arrangements include intronic-single genes (Lowe and Eddy, 1999) and independent clusters transcribed under a single promoter (Qu et al., 1999). Independent yeast snoRNA genes (whether single or clustered) are transcribed from promoters containing TATA elements (RNA pol II transcripts) and Rap1p binding sites (a transcription factor also involved in the expression of ribosomal-protein genes) (Qu et al., 1999). Intronic snoRNAs are processed via a

splicing dependent pathway, similar to vertebrates. Polycistronic snoRNA precursor transcripts must undergo cleavage to produce individual snoRNAs. The enzyme Rnt1p (RNase III) recognizes and cleaves conserved stem-loop structures found in spacer regions between yeast snoRNAs. Exonucleolytic cleavage to produce mature snoRNA ends is carried out by the exosome (3′-5′ exonuclease cleavage) and by Xrn1p and Rat1p (5′-3′ cleavage) (Chanfreau et al., 1998; Qu et al., 1999). More recently, studies have suggested that polyadenylation is an important step in the pre-snoRNA 3′ end-processing pathway (Grzechnik and Kufel, 2008; Lemay et al., 2010). Mature snoRNAs are non-polyadenylated, however the transient presence of a polyA tail may be an important part of their regulation and processing.

Plants show remarkable diversity in snoRNA gene organization. Approximately 78% of snoRNA genes in the model monocotyledonous plant *Oryza sativa* are arranged in clusters (Figure 5). Interestingly, a prevalent mode of genomic organization is intronic-clustered, an organization which is not found in vertebrates and yeast, and rarely observed in other eukaryotes. The majority of the intronic clusters are composed of multiple copies of the same snoRNA gene, suggesting gene duplication as a means by which snoRNA clusters originate (see 1.4) (Liang et al., 2002a). However, clusters can also contain both classes of snoRNAs. In the dicotyledonous plant *Arabidopsis thaliana,* snoRNAs are also predominately arranged in clusters, but the majority (59%) are independent-clusters while only 2% are intronic-clusters. *A. thaliana* is known to have a small (125 Mb), relatively compact genome compared to other plant species, with the average intron length being ~170 bp (Arabidopsis Genome, 2000). Therefore, it is not surprising that only 2% of snoRNA genes are arranged as intronic-clusters, as the average intron could only

accommodate two snoRNA genes at most. Conversely, the *O. sativa* genome (420 Mb) is approximately three times as large as that of *A. thaliana* (Goff et al., 2002), with larger introns that are able to accommodate numerous snoRNA genes.

In both monocots and dicots, the processing of the intronic-clustered snoRNA genes is splicing-independent (Leader et al., 1999). Clustered snoRNAs in plants are processed via a series of endonucleolytic cleavages followed by exonucleolytic trimming (Leader et al., 1997; Brown et al., 2008). However, the specific components and mechanistic details of snoRNA processing in plants are not well characterized. The spacer regions between snoRNAs do not show conservation in size, sequence or secondary structure. Disruption of the gene encoding the *A. thaliana* orthologue of Rnt1p did not affect polycistronic snoRNA processing (Comella et al., 2008). Genome-wide analysis of transcripts affected by mutations to exosome components revealed that polycistronic snoRNA precursors require the activity of this complex for normal processing in *A. thaliana* (Chekanova et al., 2007), a mechanism similar to snoRNA processing in yeast. The promoters that regulate the transcription of the snoRNA clusters in plants are also not characterized. An interesting genomic arrangement has been observed in some plants resulting in the expression of tRNA-snoRNA dicistronic transcripts (Kruszka et al., 2003). The snoRNA is transcribed under the tRNA promoter by RNA pol III. It is proposed that the plant endonuclease tRNase Z, which processes the 3′ ends of tRNAs, is involved in processing this transcript (Barbezier et al., 2009).

Typically, reviews of eukaryotic snoRNA gene organization (and snoRNA studies in general) focus on model organisms which are often multicellular (plants, animals and fungi), and exclude the majority of unicellular eukaryotes collectively known as protists.

Protists are a large, diverse group of organisms and consistent with their diversity, snoRNA genes show all four types of genomic organizations in the protist species in which they have been characterized.

snoRNAs have been characterized in two parasitic protozoan species belonging to the eukaryotic supergroup Excavata, *Trypanosoma brucei* and *Giardia lamblia*. In *T. brucei* (the organism known to cause African trypanosomiasis), snoRNA genes are predominately arranged in polycistronic clusters, some of which have been shown to be transcribed by RNA pol II (Uliel et al., 2004). These gene clusters often contain both classes of snoRNAs and are repeated several times in the genome (Liang et al., 2005), similar to the organization observed in plants. Only two *cis*-spliced introns have been identified in *T. brucei*, so it is likely that most snoRNA clusters localize to intergenic regions (i.e. regions between protein-coding genes) (Siegel et al., 2010). The majority of snoRNAs characterized in *G. lamblia* (the parasitic protist that causes giardiasis) are independent genes found in intergenic regions between ORFs (Yang et al., 2005; Chen et al., 2011; Hudson et al., 2012), a result that is not surprising considering the genome contains very few introns (Nixon et al., 2002; Russell et al., 2005; Morrison et al., 2007; Roy et al., 2012;). Some snoRNA genes are found in small 'clusters' of two and are expressed as dicistronic transcripts. A conserved sequence motif has been identified downstream of *G. lamblia* snoRNAs (and other ncRNAs) that is present on precursor transcripts and is predicted to play a role in their processing pathway (Hudson et al., 2012).

Plasmodium falciparum, the apicomplexan protist that causes malaria, is another parasitic protist species in which snoRNAs have been characterized. The diversity of snoRNA genomic organization is evident in this one species alone, as it has all four gene

arrangements (Chakrabarti et al., 2007; Mishra et al., 2009). Like the other organisms studied to date, there appears to be one main mode of snoRNA genomic organization – intronic-single genes, which is unique amongst the other protist organisms studied. Many of the intergenic clusters of snoRNA genes in this parasite are orientated on the same DNA strand, suggesting they are co-transcribed (Chakrabarti et al., 2007), but expression studies have not yet validated this.

In the social amoeba *Dictyostelium discoideum*, the majority of snoRNAs that have been characterized localize to intergenic regions either as single genes or clustered together (Aspegren et al., 2004). Another unicellular species in which snoRNAs have been characterized is *Chlamydomonas reinhardtii* (a unicellular green algae). snoRNA gene clusters in *C. reinhardtii* predominately localize to intronic regions, the arrangement observed in some plant species. Typically the clusters are composed of homologous snoRNA genes that are likely the result of extensive local genomic duplication events. Like plants and some protists, snoRNA gene arrangement in *C. reinhardtii* is diverse, showing all four types of genomic arrangement (Figure 5) (Chen et al., 2008).

The expression of eukaryotic snoRNA genes that localize to the intronic regions of other genes is regulated by the host gene promoter. The promoters of independent snoRNA genes (whether single or clustered) are not well characterized, but it appears these genes are typically transcribed by RNA pol II (Table 3), as TATA-like elements have been reported and some of these snoRNAs contain a 2,2,7-trimethyl guanosine (TMG, $m^{2,2,7}G$) cap (Dieci et al., 2009; Maxwell and Fournier, 1995). In eukaryotes, nascent transcripts that are synthesized by RNA pol II are modified co-transcriptionally by the addition of a 7-methyl guanosine (m^7G) 5′ cap. This cap is characteristic of

eukaryotic mRNA and functions to protect mRNA from exonucleolytic cleavage and also promotes translation initiation. For some ncRNA pol II transcripts, such as snRNAs (U6 is an exception) and some snoRNAs, the m^7G cap is hypermethylated and converted to a $m^{2,2,7}G$ cap (Cougot et al., 2004). It is thought that the presence of a hypermethylated cap on nuclear RNAs (such as snoRNAs) might prevent leakage into the cytoplasm and/or assist in nuclear localization after mitosis (Cougot et al., 2004). In eukaryotes, snoRNAs independently transcribed by RNA pol II typically contain a TMG cap, while those encoded in introns or transcribed by RNA pol III, do not (Maxwell and Fournier, 1995).

Unique cases of transcription of modification-guide snoRNAs by RNA pol III have been identified in yeast (Guffanti et al., 2006) and invertebrates (Deng et al., 2006; Isogai et al., 2007) (Table 3). As previously mentioned, plant snoRNA-tRNA dicistronic precursors are also transcribed by RNA pol III, utilizing the tRNA gene promoter (Kruszka et al., 2003). In some species of plants (Kiss et al., 1991), green algae (Antal et al., 2000) and protists (Orum et al., 1992; Fantoni et al., 1994), the rRNA-processing snoRNA U3 (see 1.2.3.2) is also transcribed by RNA pol III, while in vertebrates, invertebrates and yeast, U3 is transcribed by RNA pol II (Jawdekar and Henry, 2008).

Table 3. Transcription of eukaryotic snoRNA genes by RNA polymerase II and III

snoRNA	Transcribed by	Observed in
(Most) modification-guide	RNA Pol II	All eukaryotes*
(Some) modification-guide	RNA Pol III	Yeast *C. elegans* *D. melanogaster* Plants
U3 (Pre-rRNA processing)	RNA Pol II	Invertebrates Vertebrates Yeast
	RNA Pol III	Plants Green algae Some protists

* The promoters of autonomously expressed snoRNA genes are not well-characterized but they generally appear to be transcribed by RNA pol II (see text).

Evidently, snoRNA gene organization and expression is flexible and extremely diverse. Through the course of evolution as snoRNA genes radiated, they became positioned in diverse locations within the genomes of various species. Given the varied genomic positioning and expression strategies, it appears that independent regulation of individual snoRNAs is unnecessary. It has been suggested that only high transcription levels of snoRNAs are required (either expressed independently from strong promoters or within the introns of highly expressed genes) (Dieci et al., 2009). This is consistent with reports that yeast snoRNA genes are one of the most highly occupied by RNA pol II (Steinmetz et al., 2006) and that intronic snoRNAs often localize to highly expressed housekeeping genes (often ribosomal protein genes) (Dieci et al., 2009).

1.4 snoRNA evolution

The discovery of homologous snoRNP components in archaea indicates that RNA-guided rRNA modification is an evolutionarily ancient process that originated in the common ancestor of archaea and eukarya, approximately 2 – 3 billion years ago (Gaspin et al., 2000; Omer et al., 2000). Bacterial rRNA is also modified, but not to the same extent observed in eukaryotes. For example, *E. coli* rRNA contains 10 ψ and 4 Nm modifications, whereas yeast have 43 and 55 and humans have 91 and 106, respectively (Lafontaine and Tollervey, 1998). In contrast to archaea and eukaryotes, rRNA modification in bacteria is guided by stand-alone modification enzymes that closely resemble the tRNA modification machinery in eukaryotic organisms (Becker et al., 1997; Lecointe et al., 1998). One hypothesis is that the RNA-guided modification system evolved from stand-alone enzymes. The pseudouridine synthase of box H/ACA complexes (Cbf5) is homologous to bacterial tRNA pseudouridine synthase TruB

(Koonin, 1996). It has been hypothesized that an ancestral gene encoding a TruB-like ψ synthase duplicated and then acquired the ability to recognize a new RNA partner, likely pre-rRNA or tRNA, which could act in *trans* to target RNA sites for modification (Lafontaine and Tollervey, 1998). Repeated gene duplications and sequence divergence (see 1.4.2) then led to the diverse collection of guide RNAs observed today. Further evidence that modification guide RNPs evolved from an rRNA-derived ancestor stems from the archaeal ribosomal protein L7Ae, which binds 23S rRNA and is also an essential component of both box H/ACA and C/D RNPs (Kuhn et al., 2002; Rozhdestvensky et al., 2003).

An alternative hypothesis is that snoRNAs originated in an RNA world and co-evolved with the early ribosome (Jeffares et al., 1995). In this hypothesis, the intronic location of snoRNAs may be connected to the origin of mRNA; the 'introns-first' theory proposes that in an early RNA world, exons developed from regions between RNA genes (Penny et al., 2009). Consistent with this model, the first exonic sequences would have emerged from the regions between individual snoRNA genes. If this theory were true, snoRNAs should localize to ancient introns and the positioning would likely be conserved. A recent study examined the genomic stability of snoRNA families' relative locations across eukaryotic genomes and found that the association of snoRNAs and introns is not conserved (Hoeppner and Poole, 2012). While some snoRNA families are traceable to the Last Eukaryotic Common Ancestor (LECA), no evidence was found for positional conservation. It was concluded that individual snoRNA genes are mobile (consistent with recent studies (Weber, 2006; Luo and Li, 2007; Schmitz et al., 2008)) and can occupy any genomic location (consistent with the diverse genomic arrangements

observed in eukaryotes), as long as an adequate expression profile is maintained (Hoeppner and Poole, 2012).

The antiquity of snoRNAs is established, based on their ubiquitous distribution in archaea and eukarya. It is also evident that they continue to evolve and diversify, as the number of target molecules and functional roles expands, often in a species-specific manner. Despite the apparent rapid rate of snoRNA evolution (Weber, 2006; Luo and Li, 2007; Schmitz et al., 2008; Hoeppner and Poole, 2012), there is a stable association between snoRNAs and their target sites within vertebrates (Kehr et al., 2014). The snoRNA antisense regions required for targeted modification are under high selective pressure to maintain complementarity to the target sequence, and therefore, redundant snoRNA species and their guide regions can mutate more freely. snoRNA evolution is predicted to mainly be the result of gene duplications, followed by sequence divergence (see below).

1.4.1 Mechanisms of snoRNA evolution

It is hypothesized that snoRNA evolution occurs through a series of repeated gene duplications, followed by sequence divergence and positive selection for the maintained ability to stably associate with the required proteins (Lafontaine and Tollervey, 1998), and to exert a positive influence on ribosome assembly and/or function. While some homologous snoRNAs are observed (snoRNAs in different organisms targeting a conserved modification site), many snoRNAs guide species-specific modification sites.

In many eukaryotic genomes, the genes encoding snoRNAs are often multi-copy (Dieci et al., 2009); conversely, archaea do not appear to contain many sRNA gene copies (Dennis et al., 2001). Gene duplication is mainly the result of unequal crossing over, retroposition or chromosomal/genome duplication (Bennetzen, 2002). Duplicated genes

have four potential fates: functional redundancy, pseudogenization, subfunctionalization and neofunctionalization (Lynch and Conery, 2000). If both gene copies are maintained and conserve their original function, the genes are functionally redundant. Pseudogenes arise when mutations accumulate in one copy, rendering it non-functional. Subfunctionalization occurs when mutations lead to both gene copies adopting part of the original gene's function and neofunctionalization occurs when one gene copy retains the original function while the other gains a novel function. Duplicated snoRNA genes exhibit all four evolutionary fates (Figure 6).

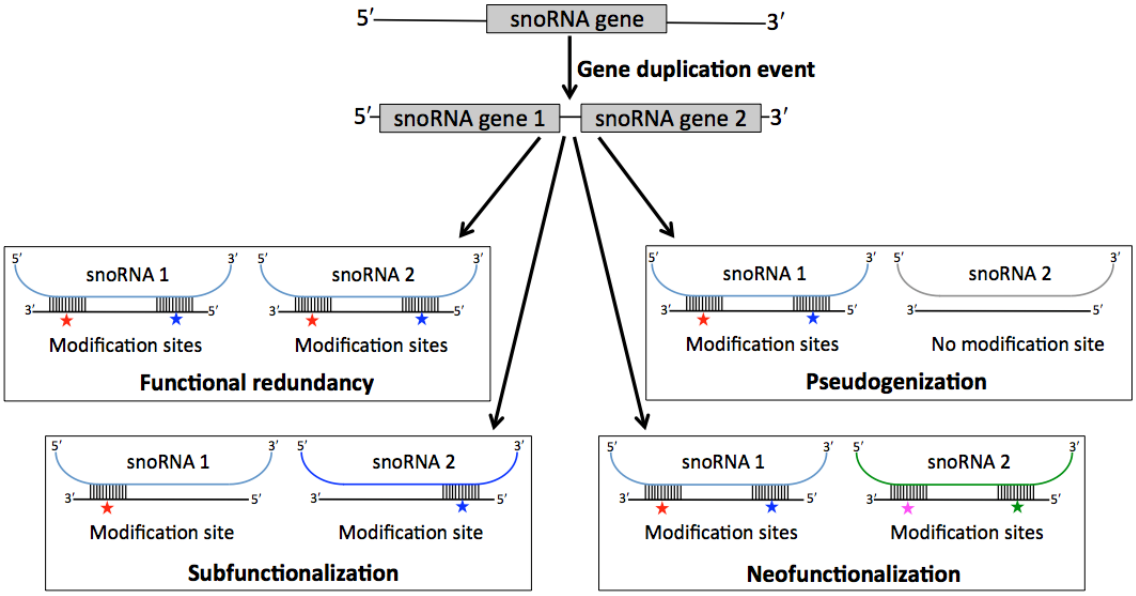


Figure 6. snoRNA evolution by gene duplication. An initial gene duplication event produces two copies of the original snoRNA, which can result in functional redundancy. Sequence divergence can lead to pseudogenization (one copy is no longer functional), subfunctionalization (two snoRNAs to perform the function of the original) or neofunctionalization (one copy obtains a new function/target). The colored stars represent unique sites of targeted modification.

Due to the high level of snoRNA gene copies in many organisms, functional redundancy is expected. Indeed it appears this is the case in numerous eukaryotes. In plants, for example, greater than 50% of identified snoRNA genes have additional full-

length copies, showing varying degrees of sequence conservation (Barneche et al., 2001; Chen et al., 2003). However, the sequence and structural elements required for functionality remain highly conserved (suggesting the extra copies are not pseudogenes). In the majority of snoRNA copies, the antisense regions used to guide nucleotide modifications are also highly conserved (Kehr et al., 2014), meaning the snoRNAs target the same site for modification; in other words, they are functionally redundant (Figure 6). snoRNA gene copies that show some sequence variation but guide the same modification site are referred to as snoRNA isoforms.

In plants, it appears that small, tandem duplications have produced snoRNA gene clusters that are often polycistronically transcribed. Entire clusters can be repeated in tandem as many as 5 times (Chen et al., 2003). In the protist *T. brucei*, snoRNA gene clusters can be repeated in the genome up to 7.5 times (Liang et al., 2005). Clusters of snoRNA genes that are duplicated together are considered 'linked'; similar linked snoRNA arrangements are conserved between *O. sativa* and *A. thaliana*, suggesting the linkage of some snoRNAs was conserved during plant evolution. Conversely, it appears some algal snoRNA clusters have been generated more recently. Two species of green algae (*Chlamydomonas reinhardtii* and *Volvox carteri*) have similar sized genomes, yet the number of snoRNA genes and snoRNA gene paralogs differ greatly. Paralogous snoRNA sequences in the same genome group together in evolutionary trees over their orthologs from the other genome (Chen et al., 2008). In combination, this suggests that the clusters emerged after the divergence of the two lineages

In placental mammals, the majority (~70%) of box C/D snoRNAs are present as single-copy genes and the remainder are present at low copy number (2-3 copies) (Makarova and Kramerov, 2009). However, in other non-placental vertebrates up to 60%

of snoRNA genes have 2 or more copies. Interestingly, a collection of snoRNAs that are conserved in mammals can be present at very high copy number – 50 isoforms were found in rhesus monkeys, 354 in mice and 1979 copies in the platypus (Zhang et al., 2010). Platypus shows incredible snoRNA mobility, with one snoRNA family present at more than 40,000 copies. Most of the gene duplications are species-specific, suggesting that most snoRNAs present in high copy number were replicated after the speciation of mammals.

It is possible that snoRNA duplication in eukaryotes is even more common than it appears. Due to functional redundancy, snoRNA copies are under less selective pressure and can mutate without consequence. Extensive sequence mutation over time may render snoRNA paralogs unidentifiable. Differences in the sequence of paralogous snoRNAs may therefore, indicate the relative timing of the gene duplication event (older duplications have longer time to mutate and diverge than more recent gene duplications).

A snoRNA can become a pseudogene if a) it contains mutations in the antisense guide region that render it non-functional, b) mutations disrupt conserved secondary structural elements required for function and/or assembly into RNPs, c) as the result of retroposition, the snoRNA gene no longer contains the necessary regulatory elements for expression and/or processing. Retroposition has recently been shown to be a mechanism of snoRNA evolution in mammals (Weber, 2006; Luo and Li, 2007; Schmitz et al., 2008). snoRNA copies can be generated via a "copy-and-paste" mechanism of retroposition. It is predicted that snoRNA retroposons (snoRTs) use the machinery of long interspersed elements (LINEs) for their mobilization. Most examined snoRTs are specific to a species or lineage, suggesting that snoRNA gene duplications by retroposition are recent events (Weber, 2006).

New functional snoRNA copies can also be generated by retroposition if the snoRT is targeted to an appropriate genomic location. As most mammalian snoRNAs are located in the introns of other genes, they do not possess their own regulatory elements. If a snoRT (of intronic origin) is targeted to an intergenic region or is inserted in the antisense orientation of a gene, it likely becomes a pseudogene. However, those positioned in the sense orientation of a new intron location could be processed into functional snoRNAs. Interestingly, snoRTs are frequently situated in introns of known genes in the proper orientation (Weber, 2006).

Double-guide box C/D RNAs (those that guide the methylation of two different RNA target sites), are more common in archaea than eukaryotes. One hypothesis is that over time, double-guide RNAs are evolving into single-guides (two independent snoRNAs to guide the modifications of one double-guide), an example of subfunctionalization. When the gene encoding a double-guide RNA duplicates, the gene copies can mutate in either the 5′ guide region (upstream of the D′ box) or 3′ guide region (upstream of the D box) so that each copy now guides one of the modification sites of the parental double-guide RNA (Figure 6). The alternative explanation is that two single function snoRNAs fused into one snoRNA capable of both functions. Given the apparent redundancy of snoRNA genes, the first explanation seems more likely to occur. However, the U85 snoRNA is an example of a single snoRNA that now targets both pseudouridylation and methylation events by possessing structural elements of both box C/D and H/ACA RNAs and thus may represent an example of a C/D snoRNA- H/ACA snoRNA fusion event (Jady and Kiss, 2001). Examples of the subfunctionalization of double-guide snoRNAs can be observed in a range of organisms including fungi (Qu et

al., 1995; Qu et al., 1999; Liu et al., 2009b), algae (Chen et al., 2008), vertebrates (Shao et al., 2009) and invertebrates (Yuan et al., 2003).

As previously mentioned, many snoRNA gene copies retain sequence identity in regions required for functionality (functionally redundant) and if enough mutations occur in these regions, the snoRNA copy is often non-functional (pseudogenes). If, however, sequences diverge in functional regions that allow the RNA to target a novel site for modification, neofunctionalization has occurred. For example, in the algae *C. reinhardtii* there is a family of snoRNAs encoded by multi-copy genes. Some of the mature snoRNAs are double-guides, while others have lost this ability and can only guide a single modification. Interestingly, two snoRNA copies belonging to this family retain the 5′-guide region but the 3′-guide region has mutated to guide a novel modification site not targeted by the other copies (Chen et al., 2008). A possible explanation is that after a snoRNA gene duplication event in *C. reinhardtii*, sequence mutations in the 3′-guide region changed the functionality of these snoRNAs so they were capable of guiding a novel rRNA modification. Similar examples have been observed in yeast (Zhou et al., 2002) and the filamentous fungi *Neurospora* (Liu et al., 2009b). In plants (*A. thaliana*) and nematodes (*C. elegans*) snoRNA genes are often present as clusters, which can themselves be duplicated several times. Analysis of clustered snoRNA copies reveals that some guide adjacent rRNA modification sites (Barneche et al., 2001; Brown et al., 2001; Zemann et al., 2006). Gene duplication, followed by sequence divergence in the guide regions, allows the snoRNA copies to target nearby rRNA sites (see Figure 15 and 3.4 for further details).

In addition to gaining new functions through altered base-pairing to novel target sites, neofunctionalization of snoRNAs can also occur if sequence mutations alter protein binding. Since snoRNAs function within RNP complexes, the recruitment and binding of protein partners depends on conserved sequence and structural motifs. Therefore, if sequence mutations alter these elements, novel protein components could bind and a novel snoRNA function could evolve. Vertebrate telomerase RNA is likely an example of such an evolutionary scenario. This telomerase RNA contains the features of snoRNAs that function as stability and localization elements, while not playing a role in RNA maturation processes. The 3′ half of the telomerase RNA in vertebrates resembles a box H/ACA snoRNA (Mitchell et al., 1999), including the 'hairpin-hinge-hairpin' secondary structure and the conserved box elements; however, other conserved sequence and structural elements that are not components of conventional snoRNAs are also located in the 3′ domain (Chen et al., 2000). The snoRNA domain is essential for telomerase function *in vivo*, where it is involved in the subcellular localization of telomerase RNA to the nucleus (Lukowiak et al., 2001). Vertebrate telomerase and snoRNAs also associate with evolutionarily conserved proteins (Dez et al., 2001; Pogacic et al., 2000), providing further evidence of an evolutionary relationship between these two unique classes of RNA.

Due to the high level of multi-copy snoRNA genes, it is apparent that novel snoRNAs are often the result of an initial gene duplication event, followed by sequence divergence. Alternative processing as a mechanism of snoRNA evolution has also been documented in mammalian cells. In the gene encoding a box C/D snoRNA, a single-nucleotide substitution in the region of the K-turn motif (see 1.2.2.1) results in the

inclusion of downstream intronic sequence into the mature snoRNA, producing a longer snoRNA with a proper K-turn (Mo et al., 2013). Through recruitment of intronic sequence into the mature snoRNA structure, a novel ncRNA evolved. While the exact function of this novel ncRNA has not yet been determined, it does show potential base-pairing complementarity to cellular RNAs, including rRNA and mRNA. The novel RNA is specific to rats and is not found in other mammals (including mice), suggesting it occurred via a recent evolutionary event after the divergence of mice and rats. This mechanism is different than the more typical mechanism of evolution by gene duplication and may explain the emergence of sequence-related snoRNAs that show length variation, relative to the canonical structure.

1.5 *Euglena gracilis* as a model system to study snoRNAs

The majority of snoRNA studies focus on archaea and model eukaryotes such as animals, plants and fungi. However, the single-celled eukaryotes collectively known as protists are an incredibly diverse group that represent most of the phylogenetic diversity in the eukaryotic domain. *Euglena gracilis* is a unicellular, freshwater flagellate that is capable of both autotrophy and heterotrophy (Ogbonna et al., 2002). Euglenids are a member of the phylum Euglenozoa, which is part of the eukaryotic supergroup Excavata (Adl et al., 2012). Euglenids share common ancestry with another group within the Euglenozoa, Kinetoplastea (which includes trypanosomes), where snoRNAs have been extensively studied (Uliel et al., 2004; Liang et al., 2005). Some euglenid species, including *E. gracilis*, possess secondary plastids that arose via secondary endosymbiosis of a green algae (Gibbs, 1978; Turmel et al., 2009). The nuclear genome of *E. gracilis* has yet to be sequenced and very little is known about its structure. There is evidence it

contains genetic material from its photoautotrophic endosymbiont (green algae) (Ahmadinejad et al., 2007), likely the result of endosymbiotic gene transfer as the endosymbiont evolved into an organelle (Timmis et al., 2004). In addition, there are genes of non-green algal origin, possibly the result of lateral gene transfer between eukaryotes (Maruyama et al., 2011). Based on DNA re-association kinetic studies, the *E. gracilis* genome is estimated to be highly repetitive and approximately 1.36 Gb in size (Rawson et al., 1979). Telomere hybridization experiments estimate there are at least 42 linear chromosomes (Dooijes et al., 2000).

The rRNA genes in *E. gracilis* are encoded as monomers on an autonomously replicating, extrachromosomal circular rDNA that can range from 800 to 4000 copies per cell (Ravel-Chapuis et al., 1985; Greenwood et al., 2001). Interestingly, the large subunit (LSU) rRNA is naturally fragmented into 14 discrete pieces (compared to 2 in most other eukaryotes) after transcription (Schnare et al., 1990; Schnare and Gray, 1990). *E. gracilis* rRNA is also the most extensively modified of any examined eukaryote to date, containing 211 2′-O-methylations and 116 pseudouridine modifications (Schnare and Gray, 2011). The LSU is more extensively modified than the non-fragmented SSU, which has a similar number of modifications as its human counterpart. The modifications are predicted to help stabilize the highly fragmented LSU during ribosome assembly (Schnare and Gray, 2011). A unique rRNA processing pathway that includes numerous cleavage and modification events, suggests that *E. gracilis* is an excellent organism in which to study snoRNAs. As parts of its genome are highly repetitive and many genes have multiple copies, *Euglena* also presents a unique opportunity to study snoRNA evolution. Furthermore, because of its position on the eukaryotic tree and relationship to

kinetoplastid protists in which snoRNAs have already been characterized (namely trypanosome species), *Euglena* is also a phylogenetically important organism in which to characterize snoRNAs.

Previous studies have isolated snoRNAs in *E. gracilis* through immunoprecipitation with antibodies to snoRNP protein components (Russell et al., 2004, 2006). *Euglena* box C/D snoRNAs are typically smaller and more uniform in size compared to other eukaryotes, more closely resembling those found in archaea (Russell et al., 2006). Of the first 70 characterized methylation-guide snoRNAs, there were surprisingly only two double-guide RNAs. *Euglena* box C/D RNAs also do not form the typical extended terminal stem structure (Figure 3a), which may be a general feature of some protist snoRNAs as those characterized in *T. brucei* (Liang et al., 2005) and *D. discoideum* (Aspegren et al., 2004) also lack this motif. Furthermore, the C′ and D′ boxes of *Euglena* snoRNAs are more degenerate than the C and D box elements such that they usually do not form a K-turn-like motif. For many snoRNA species, unique isoforms were identified, suggesting that many snoRNAs are encoded as multi-copy genes in *E. gracilis*.

As previously mentioned, *Euglena* rRNA is extensively modified and some regions (such as LSU species 6) contain densely clustered $O^{2'}$-methylated sites (Schnare and Gray, 2011). Some of the snoRNAs that guide these densely clustered modification sites have been characterized (Figure 7). It is clear that the guide regions and base-pair interactions of these snoRNAs overlap and yet, methylation at each site is highly efficient (there is no indication of only partial methylation at any site when examining the cellular

rRNA pool) (Russell et al., 2006), suggesting a high degree of coordination is required for modification at these positions.

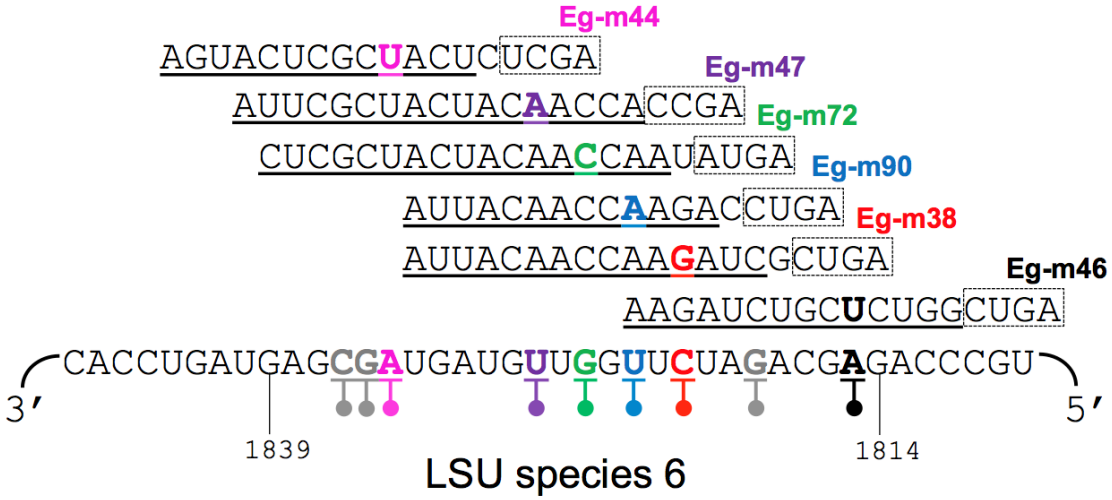


Figure 7. Clustered 2′-O-methylation sites in *Euglena gracilis* rRNA and their associated modification-guide snoRNAs. The region of LSU species 6 between position 1814 and 1839 contains densely clustered $O^{2'}$-methylated sites. The guide region of each box C/D snoRNA (labelled Eg-m) predicted to guide a specific modification site are shown in coordinating colors. The D or D′ box element is indicated with a hatched box and the nucleotides that base-pair to the rRNA are underlined. Methylated rRNA sites are underlined and highlighted with a filled circle. Eg-m38 – Eg-m47 identified in (Russell et al., 2006). Figure adapted from (Russell et al., 2006).

Pseudouridine-guide RNAs were also identified in *E. gracilis* by immunoprecipitating Cbf5p-containing complexes from cellular extracts (Russell et al., 2004). Due to technical challenges, only four snoRNAs that guide ψ modifications were initially identified; all four are single-stem RNAs (as opposed to the typical double-stem structure observed in other eukaryotes) and contain an 'AGA' box sequence element 3 nt from the 3′ end, features also characteristic of trypanosome ψ-guide RNAs (Liang et al., 2005). Interestingly, archaeal single stem ψ-guide sRNAs also often contain an AGA box in place of the ACA box (Tang et al., 2002). One unique triple-stem box H/ACA snoRNA was identified in the *E. gracilis* study that does not appear to guide any mapped ψ rRNA modification (Russell et al., 2004). Large, triple-stem box H/ACA RNAs have also been

identified in archaea, yeast and some vertebrates. The function and genomic organization of this large *E. gracilis* snoRNA was not determined.

The genomic organization of *E. gracilis* snoRNAs was not well characterized. Preliminary limited evidence indicated that some genes encoding both classes of snoRNAs can be present as clustered, tandem repeats (Russell et al., 2004). However, the prevalence of this organization for the majority of *E. gracilis* snoRNA genes, how expansive the clusters are, and how the genes are expressed, was unknown.

1.6 Functional characterization of snoRNAs using RNA interference

The number of identified ncRNAs has increased exponentially in the past decade. While some ncRNAs have well established cellular roles, the function of others remain unknown. Functional characterization of protein-coding genes often involves 'knocking-down' the mRNA, to prevent protein expression. This approach utilizes the RNA interference (RNAi) pathway, where 19-21 nt-long RNA duplexes (siRNAs), or the precursors that produce them, are introduced into the cell and associate with the RNA induced silencing complex (RISC); this results in sequence specific degradation of the complementary RNA target (Hannon, 2002; Meister and Tuschl, 2004; Tijsterman and Plasterk, 2004). RNAi typically occurs in the cytoplasm where mature mRNA species reside; however, nuclear RNAi pathways have also been reported (Meister, 2008). Using siRNAs, the nuclear ncRNA 7SK has successfully been silenced in human cells (Robb et al., 2005). Furthermore, in various trypanosome species (*Leptomonas collosoma, Leishmania major*, and *Trypanosoma brucei*), RNAi was used to knock down mature snoRNA species (Liang et al., 2003; Gupta et al., 2010). In these trypanosome studies, siRNAs were not introduced into the cells, rather, target-specific double-stranded RNA

(dsRNA) molecules were expressed from plasmids. As the number of orphan snoRNAs steadily increases, and more predicted novel functions and RNA targets continue to be uncovered, developing snoRNA inactivation/depletion methods, such as nuclear RNAi, will be important for the functional characterization of snoRNAs (and other ncRNAs).

1.7 Objectives

The overall objective of my research project was to further explore eukaryotic snoRNA diversity in terms of genomic organization, gene expression strategies and evolutionary relationships. Protist genomes remain largely unexplored in terms of snoRNA coding capacity (and genome structure in general), yet this group of eukaryotes show incredible phylogenetic diversity. My research focused on snoRNA characterization in the protist organism *Euglena gracilis.* While extensively used as a model laboratory organism, very little is known about its genome structure and gene expression strategies.

The specific objectives of my studies were as follows: **1)** Characterize novel snoRNA genes and determine their genomic arrangement in *E. gracilis*. Based on the number of modification sites in the *E. gracilis* rRNA, the previously identified snoRNAs only represent a small fraction of the predicted total number in this organism. Preliminary evidence suggested snoRNA genes in *E. gracilis* may be arranged as clusters and therefore, further identification of snoRNAs and characterization of contiguous genomic fragments will determine the prevalence of such an arrangement. **2)** Determine the mode of expression of snoRNA genes in *E. gracilis*. Based on the predicted clustered arrangement, I hypothesized that the snoRNA genes may be expressed as polycistronic transcripts, similar to what is observed in plants and trypanosome species. **3)** Study the evolution and appearance of new snoRNA genes and their target modification sites in *E.*

gracilis. An understanding of the evolutionary mechanisms contributing to the unusually large collection of snoRNAs could allow us to deduce a relationship between snoRNA gene organization and the rRNA modification patterns observed in this organism. **4)** Develop experimental and computational approaches (‘RNomics') to further characterize snoRNAs (and other ncRNAs) in *E. gracilis.* This organism is predicted to have a large, repetitive genome and may contain other numerous (and potentially novel) ncRNA species. **5)** Characterize the function of orphan snoRNAs in *Euglena* using an RNAi-based approach. By knocking down target RNA species and observing the resulting cellular effects, the functions of orphan snoRNAs may be inferred. Additionally, mechanisms and timing of snoRNA targeting of densely clustered modification rRNA regions could be examined with selective snoRNA species depletion.

Chapter 2: Materials and Methods

2.1 Growth conditions of *Euglena gracilis*

Euglena gracilis strain Z (wild-type, photosynthetic) was cultured in standard growth media (Russell et al., 2004) for optimal photoheterotrophic growth with media pH adjusted to 5.5 using phosphoric acid (Table 4). The cells were grown at room temperature in abundant sunlight and large culture volumes (>1 L) were gently agitated on a stir plate. Cell growth was recorded (OD_{600} and cell density) over a period of 14 days.

Table 4. *Euglena gracilis* wild-type media for photosynthetic growth

Trace Components	**Concentration**
$CaCl_2$	180 µM
$FeSO_4 \cdot 7H_2O$	40 µM
$MnCl_2 \cdot 4H_2O$	14 µM
$CoCl_2 \cdot 6H_2O$	12 µM
$ZnSO_4 \cdot 7H_2O$	8 µM
Sodium Molybdate	1 µM
$CuSO_4 \cdot 5H_2O$	0.8 µM
Thiamine Hydrochloride	2 µM
Vitamin B12	2.95 nM
Boric Acid	7.3 µM
Ethanol	0.2% v/v
Salt Components	
$(NH_4)_2SO_4$	7.6 mM
KH_2PO_4	2.2 mM
$MgSO_4 \cdot 7H_2O$	1.6 mM
Sodium Citrate Dihydrate	2.7 mM
Growth media modified from Russell *et al.* 2004.	

2.2 Isolation of *Euglena gracilis* genomic DNA and RNA

Cells from a 1.4 L culture (grown as described in 2.1) were collected at mid-late log phase (7-10 days of growth, OD_{600} 0.8-1.0) by centrifugation at 3220 *g* for 20 min at 4°C. The cells were then resuspended in 25 mL of resuspension buffer (RB) (25 mM EDTA-Tris, pH 8.5) and collected again by centrifugation at 1160 *g* for 10 min. The

pellet was resuspended in RB to a total volume of 10 mL. To lyse the cells, 2.5 mL of 25% SDS solution was added with gentle mixing by inversion to form a uniform suspension (2-3 min). The volume was adjusted to 25 mL with RB and vortexed until homogenous. To the cell lysate, 4 mL of 8 M sodium perchlorate was added and mixed gently by inversion. The nucleic acids were extracted twice with chloroform:isoamyl alcohol (24:1) and then precipitated with an equal volume of isopropanol. The high molecular weight DNA was spooled out using a sealed Pasteur pipette and washed with 5 mL of 80% ethanol. The DNA was collected by centrifugation at 515 *g* for 3 min and then air dried for 10 min. The dried pellet was resuspended in 10 mL of TE buffer (10 mM Tris-HCl pH 7.6, 0.1 mM EDTA) then 1 mL of 3 M sodium acetate solution was added, followed by 2 phenolic (pH 7) extractions. Genomic DNA was precipitated from the aqueous phase with ethanol and the dried pellet was resuspended in 3 mL of TE buffer. A fraction (2 mL) of the purified DNA was stored long-term in ethanol. The remaining 1 mL of DNA was treated with 165 U of RNase A (Sigma) and incubated at 37°C for 2.5 hrs, followed by phenolic extraction and ethanol precipitation. The resulting DNA pellet was resuspended in 1 mL of TE buffer and quantified.

Total *E. gracilis* RNA was extracted using TRIzol®, following the manufacturer's protocol with the following modifications. 1) Cells from 25 mL of culture (grown as described in 2.1) were collected at mid-log phase (OD_{600} 0.5-0.8) by centrifugation at 10,000 *g* for 3 min at room temperature. Cells were counted using a haemocytometer and 1 mL of TRIzol® per 1-2 × 10^7 cells was added. The cells were lysed by repeated pipetting and then incubated at room temperature for 1-2 hrs. 2) The final RNA pellet was resuspended in 100 µL of TE buffer. The purified RNA was treated with 20 U of DNaseI (NEB) in a 1 mL reaction volume and incubated at 37°C for 20 min, followed by a

phenol:chloroform (1:1) extraction and two subsequent chloroform extractions. The purified RNA was precipitated with isopropanol, washed with ethanol and the dried pellet was resuspended in 100 µL of TE buffer and quantified.

2.3 snoRNA gene identification and organization

2.3.1 PCR-mediated genomic DNA amplification (based on Moore & Russell, 2012)

Oligonucleotides were designed based on previously biochemically-isolated and sequenced snoRNA species (Russell et al., 2006) (for oligonucleotide sequences and amplification pairs, see Supplementary Table S1). Standard PCR reactions contained: 1 U of Taq DNA polymerase (NEB), 0.4 µM each primer, 400 µM dNTPs, 1X Standard Taq Reaction Buffer (NEB) and 100 ng of *E. gracilis* genomic DNA as a template. Cycling parameters were: 94°C (5 min); 35 cycles of 94°C (30 sec), 50°C * (30 sec) and 72°C (1 min); followed by 72°C (5 min) where * indicates variable annealing temperatures altered for different primer combinations. Some genomic DNA amplifications were more successful when 5% DMSO (v/v) was added. If multiple different sized products were generated in a single PCR reaction, they were individually gel-extracted using a MinElute Gel Extraction Kit (Qiagen), following the protocol provided by the manufacturer. Purified PCR products were then cloned into either pCR2.1 Topo vector (Invitrogen) using the Topo TA Cloning Kit (Invitrogen) according to the manufacturer's protocol or pJET1.2/blunt vector (Fermentas) using the CloneJET PCR Cloning Kit (Fermentas) according the manufacturer's protocol. Automated DNA sequencing of cloned PCR products was performed by Macrogen Corp USA.

2.3.2 BAC genomic DNA library screening

A bacterial artificial chromosome (BAC) library containing 15-100 kb *Euglena gracilis* genomic DNA inserts was constructed by Bionexus Inc.

2.3.2.1 Isolation of BAC DNA

Transformed *E. coli* cells were cultured in LB media supplemented with 12.5 µg/mL chloramphenicol. BAC DNA was isolated using the BACMAX DNA Purification Kit (epicentre®) according to the manufacturer's protocol with the following modifications. After digestion with RiboShredder RNase Blend®, the DNA was extracted with phenol:chloroform (1:1) followed by two chloroform extractions and then precipitated with ethanol and quantified.

2.3.2.2 PCR-mediated BAC library screening

In order to initially screen the BAC library for *E. gracilis* snoRNA-coding regions of interest, BAC library DNA was isolated from the mixed BAC host culture. Based on the provided library titer (1×10^7 cfu/µL), ~10,000,000 cfu were grown in liquid LB media and BAC DNA was isolated as described in 2.3.2.1. Alternatively, following the identification of positive hybridization signals from library screening (see 2.3.2.3), a single positive colony was used to inoculate liquid LB media and BAC DNA was isolated as described in 2.3.2.1. The isolated BAC DNA (~200 ng) was then used as template in PCR, using standard conditions outlined previously and oligonucleotide amplification pairs for specific snoRNA genes (Table S2), to verify the presence of that gene in the library or a specific isolated BAC.

2.3.2.3 Isolation of BACs containing snoRNA target sequences

Amersham Hybond-XL (GE Healthcare) membranes were placed individually on a total of 10 LB agar plates supplemented with 12.5 µg/mL chloramphenicol.

Approximately 25,000 cfu (per plate) were spread, following dilution in LB, on the surface of each membrane and grown at 37°C for 16 hrs (master plates). Colonies from these master plates were then transferred to replica membranes (Sambrook and Russell, 2001) and then all the membranes were returned to their original LB plates and incubated for 2.5 hrs at 37°C. The master plates were sealed and stored at 4°C.

Colonies on replica membranes were lysed and the DNA was fixed to the membranes (based on the procedure of (Sambrook and Russell, 2001)) with the following modifications. The optional step of exposing the membranes to 10% SDS (w/v) was performed. The membranes were only treated once with neutralizing solution (0.5 M Tris-Cl pH 7.4, 1.5 M NaCl) and the membranes were placed on a piece of 3MM Whatman® paper saturated with 2× SSPE (1× SSPE = 150 mM NaCl, 10 mM NaH_2PO_4, 1 mM EDTA) as opposed to floating them in a tray of solution. The membranes were then dried on 3MM Whatman® paper for 1 hr at room temperature. The DNA was cross-linked to the membrane by exposure to a hand-held UV lamp (254 nm) positioned one inch above each membrane for 10 min.

Radioactively-labelled hybridization probes were synthesized using PCR (Sambrook and Russell, 2001). The PCR labelling reactions contained: 2.5 U of Taq DNA polymerase (NEB), 1 µM each primer, 200 µM dTTP, dCTP, and dATP, 2 µM dGTP, 1X ThermoPol Buffer (NEB), 100 ng of *E. gracilis* genomic DNA and 50 µCi of [α-^{32}P] dGTP (3000 Ci/mmol). Cycling parameters were: 95°C (5 min), 35 cycles of 95°C (30 sec), 50°C * (45 sec) and 68°C (1 min), followed by 68°C (7 min) where * indicates variable annealing temperature that was altered for different primer combinations. The PCR products from amplification of different snoRNA sequences were usually pooled together (~4 per pool to optimize initial screening) and purified by spun-column

chromatography to separate labelled DNA from radioactive precursors (Sambrook and Russell, 2001). Briefly, a 1 mL disposable syringe was plugged with glass wool and filled with Sephadex G-50 resin (GE Healthcare), which was hydrated with TEN buffer (10 mM Tris-Cl pH 8, 1 mM EDTA pH 8, 100 mM NaCl). After equilibration of the column with TEN buffer, the pooled PCR products were purified.

The radiolabelled PCR probes were hybridized to the fixed bacterial DNA on the membranes (Sambrook and Russell, 2001), with the following modifications. After soaking the membranes in 2× SSC (1× SSC = 150 mM NaCl, 15 mM sodium citrate), they were transferred to 150 mL of 6× SSC in a plastic container with a lid. After scraping bacterial debris from the membranes, they were placed in ~25 mL of hybridization solution (1% w/v BSA, 1 mM EDTA, 0.5 M phosphate buffer, 7% w/v SDS) in glass hybridization bottles. The membranes were gently agitated in a hybridization oven for 1.5 hrs at 68°C. The purified PCR probes (4.8×10^6 cpm – 6.4×10^6 cpm) were denatured at 100°C for 5 min and then rapidly chilled in ice water. The probes were added to the hybridization solution and incubated with agitation at 68°C overnight in the oven. The filters were then transferred to 150 mL of wash solution 1 (2× SSC, 0.1% w/v SDS) in plastic containers with lids. The filters were gently agitated on a rotating platform at room temperature for 5 min, after which the solution was discarded and the wash was repeated. The filters were placed back in hybridization bottles with 100 mL of wash solution 2 (1× SSC, 0.1% w/v SDS) and were incubated at 68°C with agitation. The filters were then air-dried on 3MM Whatman® paper and visualized by phosphorimaging with a GE Typhoon.

Filters with positive hybridization signals were aligned with the corresponding master membrane and colonies in the area of the signal were transferred and co-cultured

in 5 mL of liquid LB media with 12.5 µg/mL chloramphenicol at 37°C for 16 hrs. The overnight culture was then diluted 1000× and 1 µL of this dilution was plated using a new membrane for the second round of screening (fewer cfu to form well-isolated colonies). The entire hybridization procedure (as above) was repeated and individual colonies that aligned with positive hybridization signals were isolated. BAC DNA from the positive colonies was isolated as described in 2.3.2.1.

2.3.2.4 Characterization of isolated BACs

To sequence the extremities of the *E. gracilis* snoRNA genomic repeat inserts in the isolated BACs, primer walking (for oligonucleotide sequences see Table S3) was used with automated DNA sequencing performed by Macrogen Corp USA. BAC DNA (~5 µg) was then partially digested with either 20 U, 2 U, 1 U, 0.2 U, 0.1 U or 0.02 U of EcoRI , HindIII, NcoI, or SphI (NEB) at 37°C for 30 min. Alternatively, 10 µg of BAC DNA was completely digested with 20 U of NotI (NEB) at 37°C for 2 hrs. The digested DNA was ethanol precipitated and resuspended in 20 µL of water.

To determine the orientation of the *E. gracilis* genomic DNA insert, as well as the distance between the cloning site and the first snoRNA gene, PCR was performed using BAC DNA as template (for oligonucleotide pairs and sequences, see Table S4).

***2.3.3 Bioinformatic analysis* (based on Moore & Russell, 2012)**

SPIN v1.3 of the Staden 2003 Beta Version 1 software package was used to identify regions of complementarity between PCR-amplified genomic sequences (2.3.1) and the *E. gracilis* rRNA sequences (Schnare and Gray, 1990). Regions of significant sequence similarity were then compared to mapped $O^{2'}$-methylated modification sites on the rRNA (Schnare and Gray, 2011). A methylated rRNA nucleotide found at the expected position relative to a D or D′ box element (n+5 rule), as well as the presence of

all other conserved snoRNA box elements, formed the basis for identifying a box C/D snoRNA gene. This strategy exploited the complete *E. gracilis* modification mapping data available (Schnare and Gray, 2011) and by design is biased towards identifying box C/D snoRNA genes targeting rRNA sites (modification-guide snoRNAs), as opposed to snoRNAs with other cellular functions and/or targets. The amplified genomic regions were also searched for box H/ACA-like snoRNA coding regions by manual inspection, looking for conserved sequence box elements, the potential to form secondary structural features diagnostic of H/ACA RNAs and the ability of the predicted snoRNA to form a pseudouridine pocket targeting a mapped pseudouridine position in *E. gracilis* rRNA (Schnare and Gray, 2011).

To identify previously characterized *Euglena gracilis* snoRNA coding regions in isolated BAC DNA (2.3.2.3), BLASTN searches were performed. The genomic regions surrounding characterized snoRNA genes were also analyzed for novel snoRNAs, as just described.

To identify the U14 snoRNA homolog in *A. gambiae*, the conserved region of 18S rRNA that base-pairs to Domain A of U14 in humans, yeast and plants was located in *A. gambiae* 18S rDNA. A BLAST search was used to locate a region of the *A. gambiae* genome that could base pair to this region of the 18S rRNA. Potential U14 candidates were then manually inspected for the additional presence of a C and D box element at the correct relative position to the Domain A region.

2.4 snoRNA expression studies (based on Moore & Russell, 2012)

Antisense primers (2 pmole) were designed to anneal within different snoRNA coding regions. These were annealed to 1 µg of total *E. gracilis* RNA at 65^{o}C for 5 min,

followed by 47°C for 10 min. Using Superscript II RT (Invitrogen), cDNA was synthesized at 47°C for 45 min following the manufacturer's protocol. The cDNA (5 µL of the original 20 µL reaction) was then used as a template for PCR amplification employing standard conditions outlined above (for oligonucleotide sequences and amplification pairs see Table S5). RT-PCR products were cloned into pCR2.1 Topo vector (Invitrogen) or pJET1.2/blunt vector (Fermentas) and sequenced.

To verify expression and map the 3′ ends of the identified box AGA snoRNAs, 3′ RACE was performed as previously described (Russell et al., 2006) (for oligonucleotide sequences, see Table S6). The RACE products were cloned into pCR2.1 Topo vector or pJET1.2/blunt vector (Invitrogen) and sequenced.

All genomic snoRNA sequence information was deposited in GenBank. Cluster 2 GenBank accession no. JN051179; clusters 2.1-5.2 accession nos. JN051170-JN051178, respectively; clusters 6-34 accession nos. JN051180-JN051253, respectively. Individual snoRNA sequences and isoforms are annotated.

2.5 *Euglena gracilis* small RNA and capped RNA library synthesis

2.5.1 Library construction

E. gracilis total RNA (~112 µg) was resolved on a 15% denaturing polyacrylamide gel and RNA fragments less than 400 nt were excised. The gel-slices were crushed and soaked in an equal volume mixture of oligonucleotide elution buffer (OEB = 0.5 M ammonium acetate and 10 mM magnesium acetate) and phenol (pH 6.6), and then incubated on a rotating platform at 4°C overnight. The aqueous phase was extracted twice with chloroform and the RNA was ethanol precipitated with linear acrylamide carrier (Gaillard and Strauss, 1990).

Prior to affinity purifying 2,2,7-trimethyl guanosine (TMG)-capped ncRNAs, polyadenylated mRNA was removed from ~240 µg of *E. gracilis* total RNA using the PolyA Tract mRNA Isolation System III (Promega), following the manufacturer's instructions (except the mRNA depleted fraction was collected instead of the mRNA enriched fraction). The mRNA-depleted fraction was precipitated with isopropanol and the RNA was resuspended in 100 µL of NET buffer (50 mM Tris pH 7.5, 150 mM NaCl, 1 mM EDTA).

To enrich TMG-capped ncRNAs from mRNA depleted total RNA, 40 µL of Protein A Sepharose was washed once with 500 µL of DEPC-dH_2O and three times with 500 µL of NET buffer (beads were collected by centrifugation at 400 *g* for 1 min). To the washed beads, 50 µg of anti-TMG monoclonal antibody preparation (Santa Cruz Biotechnology) was added and the volume was adjusted to 500 µL with NET buffer and incubated on a mixer at 4^{o}C for 1 hr. To remove unbound antibodies, the beads were washed three times with 500 µL of NET buffer. The mRNA depleted total RNA (~180 µg) was added and the volume was adjusted to 500 µL with NET buffer. The binding reaction was incubated on a mixer for 14 hrs at 4^{o}C. The reaction was centrifuged at 400 *g* for 1 min and the supernatant was collected. The beads were then washed 10× with 500 µL of NET buffer (the supernatant was collected after each wash). The mixture was extracted with phenol:chloroform (1:1) and then twice with chloroform. The aqueous phase was precipitated with ethanol (and acrylamide carrier) and the final RNA pellet was resuspended in 20 µL of DEPC-dH_2O.

A poly-G tail was added to the 3′ ends of the size-selected or capped RNA (Rederstorff and Huttenhofer, 2011). The tailing reaction consisted of size-selected or 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). The reaction was incubated at 37°C for 60 min. Then, 20 µL of 3M sodium acetate (pH 5.2) was added and the reaction volume was adjusted to 200 µL with DEPC-dH_2O. The reaction was extracted once with phenol:chloroform (1:1), twice with chloroform and the aqueous phase was ethanol precipitated (with 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). The reaction was incubated at 37°C for 60 min and the RNA was extracted and precipitated as described above (no additional acrylamide carrier was added).

An RNA oligonucleotide linker was ligated to the 5′ termini of the TAP-treated RNA. 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 in an ice bath 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 reaction volume was adjusted and the RNA was extracted and precipitated as described above.

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), along with 500 µM dNTPs, was incubated with 10 µL of prepared RNA from the previous step at 65°C for 5 min and then immediately chilled on ice. Using Superscript II RT (Invitrogen), cDNA was synthesized at 47°C for 60 min following the manufacturer's protocol. The cDNA was then used as template for PCR amplification, using oligonucleotides designed to anneal to the 3′ poly-G tail and the 5′ linker sequence (see Table S7 for oligonucleotide

sequences). The PCR reactions contained: 1X Phusion HF buffer (Thermo Scientific), 200 µM dNTPs, 0.5 µM each oligonucleotide, 5 µL (of the 20 µL reaction) of cDNA, and 1 U of Phusion Taq Polymerase (Thermo Scientific). Cycling parameters were: 98ºC (2 min), 35 cycles of 98ºC (10 s), 60ºC (15 s), 72ºC (30 s), with a final extension at 72ºC (7 min). The PCR product was purified by gel-extraction using the crush and soak method described previously and cloned into pJET1.2/blunt vector as previously described. Transformed *E. coli* colonies were then used as template in PCR screening, using primers that anneal upstream and downstream of the PCR cloning site. Automated DNA sequencing of these PCR products was performed by Macrogen Corp USA.

2.5.2 Removal of large subunit rRNA fragments

Blocking primer sets with a C3 spacer (3 hydrocarbon) modification at the 3′ end were designed to anneal to the end of the added 5′ linker sequence + 5′ end of each LSU rRNA fragment (see Table S8 for blocker oligonucleotide sequences). At the PCR amplification step of library preparation, 5 pmole/µL of each blocking primer was added to the reaction to prevent amplification of the unwanted rRNA species. The final resulting PCR-generated cDNA library was gel-purified using either the crush and soak method or using the E.Z.N.A ® Cycle-Pure Kit (Omega) and then submitted to Genome Quebec for further preparation and high-throughput sequencing using the Illumina MiSeq platform.

2.5.3 Bioinformatic analysis

The Illumina MiSeq sequence reads were first sorted based on the presence of the 5′ linker sequence that was added during library synthesis, 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. 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 using the Trim Ends tool in Geneious. The UCLUST algorithm (a component of the USearch software package (Edgar, 2010)) was used to cluster related sequences together based on pair-wise alignments, using an identity threshold of 0.8. To remove known *Euglena* RNAs from the newly formed sequence clusters, a database of *E. gracilis* snoRNAs, rRNA, and tRNAs (obtained from GenBank) was created. The UBlast algorithm (a component of USearch (Edgar, 2010)) was used to find hits to the database in the library sequences (using an expectation value (E-value) of 1e-9) and matches to the database were subsequently removed.

To characterize novel snoRNAs, 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' (Dsouza et al., 1997). Briefly, a consensus pattern was made based on all previously identified *Euglena* snoRNAs including size, box elements, and secondary structure potential. The Scan for Matches 'hits' were then manually screened and analyzed as previously described (see 2.3.3). Sequences that maintained conserved features of snoRNAs but did not base pair to any known modified rRNA site were also noted.

2.6 RNA silencing in *Euglena gracilis*

2.6.1 Optimization of electroporation conditions

The electroporation conditions for introducing double-stranded RNA (dsRNA) into *Euglena gracilis* cells were optimized based on previous studies (Iseki et al., 2002;

Daiker et al., 2010). The volume of cells, voltage of the pulses and the incubation conditions were individually tested (Table 5). In each treatment, ~10 µg of fibrillarin dsRNA (see 2.6.2.1) was introduced into the cells and effectiveness of silencing was determined based on the level of inhibition of cellular growth after treatment as compared to control treatments in which cells were subjected to identical electroporation conditions minus dsRNA.

2.6.2 dsRNA-induced RNA silencing

2.6.2.1 dsRNA synthesis

Templates for *in vitro* transcription were synthesized by PCR or RT-PCR (as per 2.3.1 or 2.4, respectively) with the incorporation of a T7 RNA polymerase promoter sequence into the PCR product to allow either sense or antisense RNA strand synthesis (see Table S9 for oligonucleotide sequences). The PCR products were gel-purified using the E.Z.N.A. Gel-Extraction Kit (Omega) and then *in vitro* transcribed into single-stranded RNA using the MEGAscript kit (Ambion), according to the manufacturer's protocol. To generate dsRNA, equimolar amounts of sense and antisense RNAs were annealed together by heating to 80^{o}C for 2 min, and then slowly cooling by 2^{o}C every 30 sec until the temperature reached 4^{o}C.

2.6.2.2 Electroporation of Euglena gracilis with dsRNA

Cells were grown to mid-log phase (as described in 2.1) and counted using a haemocytometer. The cells were collected by centrifugation at 5000 *g* for 5 min at room temperature and resuspended in an appropriate volume of standard growth media to allow for 100 µL electroporation sample volumes, containing 3-8 × 10^5 cells per treatment. The cells were transferred to an ice-cold electroporation cuvette (0.4-cm-gap) and 10-40 µg of dsRNA (or just antisense RNA) was added. Immediately after the addition of RNA, the

cells were electroporated using a Bio-Rad Gene Pulser at 1.2 kV, 25 μF and 200 Ω, using 2 pulses, 10 sec apart. After electroporation, the cells were incubated on ice for 10 min and then transferred to 25 mL of fresh growth media. Control cells were electroporated identically with addition of water instead of dsRNA. In a second control, cells were electroporated using the same conditions but with non-specific dsRNA. Post-electroporation, cells were sub-cultured every 5-7 days as necessary. Total RNA was isolated from cells electroporated with dsRNA at various times post-treatment, as described in 2.2

2.6.3 2´-O-methylation mapping

To determine whether a specific snoRNA had been silenced in *E. gracilis* after treatment with dsRNA, the methylation status of its target 2´-O-methylation site was assessed.

2.6.3.1 Primer extension reactions under low dNTP concentration

Antisense oligonucleotide primers were designed to anneal to rRNA regions immediately downstream of 2´-O-methylation sites of interest (for oligonucleotide sequences see Table S10). The primers were first gel-purified and then radioactively 5´end-labelled using 10 μCi of [γ-^{32}P] ATP (3000 Ci/mmol) and 10 U of polynucleotide kinase (PNK, Lucigen) at 37^{o}C for 30 min.

Labelled oligonucleotides (~5 pmole per reaction) were mixed with ~5 μg of *Euglena* total RNA and incubated at 65^{o}C for 5 min, followed by 47^{o}C for 10 min and then room temperature for 5 min. Primer extension with 200 U of SuperScript II RT was performed at 47^{o}C for 45 min, in the presence of decreasing amounts of dNTPs (1 mM, 0.04 mM or 0.004 mM).

2.6.3.2 RTL-P

Reverse transcription at low dNTP concentrations followed by PCR (RTL-P) (Dong et al., 2012) was also utilized to detect 2´-O-methylation sites. Primers were designed for reverse transcription downstream of targeted 2´-O-methylation sites, as well as for semi-quantitative PCR in the region surrounding the modified sites (for oligonucleotide sequences see Table S11). Reverse transcription reactions were performed as previously described, under normal (1 mM) and low (1-1.5 µM) dNTP concentrations. PCR reactions were performed using the standard conditions outlined previously, however, the cycle was only repeated 12 times. PCR was performed using primers R/F_D and R/F_U for each methylation site assayed (refer to Figure 23b).

Chapter 3: Results

3.1 Identification and characterization of snoRNA gene clusters in the *E. gracilis* genome (based on Moore & Russell, 2012)

Previous studies in *E. gracili*s had identified a large collection of modification-guide box C/D and a few box AGA (H/ACA-like) snoRNA sequences mainly through biochemical isolation (Russell et al., 2004, 2006); however, only 4 of these snoRNAs had been characterized at the gene organization level where they were found to be clustered with other snoRNA coding regions. This snoRNA genomic analysis was significantly extended by employing a PCR-based approach. If snoRNA genes are clustered, and if these clusters are tandemly repeated, then appropriately designed oligonucleotides annealing to the ends of individual snoRNA sequences can amplify genomic sequence between tandem snoRNA gene copies (Figure 8). The first two initial objectives were to: a) determine how prevalent the previously described clustered snoRNA gene arrangement is in *E. gracilis* and b) to identify new snoRNA-encoding genes that may reside in these amplified genomic segments.

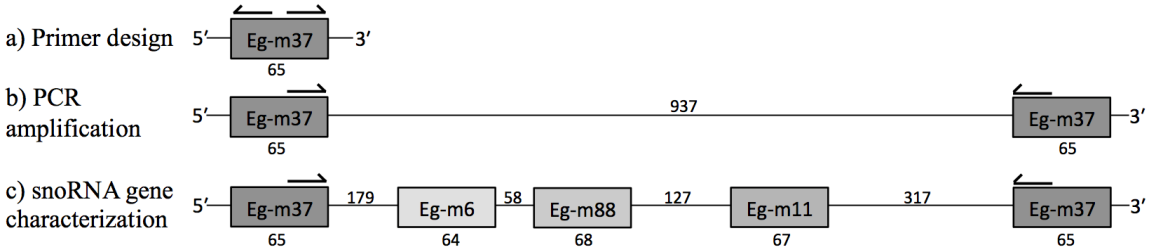
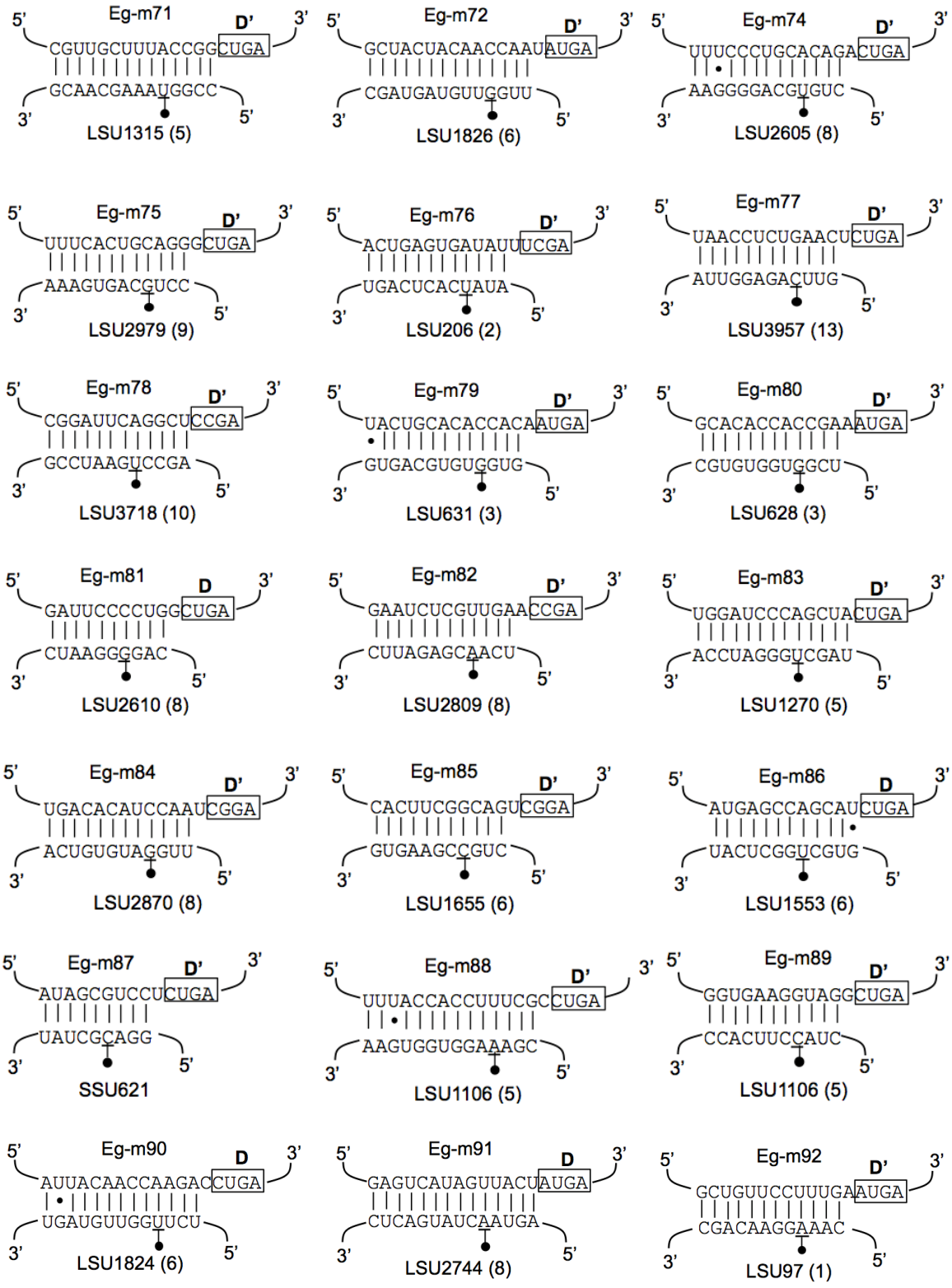
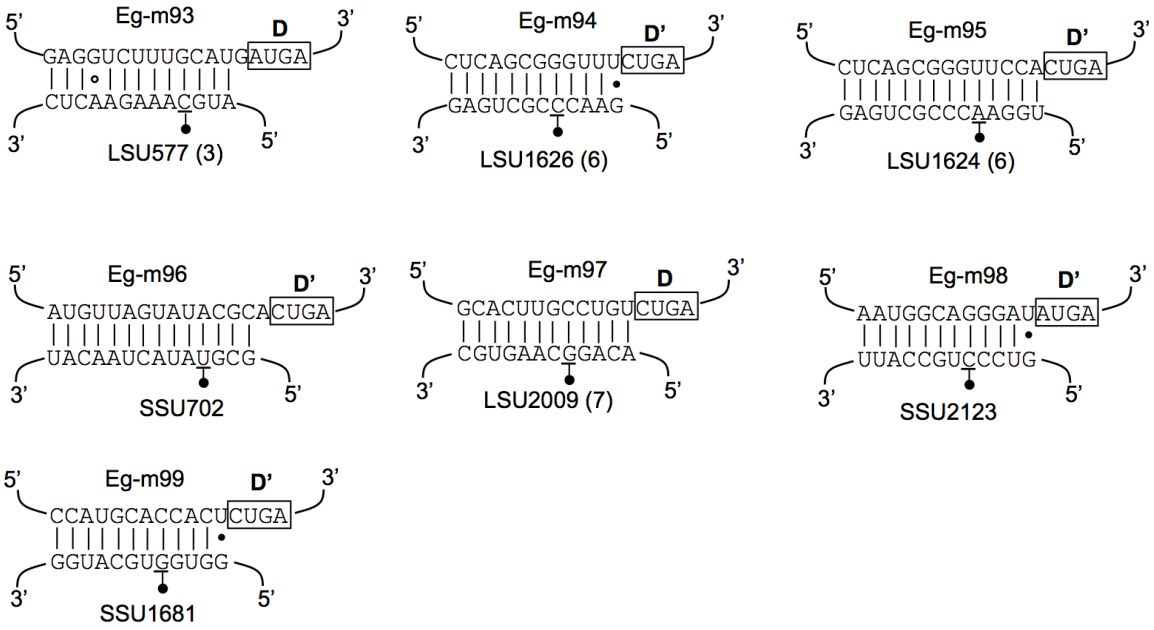


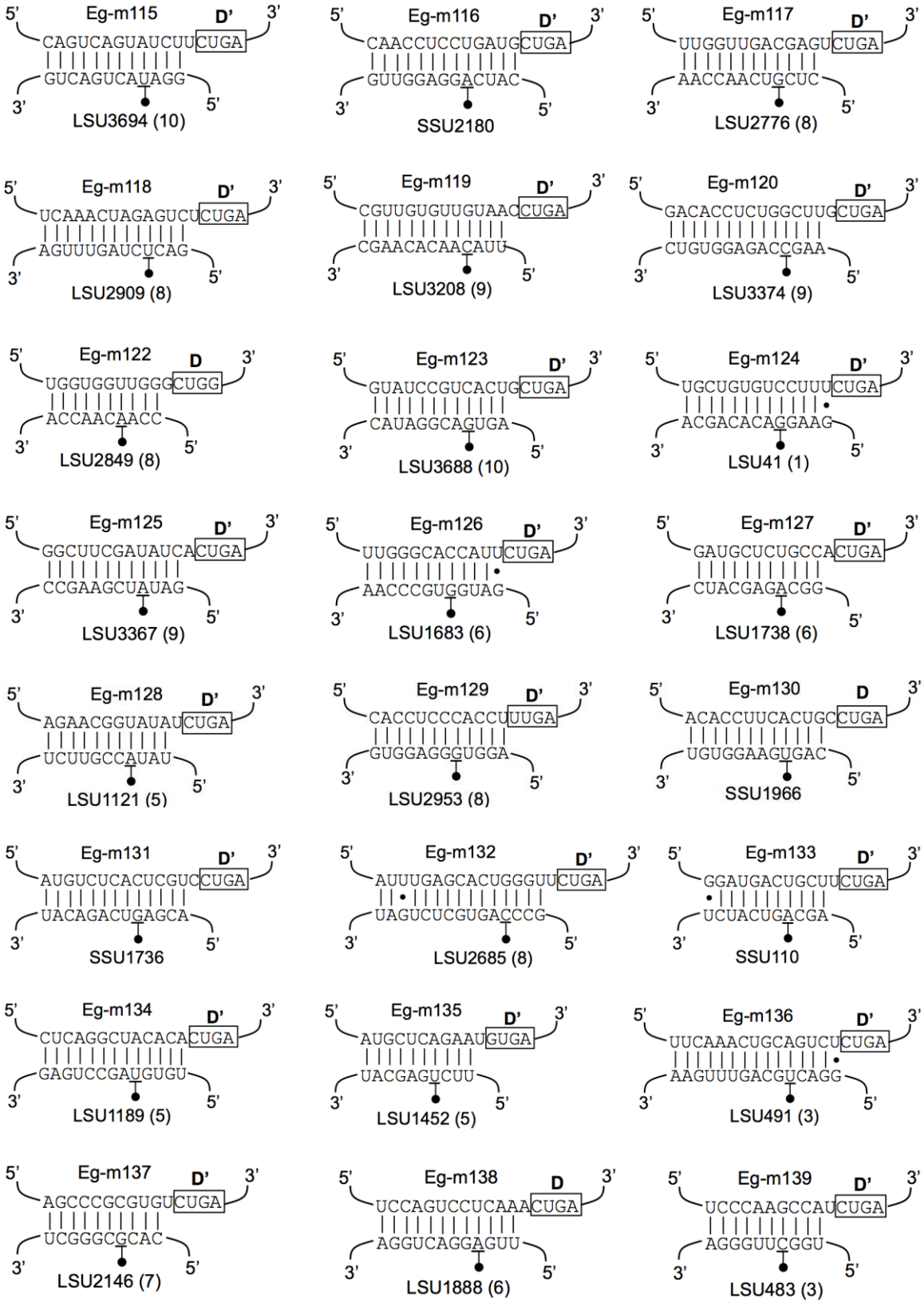
Figure 8. PCR based strategy to characterize snoRNA gene clusters in *E. gracilis*. **a)** Oligonucleotides were designed to anneal to the 5′ and 3′ ends of identified snoRNA sequences. **b**) If a tandem copy (or isoform) of the snoRNA is nearby, a genomic product can be amplified using PCR. **c)** The intervening sequence between tandem gene copies can then be analyzed for additional snoRNA coding regions. Arrows indicate primers, boxes indicate snoRNA genes, and lines represent intervening sequences. The distance (in number of base pairs) between snoRNA genes is indicated above the lines and size of snoRNA genes is indicated below the boxes. Example shown is cluster 7.2. Figure not drawn to scale (from Moore & Russell, 2012).

Of the 79 previously identified unique snoRNA species on which such PCR primers could be designed, at least 52 reside within tandemly repeated snoRNA clusters, as evidenced by successful genomic DNA amplification. It was previously observed that some *Euglena* snoRNA species consist of sequence-related isoforms (Russell et al., 2006; Charette and Gray, 2009) and therefore we are likely underestimating the frequency of tandem clustered arrangements. Those species whose closest gene copies (isoforms) contain the greatest sequence divergence in primer binding regions may not be efficiently amplified. In total, PCR-based amplification and sequencing of >100,000 nucleotides of non-overlapping *E. gracilis* genomic DNA, in snoRNA-encoding regions, was performed.

Using bioinformatic analysis, 124 unique snoRNA sequences were characterized, representing 29 previously unknown box C/D snoRNA species, 9 new box AGA Ψ-guide snoRNAs and various isoforms of this collection, as well as 48 new isoforms of previously characterized snoRNA species. A complete list of all snoRNA and isoform sequences is annotated in Supplementary Figure S1. For the newly identified snoRNA genes, we searched for regions of base-pairing complementarity to the *E. gracilis* rRNAs. All but one of the new snoRNAs has the ability to engage in substantial base-pairing interactions to a target region within the rRNA using an appropriate region of the snoRNA sequence (Figure 9 and Figure 10). In each case, a mapped modified nucleotide is present at the expected position (Schnare and Gray, 2011), based on the universally-conserved rules of snoRNA-targeted modification site selection. In some cases we did observe more limited and interrupted base-pairing potential of regions within genomic clusters to non-modified rRNA sites, but rarely were canonical snoRNA box elements also found in the expected relative positions.


Eg-m71
D'
5'
3'
CGUUGCUUUACCGGCUGA
GCAACGAAAUGGCC
LSU1315 (5)
Eg-m72
GCUACUACAACCAAUAUGA
CGAUGAUGUUGGUU
LSU1826 (6)
Eg-m74
UUUCCCUGCACAGACUGA
AAGGGGACGUGUC
LSU2605 (8)
Eg-m75
UUUCACUGCAGGGCUGA
AAAGUGACGUCC
LSU2979 (9)
Eg-m76
ACUGAGUGAUAUUUCGA
UGACUCACUAUA
LSU206 (2)
Eg-m77
UAACCUCUGAACUCUGA
AUUGGAGACUUG
LSU3957 (13)
Eg-m78
CGGAUUCAGGCUCCGA
GCCUAAGUCCGA
LSU3718 (10)
Eg-m79
UACUGCACACCACAAUGA
GUGACGUGUGGUG
LSU631 (3)
Eg-m80
GCACACCACCGAAAUGA
CGUGUGGUGGCU
LSU628 (3)
Eg-m81
D
GAUUCCCCUGGCUGA
CUAAGGGGAC
LSU2610 (8)
Eg-m82
GAAUCUCGUUGAACCGA
CUUAGAGCAACU
LSU2809 (8)
Eg-m83
UGGAUCCCAGCUACUGA
ACCUAGGGUCGAU
LSU1270 (5)
Eg-m84
UGACACAUCCAAUCGGA
ACUGUGUAGGUU
LSU2870 (8)
Eg-m85
CACUUCGGCAGUCGGA
GUGAAGCCGUC
LSU1655 (6)
Eg-m86
AUGAGCCAGCAUCUGA
UACUCGGUCGUG
LSU1553 (6)
Eg-m87
AUAGCGUCCUCUGA
UAUCGCAGG
SSU621
Eg-m88
UUUACCACCUUUCGCCUGA
AAGUGGUGGAAAGC
LSU1106 (5)
Eg-m89
GGUGAAGGUAGGCUGA
CCACUUCCAUC
LSU1106 (5)
Eg-m90
AUUACAACCAAGACCUGA
UGAUGUUGGUUCU
LSU1824 (6)
Eg-m91
GAGUCAUAGUUACUAUGA
CUCAGUAUCAAUGA
LSU2744 (8)
Eg-m92
GCUGUUCCUUUGAAUGA
CGACAAGGAAAC
LSU97 (1)


Eg-m93
D
5'
3'
GAGGUCUUUGCAUGAUGA
CUCAAGAAACGUA
3'
5'
LSU577 (3)
Eg-m94
D'
CUCAGCGGGUUUCUGA
GAGUCGCCCAAG
LSU1626 (6)
Eg-m95
CUCAGCGGGUUCCACUGA
GAGUCGCCCAAGGU
LSU1624 (6)
Eg-m96
AUGUUAGUAUACGCACUGA
UACAAUCAUAUGCG
SSU702
Eg-m97
D
GCACUUGCCUGUCUGA
CGUGAACGGACA
LSU2009 (7)
Eg-m98
D'
AAUGGCAGGGAUAUGA
UUACCGUCCCUG
SSU2123
Eg-m99
CCAUGCACCACUCUGA
GGUACGUGGUGG
SSU1681
Eg-m100
CUUGGUCCGUGUACUGA
GAACCAGGCACA
LSU936 (5)
Eg-m101
CUUGUUCCCUUGGUUGUCUGA
GGGCAAGGGAACCAACA
LSU2853 (8)
Eg-m102
UCUCGUUGAUCCACUGA
AGAGCAACUAGGU
LSU2806 (8)
Eg-m103
UGUCCCAGCCAAUCUGA
GCGGGGUCGGUU
LSU3187 (9)
Eg-m104
AACACUGAGUUUUCUGA
UUGUGACUCAAAA
SSU1037
Eg-m105
AUCAGCACUGCUUCUGA
UGGUCGUGACGAA
LSU1502 (6)
Eg-m106
UUGUGUCCUUCCCCUGA
GACACAGGAAGG
LSU39 (1)
Eg-m107
GCACGUUUCGGCUGA
CGUGCAAAGCC
LSU3315 (9)
Eg-m108
GUUGUUCCAUGCACCUGA
CAGCAAGGUACGU
LSU1039 (5)
Eg-m109
UUUCAUGGCCACCUCUGA
GAAGUACCGGUGG
LSU1608 (6)
Eg-m110
AACCAAGAUCUGUUCUGA
UUGGUUCUAGACGA
LSU1819 (6)
Eg-m111
UGAGGCAUUUGGCCUGA
ACUCCGUAAACCG
LSU2768 (8)
Eg-m112
UUUGCAAGCACUCGCUGA
GAACGUUCGUGAGC
LSU1081 (5)
Eg-m113
AGGUAGUGGACUGA
UCCAUCACC
LSU2993 (9)
Eg-m114
ACUUGCCUGUUUGGCUGA
UGAACGGACAAAC
LSU2005 (7)


Eg-m115
D'
5'
3'
CAGUCAGUAUCUUCUGA
GUCAGUCAUAGG
LSU3694 (10)
Eg-m116
SSU2180
GUUGGAGGACUAC
CAACCUCCUGAUGCUGA
Eg-m117
LSU2776 (8)
UUGGUUGACGAGUCUGA
AACCAACUGCUC
Eg-m118
LSU2909 (8)
UCAAACUAGAGUCUCUGA
AGUUUGAUCUCAG
Eg-m119
LSU3208 (9)
CGUUGUGUUGUAACCUGA
CGAACACAACAUU
Eg-m120
LSU3374 (9)
GACACCUCUGGCUUGCUGA
CUGUGGAGACCGAA
Eg-m122
D
LSU2849 (8)
UGGUGGUUGGGCUGG
ACCAACAACC
Eg-m123
LSU3688 (10)
GUAUCCGUCACUGCUGA
CAUAGGCAGUGA
Eg-m124
LSU41 (1)
UGCUGUGUCCUUUCUGA
ACGACACAGGAAG
Eg-m125
LSU3367 (9)
GGCUUCGAUAUCACUGA
CCGAAGCUAUAG
Eg-m126
LSU1683 (6)
UUGGGCACCAUUCUGA
AACCCGUGGUAG
Eg-m127
LSU1738 (6)
GAUGCUCUGCCACUGA
CUACGAGACGG
Eg-m128
LSU1121 (5)
AGAACGGUAUAUCUGA
UCUUGCCAUAU
Eg-m129
LSU2953 (8)
CACCUCCCACCUUUGA
GUGGAGGGUGGA
Eg-m130
SSU1966
ACACCUUCACUGCCUGA
UGUGGAAGUGAC
Eg-m131
SSU1736
AUGUCUCACUCGUCCUGA
UACAGACUGAGCA
Eg-m132
LSU2685 (8)
AUUUGAGCACUGGGUUCUGA
UAGUCUCGUGACCCG
Eg-m133
SSU110
GGAUGACUGCUUCUGA
UCUACUGACGA
Eg-m134
LSU1189 (5)
CUCAGGCUACACACUGA
GAGUCCGAUGUGU
Eg-m135
LSU1452 (5)
AUGCUCAGAAUGUGA
UACGAGUCUU
Eg-m136
LSU491 (3)
UUCAAACUGCAGUCUCUGA
AAGUUUGACGUCAGG
Eg-m137
LSU2146 (7)
AGCCCGCGUGUCUGA
UCGGGCGCAC
Eg-m138
LSU1888 (6)
UCCAGUCCUCAAACUGA
AGGUCAGGAGUU
Eg-m139
LSU483 (3)
UCCCAAGCCAUCUGA
AGGGUUCGGU

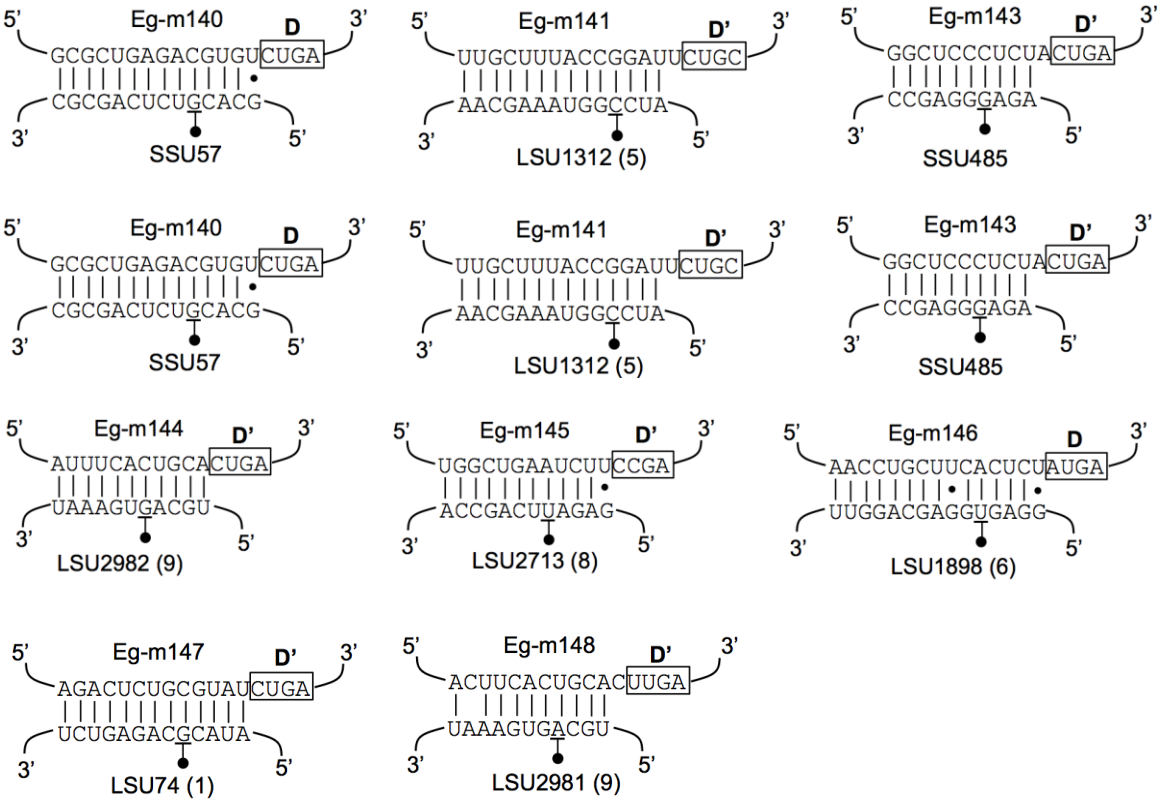
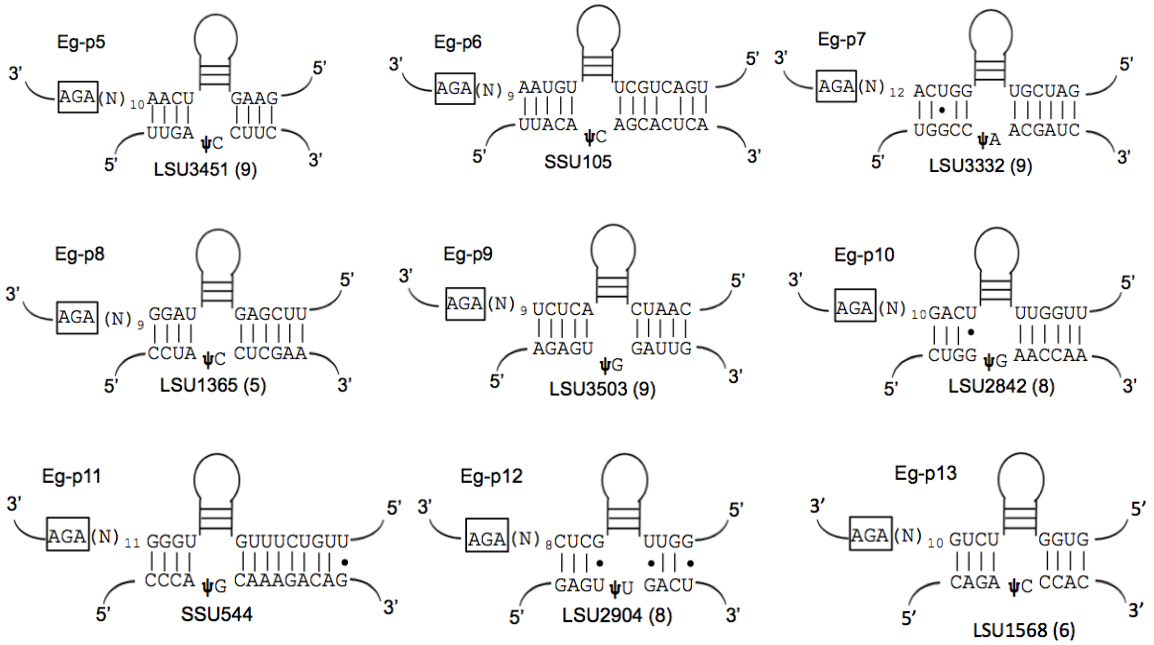


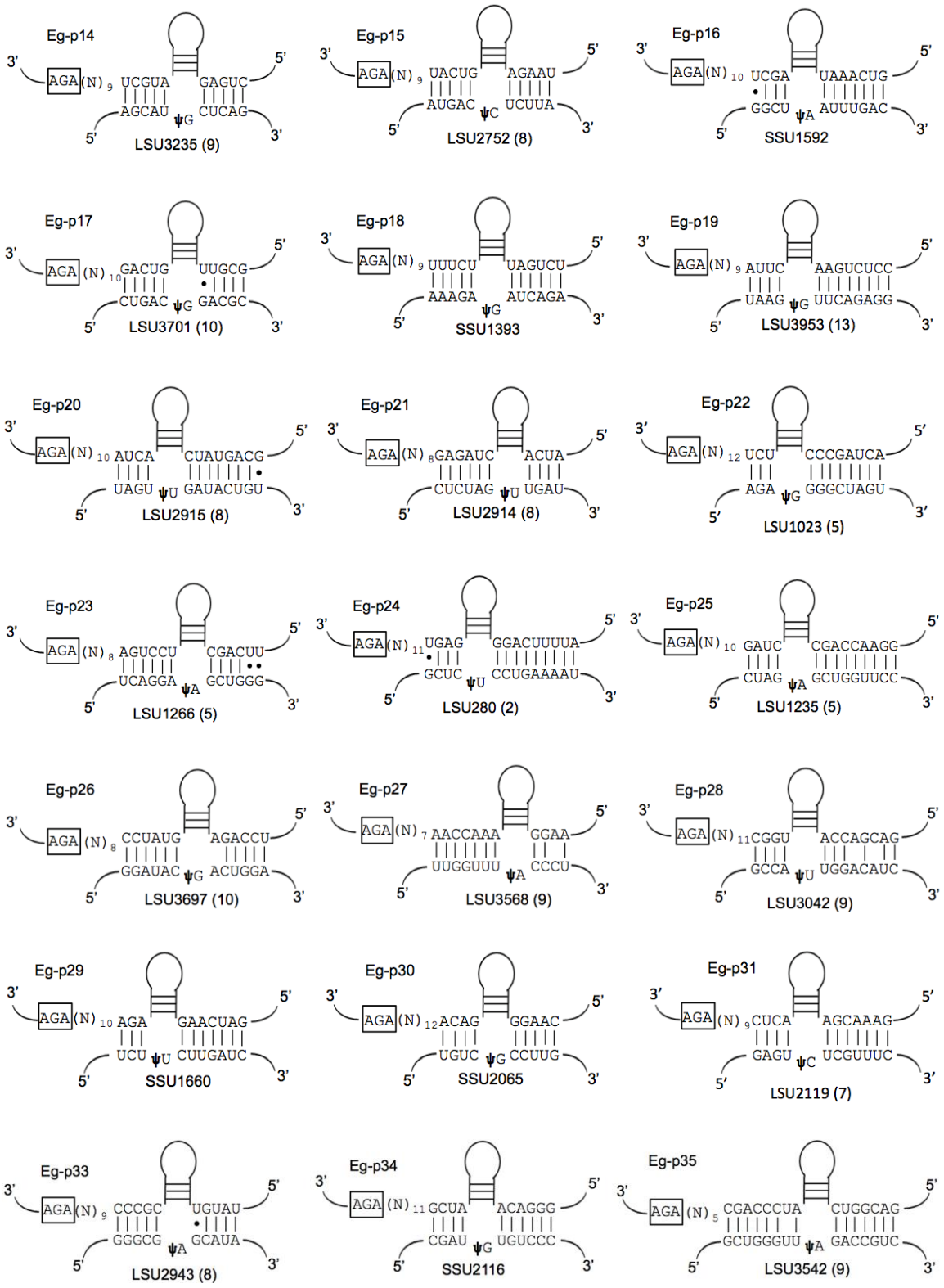
Figure 9. Identified *E. gracilis* box C/D snoRNAs and their predicted target sites in the rRNA for $O^{2'}$-methylation events. The predicted base-pairing interaction between the snoRNA (top strand) and the target region in the rRNA (bottom strand) is depicted. Experimentally confirmed methylation sites (Schnare and Gray, 2011) are underlined and highlighted with a filled circle. The predicted D or D′ box element of each snoRNA is highlighted. 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 Figure S1. Eg-m71 – Eg-m99 were identified through characterization of genomic DNA (adapted from Moore & Russell, 2012) whereas Eg-m100 – Eg-m148 were identified through ncRNA library characterization. Eg-m73 is not shown here as no rRNA target was found. Eg-m121 is shown in Figure 21 and Eg-m142 shown in Figure 22.

The characteristics of the new box C/D snoRNAs identified in this study (Figure 9) are consistent with the main features highlighted previously (Russell et al., 2006) (See 3.7 and Table S12 for details). Isoforms were characterized for nearly half of the newly identified box C/D RNAs, emphasizing the extraordinarily large complement of snoRNAs in this organism. The majority of these snoRNAs target LSU rRNA sites which is consistent with the higher density of modification mapped in the various LSU rRNA

species relative to the single SSU rRNA species (Schnare and Gray, 2011). Our data confirms that *Euglena* contains an exceptionally small number of box C/D snoRNA species capable of acting as double-guide RNAs; that is, targeting two different modification sites in rRNA using two different regions of the same snoRNA. We did not identify any new double-guides in this study (only isoforms of the previously described double-guides Eg-m1 and Eg-m14). For single-guide snoRNAs, the regions not being used for modification targeting do not appear to have alternative rRNA biogenesis roles since they do not exhibit any extensive rRNA base-pairing potential.

All of the pseudouridine-guide RNAs identified here contain a single-hairpin and an "AGA" box motif, structurally similar to those previously identified (Russell et al., 2004) (Figure 10 and Figure S3). The results of 3′ RACE experiments verify that the 3′ ends of the newly identified H/ACA-like snoRNAs are 3 nt downstream of the AGA sequence element and confirm that the predicted snoRNA genes are indeed expressed (Figure 11; see Figure S1 for sequence data).




Eg-p14
AGA (N) 9 UCGUA GAGUC
AGCAU ψG CUCAG
LSU3235 (9)
Eg-p15
AGA (N) 9 UACUG AGAAU
AUGAC ψC UCUUA
LSU2752 (8)
Eg-p16
AGA (N) 10 UCGA UAAACUG
GGCU ψA AUUUGAC
SSU1592
Eg-p17
AGA (N) 10 GACUG UUGCG
CUGAC ψG GACGC
LSU3701 (10)
Eg-p18
AGA (N) 9 UUUCU UAGUCU
AAAGA AUCAGA
ψG
SSU1393
Eg-p19
AGA (N) 9 AUUC AAGUCUCC
UAAG ψG UUCAGAGG
LSU3953 (13)
Eg-p20
AGA (N) 10 AUCA CUAUGACG
UAGU ψU GAUACUGU
LSU2915 (8)
Eg-p21
AGA (N) 8 GAGAUC ACUA
CUCUAG ψU UGAU
LSU2914 (8)
Eg-p22
AGA (N) 12 UCU CCCGAUCA
AGA ψG GGGCUAGU
LSU1023 (5)
Eg-p23
AGA (N) 8 AGUCCU CGACUU
UCAGGA ψA GCUGGG
LSU1266 (5)
Eg-p24
AGA (N) 11 UGAG GGACUUUUA
GCUC ψU CCUGAAAAU
LSU280 (2)
Eg-p25
AGA (N) 10 GAUC CGACCAAGG
CUAG ψA GCUGGUUCC
LSU1235 (5)
Eg-p26
AGA (N) 8 CCUAUG AGACCU
GGAUAC ψG ACUGGA
LSU3697 (10)
Eg-p27
AGA (N) 7 AACCAAA GGAA
UUGGUUU ψA CCCU
LSU3568 (9)
Eg-p28
AGA (N) 11 CGGU ACCAGCAG
GCCA ψU UGGACAUC
LSU3042 (9)
Eg-p29
AGA (N) 10 AGA GAACUAG
UCU ψU CUUGAUC
SSU1660
Eg-p30
AGA (N) 12 ACAG GGAAC
UGUC ψG CCUUG
SSU2065
Eg-p31
AGA (N) 9 CUCA AGCAAAG
GAGU ψC UCGUUUC
LSU2119 (7)
Eg-p33
AGA (N) 9 CCCGC UGUAU
GGGCG ψA GCAUA
LSU2943 (8)
Eg-p34
AGA (N) 11 GCUA ACAGGG
CGAU ψG UGUCCC
SSU2116
Eg-p35
AGA (N) 5 CGACCCUA CUGGCAG
GCUGGGUU ψA GACCGUC
LSU3542 (9)

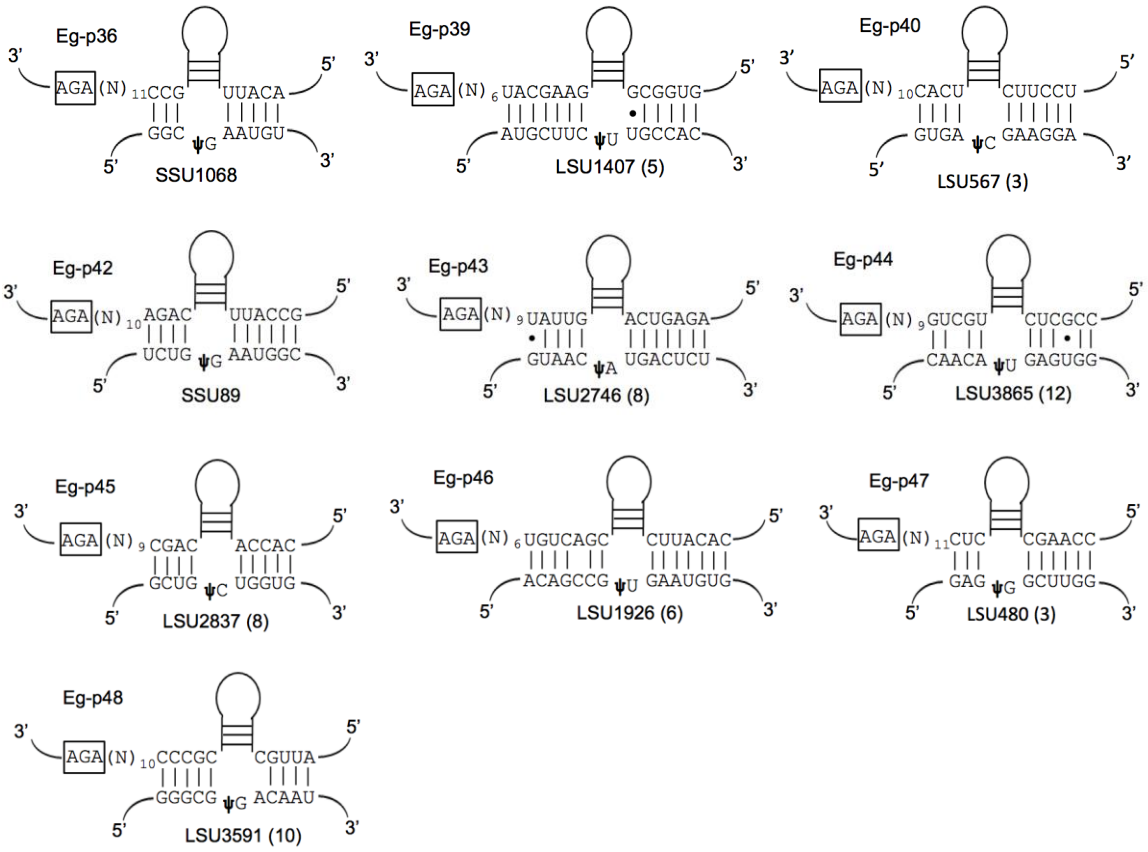


Figure 10. Identified box AGA snoRNAs and their predicted rRNA target sites for pseudouridine (Ψ) formation in *E. gracilis*. Bipartite base-pairing interactions of the snoRNA (top strand) and the region in the rRNA (bottom strand) are shown, with the intervening stem-loop structure within the snoRNA shown schematically. Experimentally confirmed pseudouridine sites are indicated as "Ψ". The AGA box element is highlighted and the number of nucleotides (N) to the base-paired region is indicated. Full-length snoRNA sequences are shown in Figure S1. Eg-p5 – Eg-p13 were identified in genomic DNA (adapted from Moore & Russell, 2012); Eg-p14 – Eg-p48 were identified from the ncRNA library. Eg-p38 shown in Figure 21. Eg-p32, Eg-p37, and Eg-p41 shown in Figure 22.

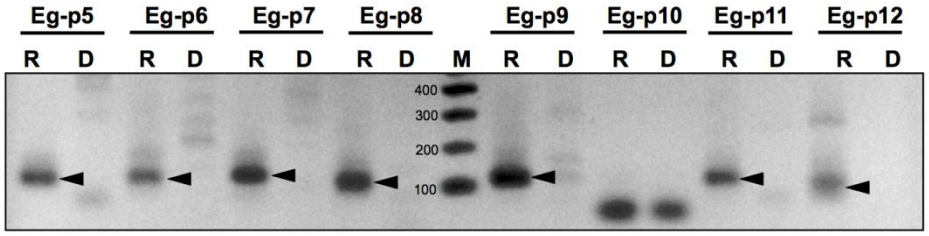


Figure 11. Expression analysis and 3′ end mapping of *Euglena* box AGA snoRNAs. The RT-PCR products have been separated on a 2% agarose gel and visualized by ethidium bromide staining. The lanes labelled "R" contain products obtained when *E. gracilis* total poly-A tailed RNA was used as template during the reverse transcription step of the 3′ RACE procedure whereas "D" lanes include addition of *E. gracilis* genomic DNA instead of RNA during this experimental step. Presence of bands of the expected size (indicated with filled arrowheads) observed only in the "R" lanes verifies that these products were produced from RNA species. M = 2-log DNA ladder (from Moore & Russell, 2012).

The high success rate of our PCR-mediated amplification indicates that snoRNA genes arranged in tandemly repeated clusters is a common genomic organization mode in *E. gracilis*. We characterized 84 unique clusters, each encoding two or more different snoRNAs, as detailed in Figure S2. In many instances, we were able to identify different but sequence-related clusters. Clusters were considered related if they shared at least two of the same snoRNA genes (often present as isoforms), but were more variable in the intergenic regions. The gene order in related clusters is highly conserved. This suggests that entire clusters of snoRNA genes are often repeated in the genome, consistent with previous limited observations that modification-guide snoRNAs are encoded by multi-copy genes in *E. gracilis* (Russell et al., 2004). Because of the repetitive nature of many of our sequenced portions of the *E. gracilis* genome, we were unable to confidently assemble related clusters into genomic contigs. Our data also indicates that box AGA snoRNA genes are found intermingled with box C/D snoRNA genes and this mixed clustered arrangement is most similar to what has been observed in *Trypanosoma brucei* (Liang et al., 2005), a distant but specific relative to *E. gracilis* (Simpson and Roger,

2004; von der Heyden et al., 2004). However; although the number of Ψ guide RNAs is predicted to be less than box C/D RNAs in *Euglena* based on differences in rRNA modification type frequency, numbers of identified box AGA RNA genes in these clusters is significantly lower than the expected difference. This may be the result of amplifying genomic clusters using primers primarily designed on box C/D RNA sequences.

3.2 Large-scale arrangement of snoRNA gene clusters in the *E. gracilis* genome

The PCR-based approach used to identify snoRNA genes in *E. gracilis* yielded a large collection of unique gene clusters. However, this approach is limited, as PCR preferentially amplifies the smallest gene repeat, making it difficult to determine the expansiveness of the tandem repeats (tandem repeats of snoRNA gene clusters will be referred to as 'snoRNA gene arrays'). To determine how large snoRNA gene arrays can be in *Euglena*, a bacterial artificial chromosome (BAC) library containing *E. gracilis* genomic DNA fragments (~15-100 kb) was constructed. PCR was initially used to screen the library for *E. gracilis* snoRNA-coding regions of interest. Of the 57 unique regions that gave positive results in PCR amplifications of *Euglena* total genomic DNA (previous results, 3.1), 22 were confirmed to be present in the BAC library, as evidenced by the successful amplification of a product of the predicted size (Table S13). *Euglena gracilis* nuclear DNA contains an unusual base modification, β-D-glucopyranosyloxymethyluracil (or Base J), that replaces ~1% of thymidine residues (Dooijes et al., 2000; Borst and Sabatini, 2008). Previous attempts to construct a cosmid *E. gracilis* genomic DNA library were unsuccessful and it is predicted that Base J was likely inhibiting the packaging of

genomic DNA into phage particles. Base J may also affect BAC library representation if it has an inhibitory effect during library synthesis. Based on these preliminary results, the library was then screened through hybridization with radioactive probes designed to anneal to snoRNA-coding regions confirmed to be represented in the library by PCR. Four BACs, each containing unique genomic regions, were isolated and the snoRNA gene arrays were characterized.

The first genomic region characterized from the BAC library encodes snoRNA gene Cluster 23 (termed BAC 23) (Figure 13a), which was first characterized using the PCR approach. The *Euglena* genomic DNA fragment isolated from the BAC is approximately 17 kb in size (Figure 12). Through sequencing by primer walking and PCR amplification, the 5′ and 3′ extremities of the DNA insert were characterized. The first snoRNA gene (encoding Eg-m56) is located 2,369 nt downstream from the BAC 5′ cloning site and another copy of Eg-m56 is present 2,273 nt upstream from the 3′ cloning site (Figure 13b). It appears we have mapped the entirety of this snoRNA gene array, as the genomic sequences upstream and downstream of the flanking snoRNAs are unrelated to the cluster and are also unrelated to each other. Very little is known about how gene expression is regulated in *Euglena*, making it difficult to conclusively identify transcriptional promoter and terminator elements in these regions.

To determine if Cluster 23 is repeated in tandem, forming a large snoRNA gene array across the ~11.7 kb of uncharacterized sequence between the flanking snoRNA genes, restriction digests were performed with HindIII enzyme that cleaves once between the Eg-p7 and Eg-m56 genes (Figure 13a). Complete enzyme digestion of this BAC with HindIII produced six distinct products (Figure 13c, lane 1) and the identity of these

fragments (based on expected size) was determined by examining the HindIII recognition nucleotide sequences in the vector and the sequenced 5′ and 3′ extremities of the genomic insert. The prominent ~600 nt band corresponds to the size of the Cluster 23 repeat between HindIII cut sites, and the stronger intensity of this band relative to the larger products generated from the vector is consistent with multiple copies of Cluster 23 being present in the genomic insert. Based on the size of the uncharacterized sequence between the flanking Eg-p7 snoRNA genes (~11.7 kb), it is predicted that up to 20 tandem copies of this gene cluster repeat could be present.

To further characterize this large, repetitive snoRNA gene array, the BAC DNA was partially digested with HindIII, which resulted in a ladder of bands (14 of which could be clearly resolved by gel-electrophoresis) where each product differed in size by 600 nt (the size of the Cluster 23 HindIII fragment) (Figure 13c, lanes 2, 3, 7 & 8). To determine the extent of sequence variability within each "repeat" unit and the number of unique repeat variants, the 600 nt HindIII digest product was cloned and 69 of those clones were sequenced. Fourteen unique sequence variants of this repeat (with >96% sequence identity) were observed, containing 4 isoforms of Eg-m56 (Eg-m56, Eg-m56.2-56.4) and 6 isoforms of Eg-p7 (Eg-p7, Eg-p7.2-7.6). The sequence variation between snoRNA isoforms does not occur in the guide regions and mutations to the canonical box sequences only occur in 2 of the isoforms (Eg-p7.3 and Eg-p7.6) (Figures S1 & S4). Hence, snoRNA gene Cluster 23 is repeated in tandem at least 14 times in this array; however, this is likely an underestimation. Some of the tandem clusters may be 100% identical in sequence and therefore not distinguishable by the shotgun cloning and sequencing of the 600 nt digest product. The other possibility is that some clusters have a sequence change in the HindIII site and would not be present in the 600 nt digest product

pool. Consistent with this, even under complete digestion conditions, a band at ~1200 nt (lane 1) represents a few clusters that are unable to be cut at an altered internal Hind III site (as confirmed through sequencing of this product). As expected, this band increases in intensity in partial digest conditions (cf. lanes 1 and 2). Given that snoRNA genes Eg-m56 and Eg-p7 are still present together at both ends of the array, in combination with the restriction enzyme digest patterns, it is probable that there are in fact 20 copies of the repetitive snoRNA cluster unit in this array (to make up the distance between Eg-p7 genes in Figure 13b).

The region encoding snoRNA gene Cluster 2 (previously characterized using the PCR approach, Figure 13d) was also characterized in the BAC DNA library (BAC 2). The cloned genomic fragment is ~40 kb (Figure 12), with the first snoRNA gene (Eg-m2) found 402 nt downstream of the 5′ cloning site and the last snoRNA gene (Eg-m1) 1,588 nt upstream of the 3′ cloning site (Figure 13e). Again, the outward sequences flanking the first characterized snoRNA gene and the last snoRNA gene are unrelated, indicating this BAC insert likely contains a complete and very large snoRNA gene array. Using the same partial restriction enzyme digest strategy described previously, the genomic fragment was further characterized using SphI, which cleaves once in the intergenic sequence between Eg-m2 and Eg-p2, Figure 13d). Complete digestion yielded a range of products, including a prominent ~550 nt band, which corresponds to the size between SphI cut sites in the snoRNA gene cluster (Figure 13f, lane 1). It is likely this band is a collection of snoRNA gene repeats that constitute a significant fraction of a large array that spans the 37.5 kb of uncharacterized genomic sequence (it is predicted up to 50 copies of this gene repeat may be present).

Partial restriction digests produced a ladder of bands, each product differing in size by ~550 nt (Figure 13f, lanes 2 & 8), indicating the presence of multiple tandem copies of this snoRNA gene cluster. The abundant 550 nt product was gel-extracted, cloned and 45 clones were sequenced. Thirty-three variants of gene Cluster 2 were characterized, which included 6 new isoforms of Eg-p2 (Eg-p2.1-2.6), 6 new isoforms of Eg-m1 (Eg-m1.3-1.8) and 6 new isoforms of Eg-m2 (Eg-m2.3-2.8) (Figures S1 & S4). The guide region is conserved in the isoforms (with the exception of Eg-p2.2 and Eg-m2.8 where 1 nt mutations are present) and the canonical box elements are also highly conserved. Consistent with the previous example, the level of sequence identity between the gene cluster variants is very high (>93%).

A third snoRNA gene cluster (Cluster 25, Figure 13g) was characterized from the BAC library (BAC 25). The procedure as previously described for the other BACs was followed. The 3′ end of the ~12 kb insert (Figure 12) was sequenced through primer walking and two Eg-m18 gene isoforms were identified (Figure 13h); however, the sequencing quality from the 5′ end was poor and not interpretable. In this case, the snoRNA gene array appears to be incomplete – genomic fragmentation during BAC library construction has occurred at an internal position within this snoRNA gene array for this particular genomic insert into the vector (Figure 13h). Based on the partial restriction digest pattern, only 4 repeats (~550 nt for each repeat) of this cluster are present in this cloned genomic fragment (Figure 13i, lane 2). This was further supported by sequencing the abundant ~550 nt product; 29 clones were sequenced but only 4 unique variants of this repeat were observed, including 2 new isoforms of Eg-m18.

The genomic region encoding the snoRNA Eg-h1 was also characterized from the BAC library. Eg-h1 is an unusual triple-stem box H/ACA snoRNA that does not appear to guide any known rRNA pseudouridine modifications (Russell et al., 2004; Schnare and Gray, 2011). It more closely resembles box H/ACA RNA structure characterized in yeast, vertebrates and some archaea. The function of Eg-h1 is still unknown, but based on its ability to base pair to regions of rRNA (non-modified sites), it is predicted to instead play a role in *Euglena*'s unique rRNA processing (cleavage) pathway. Prior to this study, the genomic organization of Eg-h1 was unknown. The genomic organization of the encoding region for Eg-h1 was of particular interest because the genomic context of regions encoding the *Euglena* U3 snoRNA, a known rRNA processing snoRNA, have been previously characterized (Charette and Gray, 2009) and the genomic organization is strikingly different to the genes encoding modification-guide snoRNAs in *Euglena*. The BAC clone containing a genomic fragment encoding Eg-h1 was isolated and found to be approximately 18 kb in size (Figure 12); the first Eg-h1 gene copy is 713 nt downstream of the 5′ cloning site and another gene copy is located 1,402 nt upstream of the 3′ cloning site (Figure 13k). Again, it appears that both ends of a large snoRNA gene array encoding multiple copies of Eg-h1 have been mapped.

In this case, the restriction enzyme NcoI that cleaves in the intergenic regions directly upstream of Eg-h1 genes (Figure 13j), was used to estimate the number of tandem repeats. Complete digestion of this BAC with NcoI yields 5 distinct bands (Figure 13l, lane 1), which can all be mapped to the vector or the 5′ and 3′ extremities of the genomic insert that were characterized by primer walking. However, it is clear that the additive size of these bands does not equate to 18 kb (the size of the genomic insert). The

prominent ~480 nt band is the size of the Eg-h1 gene repeat produced after complete digestion with NcoI (Figure 13l) and likely represents multiple copies of the repetitive unit. Maximally, the uncharacterized region (~15 kb) between the flanking copies of Eg-h1 could encode up to 30 tandem copies of this gene repeat. Partially digesting this BAC with NcoI resulted in a series of bands each differing by ~480 nt (Figure 13l, lanes 2, 3,7 & 8), again indicative of a snoRNA gene array.

The prominent 480 nt band was gel-extracted, cloned and 82 individual clones were sequenced. Twenty-two variants of this gene repeat showing >90% sequence identity, were identified and 6 new isoforms of Eg-h1 were characterized (Figure S4). From this, we can conclude that at least 22 repeats of the Eg-h1 gene cluster are present, forming a very large snoRNA gene array. Again, it is likely this array is even larger since Eg-h1 genes are present at both ends (up to 30 copies based on the size of the insert), some repeats may be identical in sequence and/or 82 sequenced clones is not enough excess sequencing coverage to ensure the detection of every repeat variant.

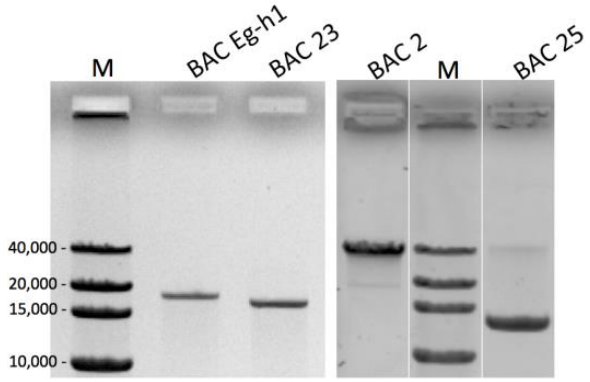
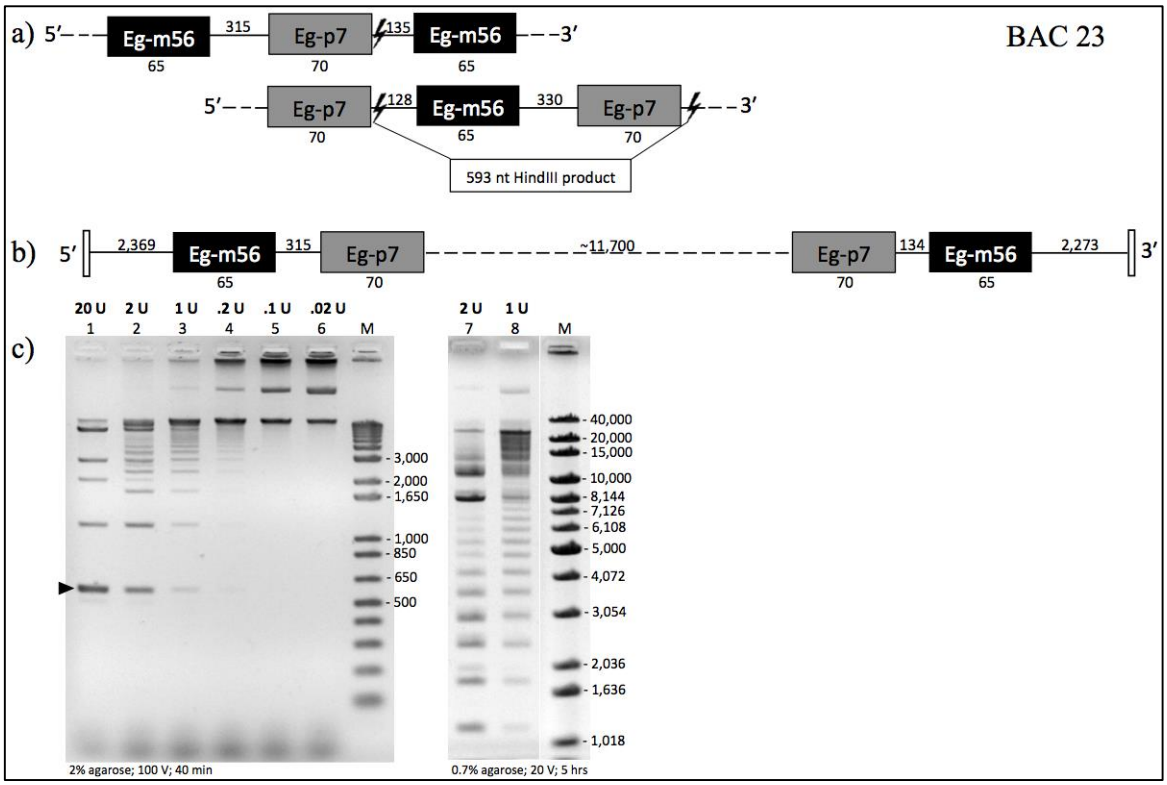
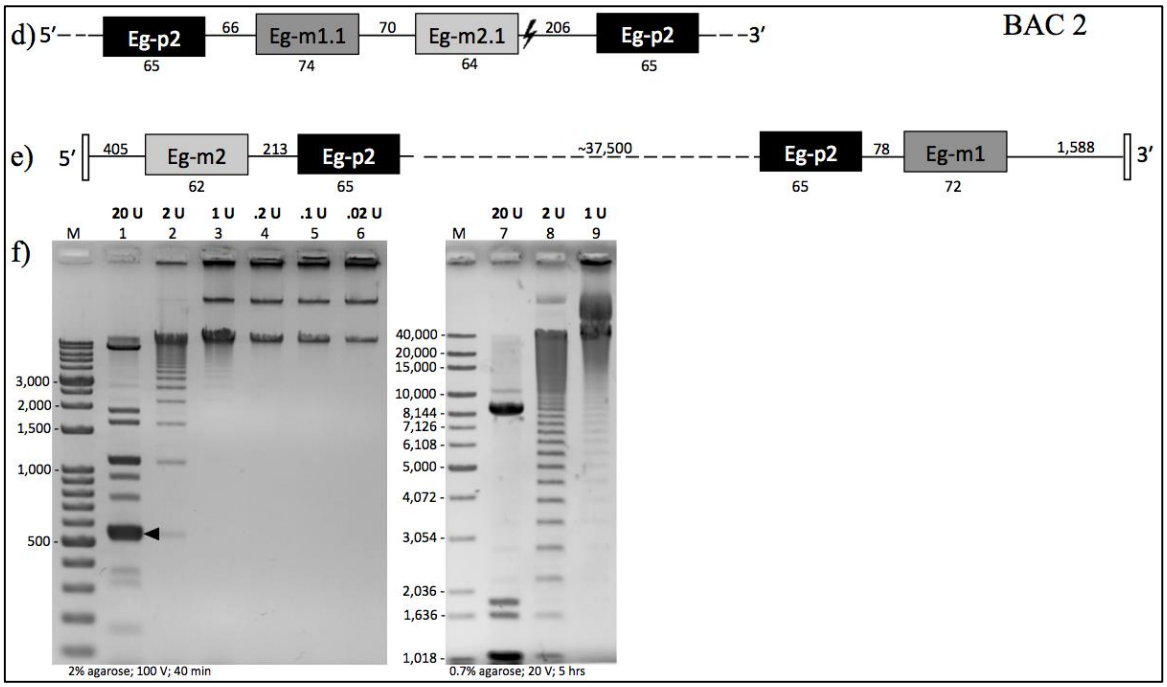


Figure 12. Determining the size of cloned *E. gracilis* snoRNA-encoding genomic DNA fragments. Restriction enzyme digestion products have been resolved on a 0.7% agarose gel and visualized by ethidium bromide staining. BACs containing snoRNA coding regions of interest were isolated and the DNA was digested with the enzyme NotI, which cleaves upstream and downstream of the cloning site. The bands indicate the size of the DNA insert (+ 623 nt of vector). M = 1 kb DNA extension ladder.


a) 5′
Eg-m56
65
315
Eg-p7
70
135
Eg-m56
65
3′
BAC 23
5′
Eg-p7
70
128
Eg-m56
65
330
Eg-p7
70
3′
593 nt HindIII product
b) 5′
2,369
Eg-m56
65
315
Eg-p7
70
~11,700
Eg-p7
70
134
Eg-m56
65
2,273
3′
c)
20 U 2 U 1 U .2 U .1 U .02 U
1 2 3 4 5 6 M
3,000
2,000
1,650
1,000
850
650
500
2% agarose; 100 V; 40 min
2 U 1 U
7 8 M
40,000
20,000
15,000
10,000
8,144
7,126
6,108
5,000
4,072
3,054
2,036
1,636
1,018
0.7% agarose; 20 V; 5 hrs


d) 5′
Eg-p2
65
66
Eg-m1.1
74
70
Eg-m2.1
64
206
Eg-p2
65
3′
BAC 2
e) 5′
405
Eg-m2
62
213
Eg-p2
65
~37,500
Eg-p2
65
78
Eg-m1
72
1,588
3′
f)
M 20 U 2 U 1 U .2 U .1 U .02 U
1 2 3 4 5 6
3,000
2,000
1,500
1,000
500
2% agarose; 100 V; 40 min
M 20 U 2 U 1 U
7 8 9
40,000
20,000
15,000
10,000
8,144
7,126
6,108
5,000
4,072
3,054
2,036
1,636
1,018
0.7% agarose; 20 V; 5 hrs

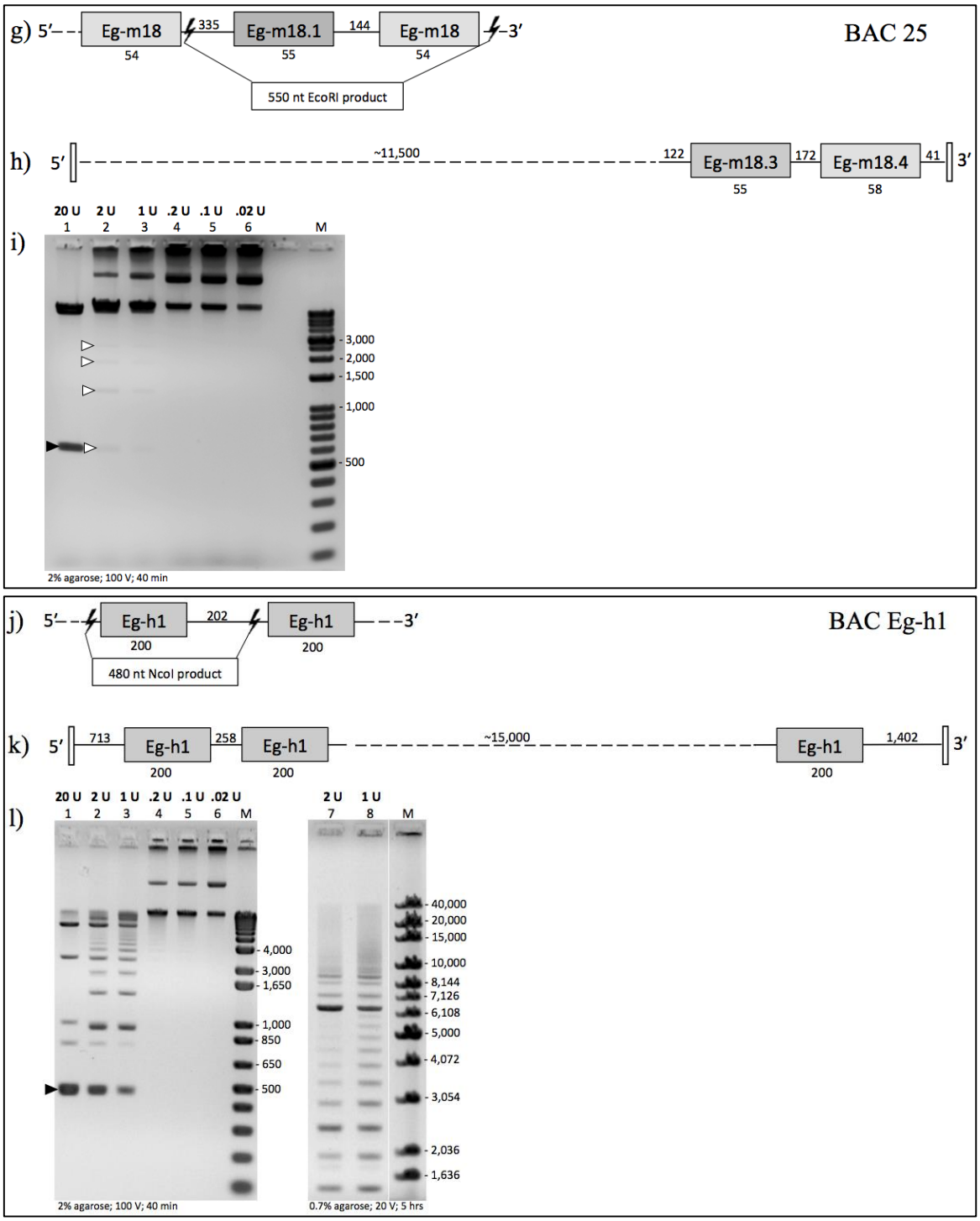


Figure 13. Analysis of the large-scale arrangement of snoRNA gene clusters in *E. gracilis*. Gene clusters were characterized through sequencing and restriction digests of isolated BAC DNA. Grey boxes represent snoRNA coding regions, white boxes represent vector sequence, solid lines are known intergenic sequence, dashed lines indicate unknown sequence and black bolts represent restriction enzyme cleavage sites. Numbers above the lines are the size of intergenic regions (in bp) and numbers below the boxes are the size of coding regions (in bp). **(a, d, g & j)** Gene clusters previously characterized

from PCR-amplified genomic DNA. The restriction enzyme cleavage sites and predicted size of the digest products (if known) are indicated. **(b, e, h, & k)** Characterized regions of the cloned genomic fragments. The 5′ and 3′ extremities were sequenced by primer-walking and/or characterized by PCR amplification, followed by cloning and sequencing. **(c, f, i & l)** Restriction digest products resolved by agarose gel-electrophoresis (conditions are indicated). The amount of enzyme used (in units, U) is indicated. Products indicated with black arrow heads are predicted abundant gene repeats (these bands were extracted, cloned and sequenced, see main text). Bands indicated with white arrow heads highlight the faint partial digest products of BAC 25.

Collectively, the BAC genomic library results show that snoRNA gene clusters can exist as very large, dense 'arrays' in *Euglena*. It appears that some of these arrays may even exceed 40 kb in size. The sequence conservation between gene copies within these arrays is high, with most of the variation occurring in intergenic regions or in gene positions not predicted to significantly affect the function of the expressed RNA. Assuming that Eg-h1 is a processing snoRNA, these results show that genes encoding both types of snoRNAs (processing snoRNAs and modification-guides) are highly repetitive and are found in large arrays. While it is known that snoRNA genes in some eukaryotic species are clustered together (Dieci et al., 2009), the expansiveness and highly repetitive nature of snoRNA gene arrays in *E. gracilis* is very unusual.

3.3 Identification of a U14 snoRNA homolog in *E. gracilis* (based on Moore & Russell, 2012)

Through analysis of the intervening genomic sequence between tandem snoRNA gene copies, we have now identified a U14-like snoRNA in *E. gracilis*. Within one previously known snoRNA gene cluster (Russell et al., 2004), an imperfect direct repeat was identified. Although the repeated sequence contained predicted C and D box elements, it was significantly longer than the typical biochemically-isolated *Euglena* modification-guide box C/D snoRNAs, and it did not have significant base-pairing

potential to any mapped modified regions of the rRNA. It was found that a region in this RNA has the ability to extensively base-pair to the same conserved region of the SSU 18S rRNA known to interact with the processing domain of the U14 snoRNAs identified in *H. sapiens, S. cerevisiae, O. sativa* and *D. melanogaster*.

U14 belongs to the box C/D snoRNA subclass involved in pre-rRNA cleavage events, and is unique in that it is known to be a dual-function snoRNA in yeast, vertebrates and plants (Li et al., 1990; Kenmochi et al., 1996; Jiang et al., 2002). It acts as both a pre-rRNA "chaperone" which is important for rRNA processing, as well as a methylation guide RNA. These events are mediated using two different snoRNA anti-sense elements to the 18S rRNA (Liang et al., 1997). Domain A, a ~13 nt antisense element to the conserved region in the 18S rRNA, guides RNA cleavage events of the precursor transcript. This element is present in the *E. gracilis* U14-like snoRNA (Figure 14). However, the *E. gracilis* U14 appears to lack Domain B functionality, which is to guide the methylation of a conserved cytosine residue (human Cm462) in the 18S rRNA (Figure 14). In yeast, it has been shown that Domain A, but not Domain B, is required for cell viability (Jarmolowski et al., 1990).

The *Euglena* case is similar to what is observed in *D. melanogaster* (Yuan et al., 2003) and *A. gambiae*, whose U14 also lacks Domain B functionality. However, in both Diptera species, a different box C/D snoRNA species has been identified which can guide methylation at this site (Yuan et al., 2003). In *E. gracilis*, this 18S rRNA residue does not appear to be methylated (Schnare and Gray, 2011). While the dual processing and guide functions of the U14 in humans, plants and yeast apparently have been functionally split into two distinct snoRNAs in Diptera, the methylation function seems to be lost altogether

in *Euglena*, suggesting that modification at this site may not be essential, as seems to be the case in yeast (Jarmolowski et al., 1990).

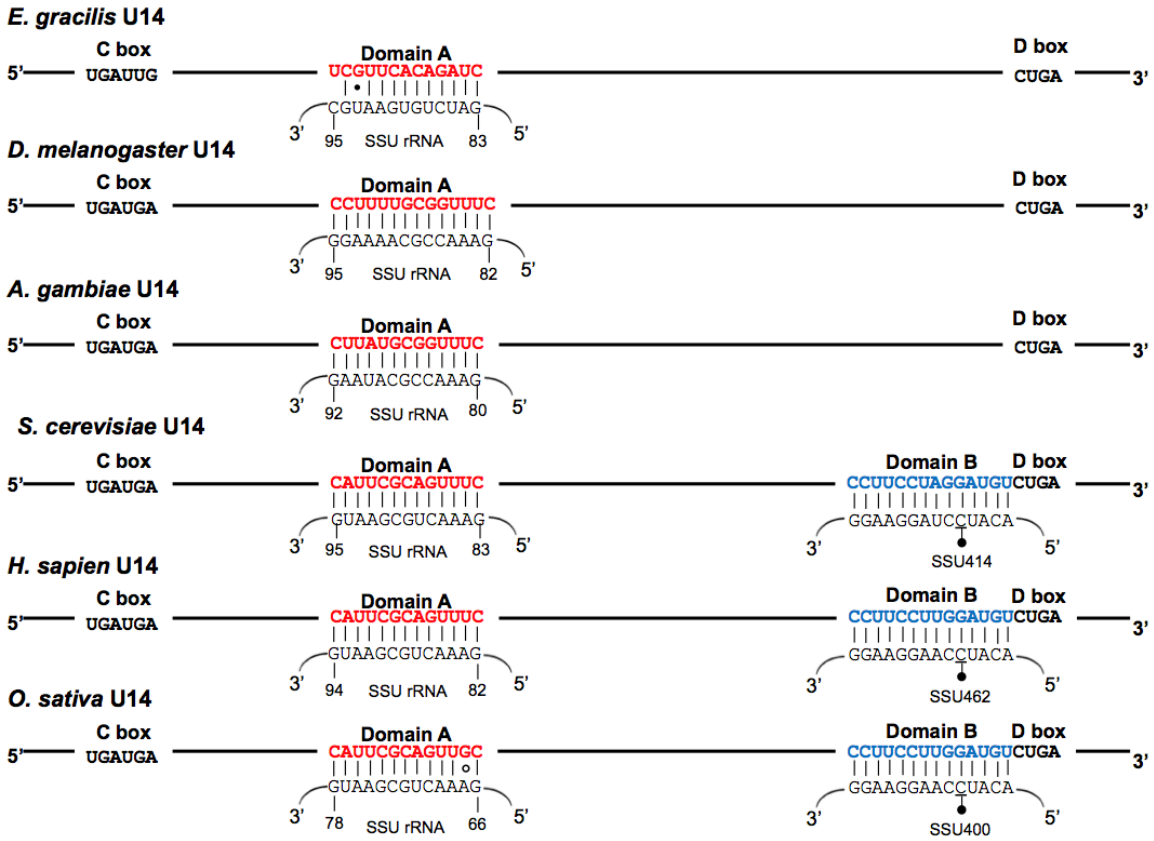


Figure 14. Identification of a U14 snoRNA homolog in *E. gracilis*. Predicted base-pairing between the *E. gracilis* U14 snoRNA (top strand) and the conserved target region(s) in the 18S SSU rRNA (bottom strand) is shown. Also shown are previously determined base-pairing interactions that occur between *H. sapiens* (Genbank Accession No. D88010), *S. cerevisiae* (X96815), *O. sativa* (AF332622), and *D. melanogaster* (NR_001639) U14 snoRNAs and their respective SSU rRNAs. The *A. gambiae* U14 homolog was identified in a previously annotated cDNA sequence (BX030536.1). For those organisms containing a second region of U14 snoRNA complementarity (Domain B) to the SSU rRNA, the target rRNA nucleotide for $O^{2'}$-methylation is underlined and highlighted with a filled circle (from Moore & Russell, 2012).

As U14 was initially characterized in a snoRNA gene repeat, it would suggest the genomic organization of U14 in *Euglena* is tandemly repeated clusters. Three sequence isoforms of U14 were initially identified in the gene cluster; additional sequencing and characterization of PCR and RT-PCR products revealed an additional 12 isoforms of U14 (Figure S1). While the large-scale genomic organization of U14 has not been

characterized, it is evident that the gene encoding U14 is multi-copy and at least some isoforms are found together with modification-guide snoRNAs in tandemly repeated clusters. In other eukaryotes, multiple isoforms of U14 have been identified, but not to the same extent as observed in *Euglena* (for example, four in maize and two in *Drosophila*). The organization of U14 and modification-guide snoRNA genes together has also been observed in plants (Leader et al., 1997). Interestingly, the genes encoding other processing snoRNAs in *Euglena*, such as U3 (Charette and Gray, 2009) and Eg-h1 (3.2), are multi-copy but are not found with genes encoding modification-guide snoRNAs.

3.4 Evolution of *E. gracilis* box C/D snoRNA genes and modification sites (based on Moore & Russell, 2012)

Mapping studies in *E. gracilis* have shown that this organism has an unusually large number of modified rRNA sites, most of which are clustered to distinct regions with some such regions containing the highest density of modification of any organism examined. Therefore, we wondered whether the pattern of snoRNA gene organization, and more specifically our finding of a preponderance of snoRNA isoforms, can explain *Euglena*'s rRNA modification profile. snoRNA isoforms could arise through evolution of gene copies generated through duplications of small genomic segments (containing a single snoRNA gene) or through duplications of a snoRNA gene cluster. Isoforms targeting the same rRNA modification site would show sequence changes in regions not affecting rRNA base-pairing interactions and we see numerous such cases in *Euglena* (refer to Figure S1). However, if sequence evolution occurred in the region that is complementary to the rRNA target, a snoRNA isoform may now target a new site in the rRNA due to altered base-pairing potential (neofunctionalization).

We have now characterized seven clear examples of box C/D snoRNA gene isoform evolution showing sequence divergence in the region that base pairs to the target rRNA (Figure 15 and Figure S5). In the first example (Figure 15a), two adjacent snoRNA genes encoding Eg-m80 and Eg-m79 target adjacent rRNA sites Gm628 (LSU 3) and Gm631 (LSU 3), respectively. This is the result of a 3 nucleotide insertion/deletion between the two snoRNAs in the guide region directly upstream of the D′ box, which alters target site specificity based on the conserved n+5 rule of methylation target site selection. Interestingly, Eg-m79 also shows further sequence divergence in the 5′ region of its guide element predicted to allow an rRNA base-pairing interaction of similar length to Eg-m80 (also refer to Figure 15a). We hypothesize this evolutionary scenario arose through a small scale genomic duplication of the region between Eg-m16 and Eg-m81.

Similar to the previous example, the box C/D RNAs Eg-m94 and Eg-m95 in Cluster 28 (Figure 15b) target adjacent sites Cm1626 (LSU 6) and Am1624 (LSU 6), again displaying altered target site specificity primarily through shifts in relative position to the D′ box guide element. Due to their relatively close genomic spacing (660 bp), we once again speculate this represents a tandem gene duplication event. However, in the region spanning Eg-m70 gene copies (Figure 15b), we also discovered the pseudouridine-guide RNA Eg-p11. The 2′-O-methylation site guided by Eg-m94 is *Euglena* specific, while the site guided by Eg-m95 is present in at least one other eukaryote (Schnare and Gray, 2011), therefore it is hypothesized that Eg-m95 is the ancestral snoRNA.

The third example (Figure 15c) appears to have arisen through a snoRNA gene cluster duplication spanning three different snoRNA species; closely-related isoforms can be found for all genes when comparing Cluster 12.5 and 12.4. Based on the much higher

degree of sequence identity between the snoRNA gene isoforms in this example, it appears to have been a much more recent duplication event than would be the case for Clusters 27 and 28. This is readily apparent by comparing Eg-m38 and Eg-m90, which show little sequence variation other than the insertion/deletion that has altered target site specificity.

The fourth illustrated example (Figure 15d) is similar to the first two, in that it appears to have arisen through a tandem gene duplication event. Within cluster 10, three isoforms of the double-guide snoRNA Eg-m14 (note: this modification guide snoRNA was named prior to the discovery of the *Euglena* U14 homolog) are present in a ~950 bp region (see Figure S1 for sequences of all identified Eg-m14 isoforms). Eg-m14.2 and Eg-14 target the same sites in the rRNA for methylation (Cm1204 and Gm1217, LSU 5). Due to a single nucleotide indel in the guide region upstream of the D box, Eg-m14.3 guides the methylation of an adjacent nucleotide (Cm1218, LSU 5), while maintaining the ability to target Cm1204 (LSU 5). In all identified cases (see Figure S5 for additional examples), the most conserved regions between isoforms (which we have arbitrarily designated a new snoRNA species if targeting a different rRNA nucleotide) are the guide regions specifying modification and the box elements required for snoRNP protein interaction.

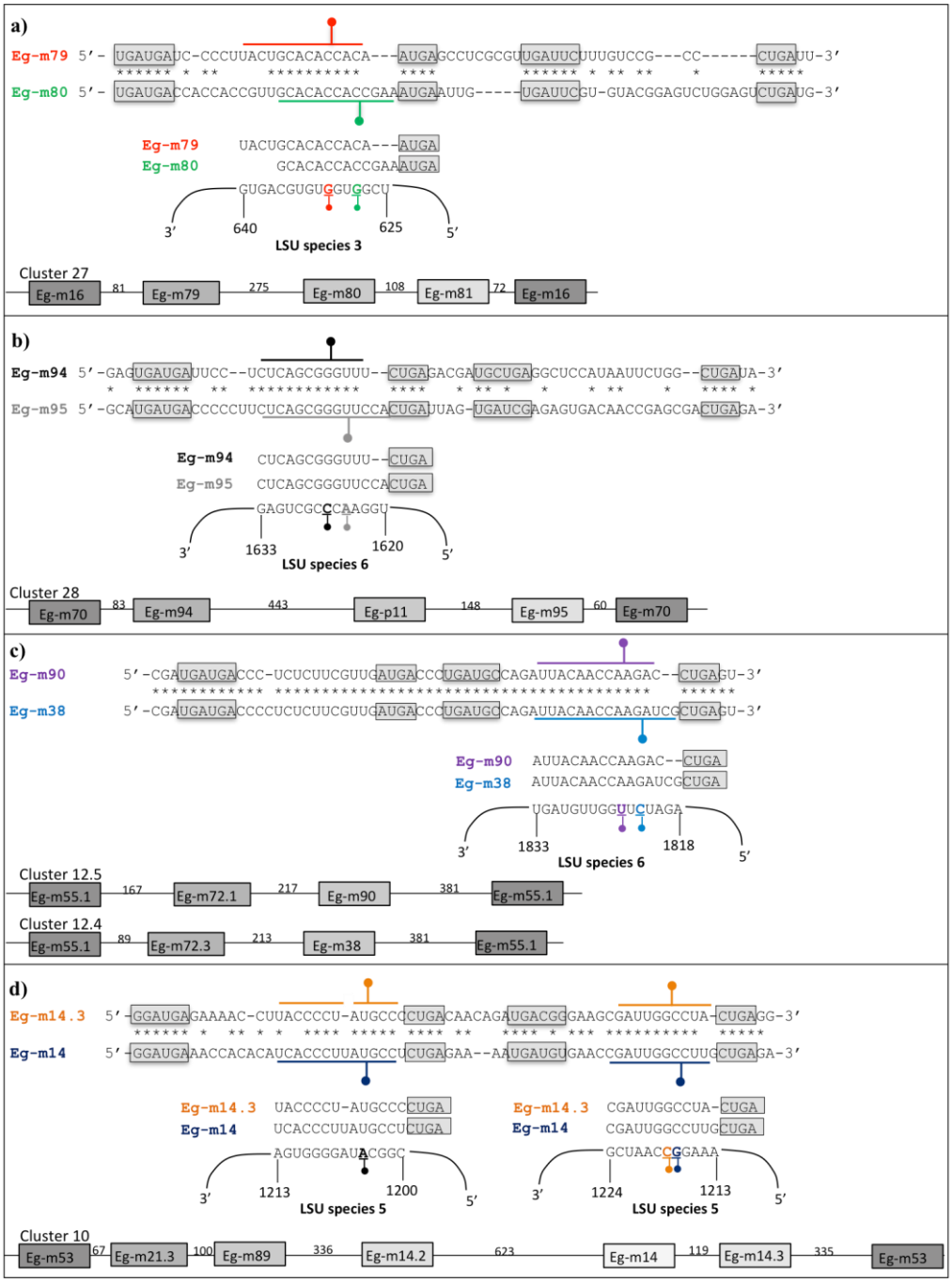


Figure 15. Gene duplication as a mechanism for box C/D snoRNA evolution and methylation clustering in *E. gracilis*. snoRNA species related through predicted sub-genomic duplication events are aligned with predicted C, D′, C′ and D box elements highlighted. Asterisks indicate positions of nucleotide identity between snoRNA pairs and dashes, positions of indels. The region of each snoRNA that base-pairs to the rRNA to target a site for $O^{2'}$-methylation (circle) is indicated with a line. The genomic clusters that contain the related snoRNA sequences are also illustrated schematically. The specific base-pairing interactions predicted to occur to target the methylation events, are shown in more detail below the snoRNA sequence alignments to illustrate how indels have altered rRNA target site specificity via the snoRNA "n+5 rule" (from Moore & Russell, 2012).

3.5 *E. gracilis* snoRNA genes are polycistronically transcribed (based on Moore & Russell, 2012)

Using an RT-PCR based approach, we investigated whether the snoRNA gene clusters are transcribed as polycistronic RNAs. Based on the close proximity of multiple adjacent snoRNA genes and their being encoded on the same DNA strand, we predicted that polycistronic transcription followed by individual snoRNA cistron removal may be part of the snoRNP biogenesis pathway in *Euglena*. This would be similar to the means by which many *Trypanosome* snoRNAs appear to be processed. Clustered snoRNAs have also been characterized in plants (Brown et al., 2003) and yeast (Qu et al., 1999).

Using appropriate oligonucleotides targeting 17 different snoRNA clusters, we could detect RNA precursor transcripts containing more than one snoRNA sequence (Figure 16 shows polycistronic precursor transcripts for eight clusters; for specific oligonucleotide information and binding position see Figure S2. For additional gel-images see Figure S6). For example, polycistronic transcripts corresponding to the genomic region Cluster 7.2 from Eg-m37 to Eg-m88 (lane 1, Figure 16a) and from Eg-m6 to Eg-m11 (lane 3) were detected. In all cases, product bands of the expected size were visualized and then individually cloned and sequenced to verify their identities. Importantly, amplification was only observed upon addition of reverse transcriptase (RT+) indicating these products were obtained from RNA templates and not from contaminating genomic DNA in the RNA sample.

Some of the polycistronic transcripts appear to be quite large (750 nt) such as Cluster 27.3 (lane 11, Figure 16a) and in many instances we were able to detect transcripts containing a minimum of three different snoRNA sequences. We currently cannot accurately estimate how large these precursor transcripts may actually be, as we found

that the efficiency of amplification decreased quite dramatically for products larger than 400 nt in length (Figure 16a cf. lane 1 to lanes 7, 9, 11).

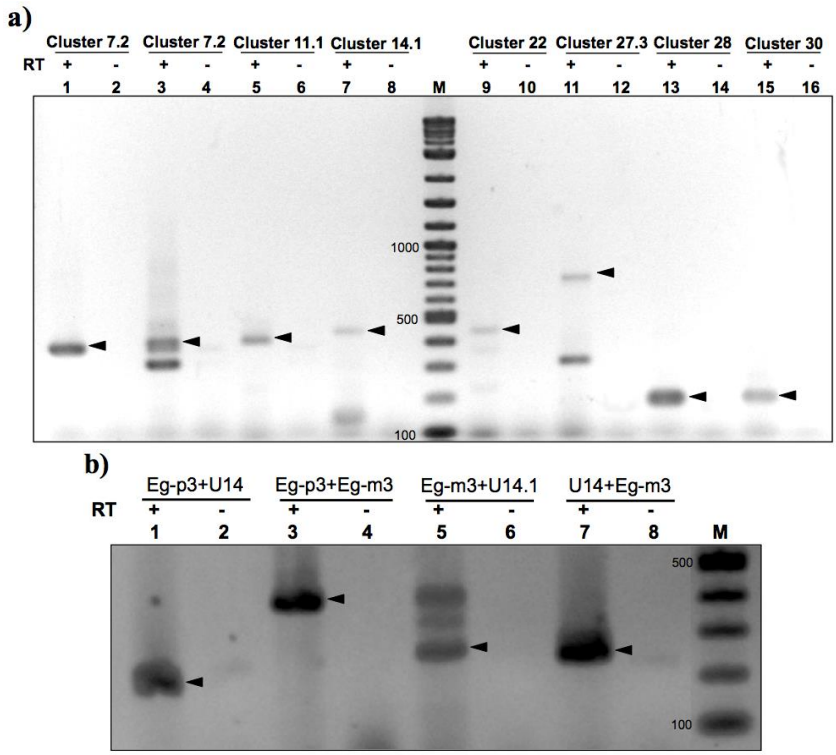


Figure 16. Polycistronic expression of snoRNA genes in *E. gracilis*. RT-PCR products produced from precursor polycistronic snoRNA transcripts have been separated on a 2% agarose gel. Reactions were performed either including reverse transcriptase enzyme (RT+) or without (RT-) to confirm absence of possible DNA contamination in the RNA sample that would result in unwanted amplification from genomic DNA during the PCR step. Table S5 lists detailed oligonucleotide primer pair sequences and Figure S2 describes genomic snoRNA cluster information and oligonucleotide primer positions. Filled arrowheads indicate products of predicted size for amplification between different snoRNA species located on the same RNA transcript. **a)** Lanes 1 and 3 are different oligonucleotide primer combinations within cluster 7.2 (Lane 1: oAR140+oAM41; Lane 3: oAM42+oAM37; see Figure S2 for more detail). **b)** All lanes are different oligonucleotide primer combinations within cluster 3. M = 2-log DNA ladder (adapted from Moore & Russell, 2012).

Using oligonucleotides that target snoRNA gene cluster 3, we were able to detect polycistronic transcripts containing U14, Eg-p3 and Eg-m3 (Figure 16b). This situation is different than the other examples, as U14 is a predicted processing snoRNA in *Euglena* (3.3), while Eg-p3 and Eg-m3 guide rRNA modifications. This situation is similar to what

has previously been observed in maize, where U14 is polycistronically transcribed with other snoRNAs (Leader et al., 1997).

3.6 A method to prevent amplification of unwanted RNAs when constructing an RNA library for RNA-Seq

An increasingly common strategy to identify non-coding RNAs is to utilize deep sequencing technologies in the form of RNA sequencing (RNA-Seq) (Mortazavi et al., 2008; Wang et al., 2009). Briefly, a population of RNA (either total or selected, such as Poly-A enriched or size-selected) is converted to a cDNA library, utilizing adaptors added to the ends of the RNA molecules. This cDNA library is amplified by PCR and the amplified products are sequenced using high throughput sequencing technology. An advantage of enriching a sample for Poly-A RNA or selecting small RNA, is that very abundant cellular RNAs that would otherwise dominate the sequence reads (such as rRNA) can be avoided. Commercial kits have also been developed to remove rRNA prior to sequencing; however, they are only available for limited model laboratory organisms.

To identify remaining uncharacterized *Euglena* snoRNAs, as well as other small ncRNAs, we utilized an RNA-Seq approach. However, in *E. gracilis* the large subunit rRNA is naturally fragmented into 14 smaller pieces, many of which are in the size range of small ncRNAs. Therefore, most of the sequence reads from a size-selected RNA-Seq experiment would be mature LSU rRNA species or their most stable degradation products.

Using a strategy similar to that previously described for eliminating unwanted DNA sequences from environmental samples (Vestheim and Jarman, 2008), the quantity of rRNA sequence reads in the RNA-Seq data was greatly reduced. The 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 unwanted sequences. A set of blocker oligonucleotides were designed to anneal to the end of the added 5′ linker sequence (13 nt) + 5′ end of each LSU rRNA fragment (13 nt) (Figure 17a). By adding an excess (10X) of blocker oligonucleotide relative to adaptor-specific oligonucleotide, amplification of specific LSU sequences was greatly diminished (Figure 17b and Figure S7).

The efficiency of blocking was first assessed by shotgun cloning the cDNA library products and sequencing 178 clones (using Sanger sequencing). When no blocker oligonucleotides were used 40% of the sequence reads were legible and greater than 20 nt long (termed 'informative sequences') (Figure 18). Of these informative sequences, 75% were rRNA fragments. Initially, blocker oligos were only designed to the 10 smallest LSU fragments, whose sizes fell in the range of small ncRNAs. 84% of the sequence reads were informative, with 92% of them being rRNA. Most of the rRNA sequences were fragments of the largest LSU species, likely stable degradation products. Next, blocker oligonucleotides were designed to all 14 LSU fragments (and subsequently used in library construction). Now, 62% of these sequences were informative and of these, only about 5% were rRNA (Figure 18), while 37% were tRNA. This indicates that adding blocker oligonucleotides to all 14 LSU fragments at the amplification step of library synthesis greatly reduced the number of rRNA sequences in the final library. As rRNA is the most abundant cellular RNA, RNA-Seq data from total RNA samples (no rRNA depletion step) typically contain >90% rRNA reads (He et al., 2010; Chen and Duan, 2011; O'Neil et al., 2013; Peano et al., 2013). Therefore, the strategy described here is

very useful for RNA-Seq experiments in organisms for which no commercial rRNA depletion kits are available.

a)

5′ linker & 3′ tail

5′ gcugauggcgaugaaugaacacugcguuugcuggcuuugaugaaaCCAACACCCCGCCCAGACCAGGCU
GGAUCUGCGGCGAGUGUAAGUGUUCAGAGGUUAUGggg$_{(N)}$ 3′

cDNA

5′ ctcccgcttccagatctcgagc$_{(15)}$g/a/tCATAACCTCTGAACACTTACACTCGCCGCAGATCCAGC
CTGGTCTGGGCGGGGTGTTGGtttcatcaaagccagcaaacgcagtgttcattcatcgccatcagc 3′

Oligonucleotide design

b)

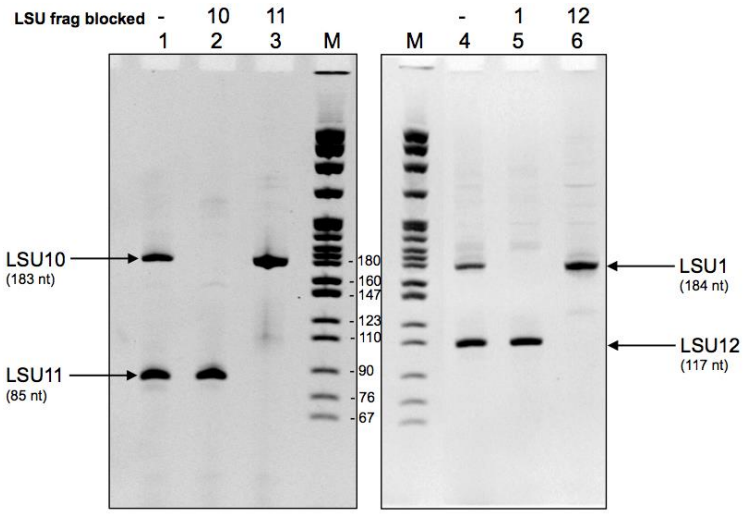


Figure 17. Primer blocking strategy to prevent rRNA amplification during small RNA library preparation. a) After the 5′ linker and 3′ poly G tail are added to the RNA, it is reverse transcribed into cDNA. 'Universal' oligonucleotides that anneal to the linker and tail are used to amplify the cDNA products. A blocker oligonucleotide primer is designed to prevent extension of one of the universal oligos. Lower case letters represent nucleotides that were added (blue = 5′ linker, red = 3′ tail), upper case letters represent rRNA sequence (black = LSU fragment 13), arrows indicate 'universal' oligonucleotides and the X-arrow represents the blocker oligonucleotide. **b)** 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.

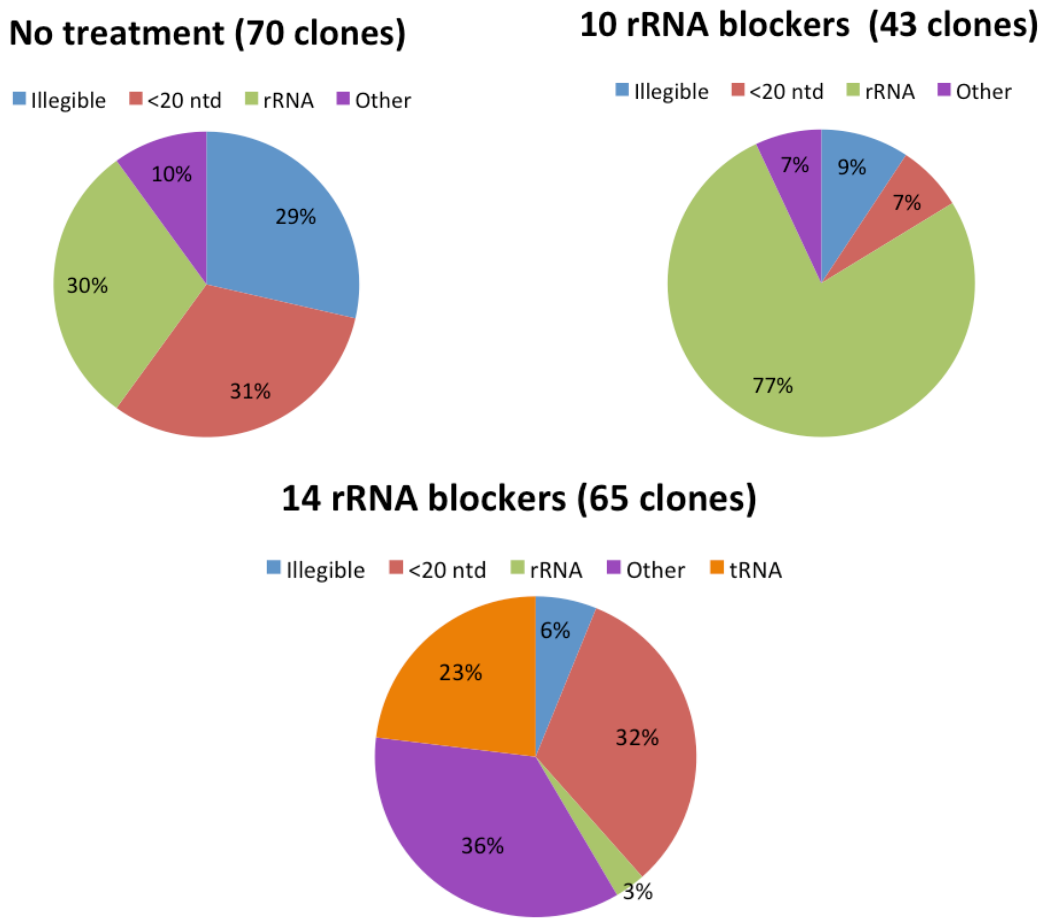


Figure 18. Sequencing results of a small RNA library after using primer blocking to prevent amplification of rRNA fragments from *Euglena gracilis*. PCR-amplified library products were cloned and sequenced using Sanger sequencing technology to assess the efficacy of the primer blocking strategy. Sequence reads that were legible and longer than 20 nt were considered informative sequences.

3.7 Identification of snoRNAs in *Euglena gracilis* by RNA-Seq

The LSU-depleted small RNA library was sequenced using paired end 250 nt sequencing on an Illumina MiSeq platform (Genome Quebec). The number of individual sequence reads after the various bioinformatic 'clean-up' steps (eliminating adapter sequences and known RNAs) and clustering of related sequences is listed in Table S14. A BLAST search was performed to find previously annotated *Euglena* RNAs in the library; approximately 4% of the reads were rRNA, 19% were snRNAs, 1.4% were known tRNAs

and <1% were known snoRNA sequences. Reads between 50 – 80 nt were scanned for snoRNA sequence and structural features. This generated 193 'matches' for candidate box C/D snoRNAs and 895 'matches' for box AGA snoRNAs. After further manual inspection of these matches and searching for corresponding modified rRNA sites, 49 novel box C/D snoRNAs and 35 box AGA RNAs were identified (Figures 9 and 10 and Figure S1). In addition, isoforms of previously identified snoRNAs were also identified, indicating that the algorithms employed for removal of known RNAs during the clean-up steps were not completely effective. Cumulatively, including all biochemically, genomically, and RNA-Seq identified RNAs, we have now characterized the snoRNAs that guide approximately 75% of the $O^{2'}$- methylated sites and 45% of pseudouridylated sites in *E. gracilis* rRNA (Russell et al., 2004, 2006; Moore and Russell, 2012).

The 49 new C/D box snoRNAs identified by RNA-Seq exhibit the same conserved features as the other box C/D RNAs in *Euglena* (see section 3.1). The predicted snoRNAs range in size from 56 – 91 nt, with an average size of 64 nt. The length of the box C/D snoRNA/rRNA interactions range from 9 – 15 bp, with an average length of 12 bp. The majority of box C/D RNAs utilize a guide element upstream of the D′ box to target modification. The nucleotide directly upstream of the adjacent D or D′ box often does not pair to the rRNA target site, and there are rarely any mismatches to disrupt the continuity of pairing between the snoRNA guide region and the rRNA. Again, no double-guide RNAs were identified. As expected, numerous isoforms of each box C/D RNA species were identified in the library, further highlighting the extraordinary abundance of snoRNAs in this organism.

Using the PCR-based approach (3.1) only 9 box AGA RNA species were identified (bringing the total number characterized in *E. gracilis* to 13), highlighting the

more difficult nature of identifying these RNAs compared to box C/D RNAs in *Euglena.* Bioinformatic analysis of the small ncRNA library has significantly increased this number to 48. The size range of the predicted box AGA RNAs is 60 – 72 nt, with an average size of 66 nt, an unusually uniform size distribution compared to Ψ guide RNAs characterized in most other eukaryotes. All of these RNAs contain a single extended hairpin structure and AGA box motif at the 3′ end, as these were the primary search parameters for bioinformatically identifying these RNAs in the library. As such, the possibility exists that there are other Ψ-guide snoRNAs in *Euglena* that more resemble "canonical" eukaryotic box H/ACA snoRNAs. It has previously been observed that the interaction between the target rRNA and the snoRNA positions the target uridine (and Ψ pocket) 13 – 16 nt from either an ACA or H box sequence (Ganot et al., 1997a; Ni et al., 1997). The *Euglena* snoRNAs described in this study share this property except for Eg-p7, Eg-p30 and Eg-p37 (17-18 nt, Figure 10). Since other isoforms of these snoRNAs may exist that more optimally position the target uridine relative to the AGA box sequence, we are hesitant to speculate that structural differences in *Euglena* Ψ-guide snoRNP structure would allow this variation.

The increased number of characterized box AGA snoRNAs now allows for a more thorough examination of the common structural features of these RNAs in *E. gracilis* (see Table S15). The basal stem (P1) is 4 – 9 bp (with an average of 7 bp) and the majority have predicted canonical base-pairing interactions. There are only seven instances where the P1 stem may be interrupted by bulged nucleotides and only two contain mismatches. The average size of the apical stem (P2) 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.

Indeed, when the nucleotide changes observed in the various isoforms of a box AGA RNA species are mapped onto the RNA secondary structure, many sequence changes occur in the P2 stem but very rarely do sequence changes disrupt the integrity of the P1 stem (Figures 19 and S3). This suggests that more sequence and structural variability is tolerated in the P2 stem in *Euglena* snoRNAs. This is consistent with what has been observed in yeast, where the P1 stem is essential for snoRNA accumulation while the P2 stem does not contribute significantly to snoRNA stability (Balakin et al., 1996; Bortolin et al., 1999). However, both stems are essential for the pseudouridylation reaction (Bortolin et al., 1999). Also noteworthy is the fact that nucleotide substitutions between isoforms often produce G•U base pairs or sequence changes in the apical loop region (Figure 19) neither of which would be expected to disrupt essential structural/sequence motifs for snoRNA functionality.

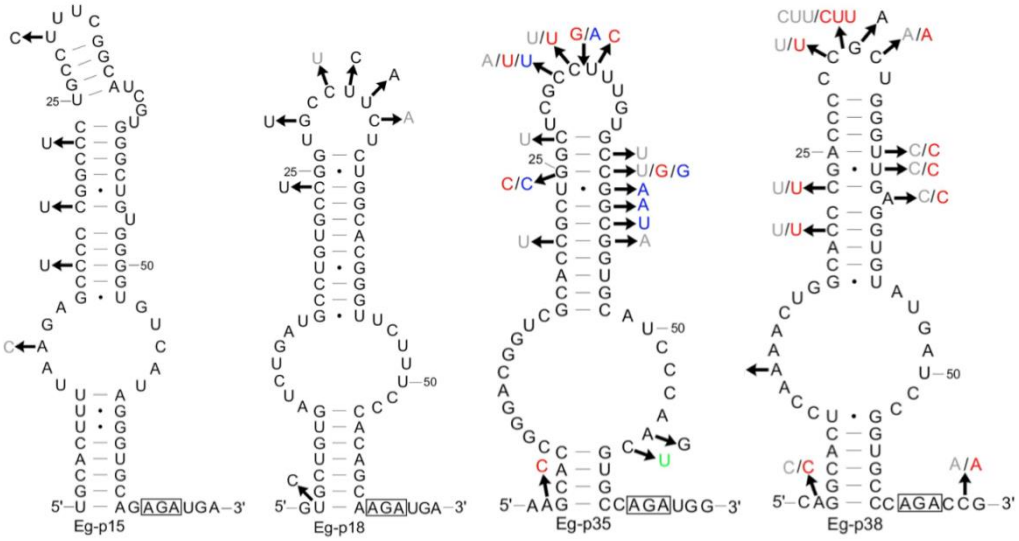


Figure 19. Examples of predicted secondary structures of *E. gracilis* box AGA modification guide snoRNAs. The boxed nucleotides highlight the AGA box elements and arrows indicate sequence variation between characterized isoforms. Nucleotide changes often occur in the stem above the pseudouridylation pocket, in the apical loop region and/or create G•U base-pairs. Each different color represents the nucleotide changes present in a single isoform. Black = Eg-p#.1; Grey = Eg-p#.2; Red = Eg-p#.3; Blue = Eg-p#.4; Green = Eg-p#.5. See Figure S3 for additional examples.

Some of the sequence reads extend more than 3 nt downstream of the AGA sequence motif and some were truncated, terminating 1 or 2 nt downstream of the box element (Figure S1). It is unlikely these represent the mature 3′ ends of these snoRNAs, as some well-characterized snoRNAs (i.e. snoRNAs whose 3′ ends were precisely mapped by 3′ RACE experiments, see section 3.1) whose sequences appeared in the library, had non-accurate 3′ ends mapped by RNA seq. Extended ends might represent processing intermediates or be the result of the tailing and sequencing procedures. As previously mentioned (see Methods), it was sometimes difficult to determine the mature 3′ end of the sequence reads during the clean-up step of analysis, as the sequence quality was very poor around the added poly-G stretch. It was noted that 50% of the newly identified box AGA snoRNA sequences contain a uridine immediately downstream of the AGA box (‘AGAUNN’) and 32% end with ‘AGAUGN’ (Figure S1).

One particularly interesting AGA box snoRNA was identified while analysing snoRNA isoforms. While examining consensus sequence reads of a bioinformatically clustered single snoRNA group (later termed Eg-p33), it was evident that the sequences were related but could in fact be further clustered into two distinct groups (Figure 20). Upon further analysis, we discovered that the sequence differences change the guide-region of the snoRNA that base-pairs with the target rRNA; thus, these 2 differentiated snoRNA clusters constitute snoRNA species that guide two different Ψ sites (the second snoRNA species is now designated Eg-p48). It appears that these Ψ-guide snoRNAs are evolutionarily related and the encoding genes have evolved by the same mechanism characterized for box C/D snoRNA and rRNA target site evolution in *Euglena* (through gene duplication followed by sequence divergence) (see section 3.4).

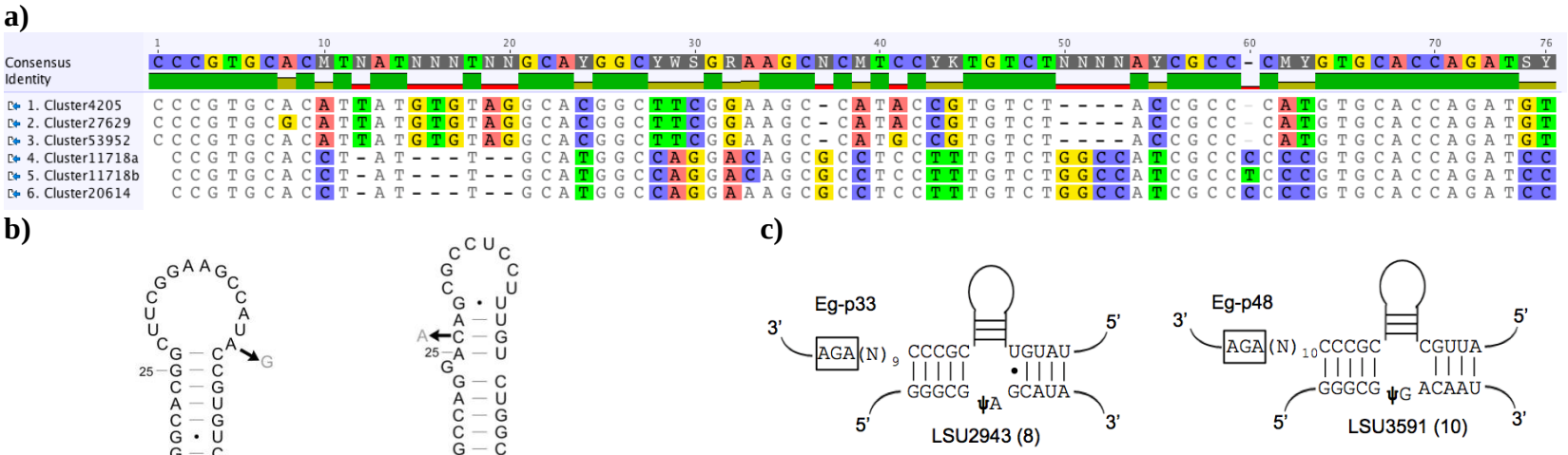


Figure 20. Evolutionarily-related AGA box snoRNA sequences.
a) Sequence alignment (Geneious software) of six related RNA-Seq reads that were bioinformatically clustered together as a single AGA box snoRNA (later denoted Eg-p33). These snoRNA sequences are obviously similar, yet there is enough sequence divergence to allow them to target different rRNA modification sites. Clusters 11718 and 20614 are clearly unique from the other three (4205, 27629 and 53952) and based on nucleotide changes that occur in the pseudouridylation pocket, constitute a novel snoRNA species (later designated Eg-p48). Cluster 4205, 27629 and 53952 = Eg-p33 isoforms. Clusters 11718 and 201614 = Eg-p48 isoforms. **b)** Predicted secondary structures of Eg-p33 and Eg-p48, with nucleotide changes mapped. Black nucleotides = Eg-p#.1; Grey nucleotides = Eg-p#.2. **c)** Eg-p33 and Eg-p48 AGA box RNAs with their predicted target rRNA pseudouridylated sites. Schematic annotation as in Figure 3.

Two of the novel snoRNAs identified by RNA-Seq target modification sites found at the 3′ extremities of rRNA species. Eg-m121 guides 2′-O-methylation at position LSU3906, which is the 3′ end of LSU fragment 12. Base pairing between the snoRNA and rRNA extends into the intergenic (spacer) region between LSU fragments 12 and 13 (Figure 21a). Eg-p38 guides Ψ formation at position SSU2305 (the 3′ end of the SSU), the base-pairing interaction between the snoRNA and rRNA extends into the ITS 1 region (Figure 21b). In fact, the entire interaction between the 5′ region of the pseudouridine pocket and the rRNA occurs in the ITS region. Therefore, the interaction between snoRNA and rRNA must occur before pre-rRNA cleavage to remove spacer regions required to generate the mature 3′ ends of these rRNA species. It is commonly suggested that snoRNA-rRNA base-pairing interactions occur very early in the ribosome biogenesis pathway (Watkins and Bohnsack, 2012) and in yeast, there is evidence that some rRNA modifications occur co-transcriptionally (Kos and Tollervey, 2010). The discovery of these *Euglena* snoRNAs provides further evidence that snoRNA binding occurs prior to some rRNA cleavage events. This is especially interesting in *E. gracilis*, as the LSU is highly modified and naturally fragmented into 14 pieces. Therefore, the coordination of these events, that is, the relative timing of individual modification and rRNA cleavage events, may be particularly important in the ribosome biogenesis pathway of this organism.

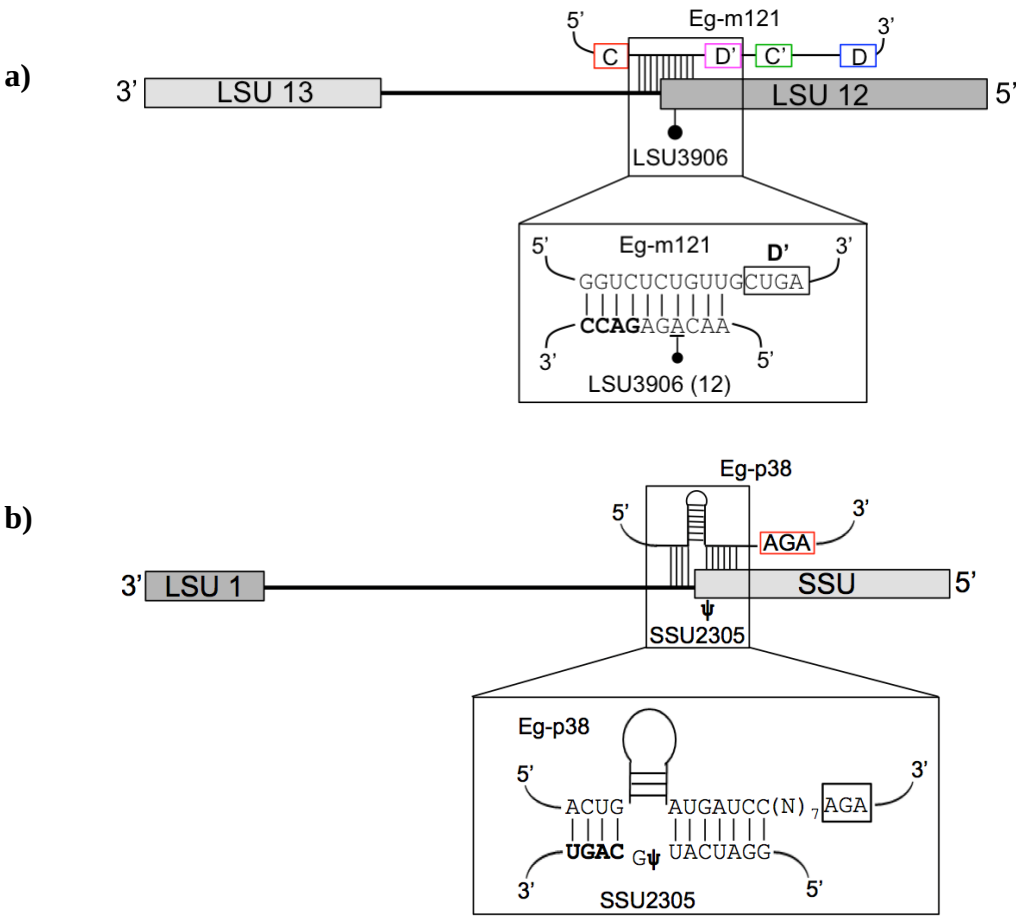


Figure 21. Identified *E. gracilis* snoRNAs whose guide regions base-pair with pre-rRNA intergenic sequence. Two snoRNAs were identified that each guide a modification found at the 3′ extremity of an rRNA subunit and as a result, their guide regions extend into intergenic sequence. **a)** snoRNA Eg-m121 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 between LSU fragments 12 and 13. **b)** snoRNA Eg-p38 guides a Ψ modification at position U2305, which is located 2 nucleotides from the mature 3′ end of the SSU. The 5′ region of the pseudouridine pocket base-pairs with the internal transcribed spacer (ITS 1) region between the 3′ end of the SSU and the 5′ end of the 5.8S rRNA (LSU 1). In the inset figures, the nucleotides found in the intergenic regions are indicated in bold.

A library of 2,2,7-trimethyl guanosine (TMG or $m^{2,2,7}G$) cap enriched RNAs was also screened for predicted *Euglena* snoRNAs. As it was unknown which RNA polymerase transcribes snoRNAs in *Euglena*, the cap-enriched library was screened for characterized snoRNA species. Of the known snoRNAs, approximately 89% of box C/D

(including U14 and U3) and 65% of box AGA RNAs (including Eg-h1) were identified. It was also confirmed that all the *E. gracilis* snRNAs (except U6) were present in the library. In other eukaryotes, snoRNAs independently transcribed by RNA pol II typically contain TMG caps, while those encoded in introns or those transcribed by RNA pol III do not (Maxwell and Fournier, 1995). As it has been reported that anti-TMG antibodies also weakly recognize m^7G and $m^{2,7}G$ caps (Simoes-Barbosa et al., 2008; Simoes-Barbosa et al., 2012), we cannot conclusively say that all of the snoRNAs present in this library contain TMG caps; however, it does confirm they are likely transcribed by RNA pol II.

A collection of predicted snoRNAs show limited, less than optimal base-pairing to mapped rRNA modification sites (Figure 22). Given the weak snoRNA-rRNA interactions, it is unclear if they actually guide these modifications. It is possible that there are additional isoforms of these snoRNAs with sequence variation in the guide regions that would improve base-pairing to modified rRNA positions. Alternatively, they may be involved in modifying other cellular RNA species (see below).

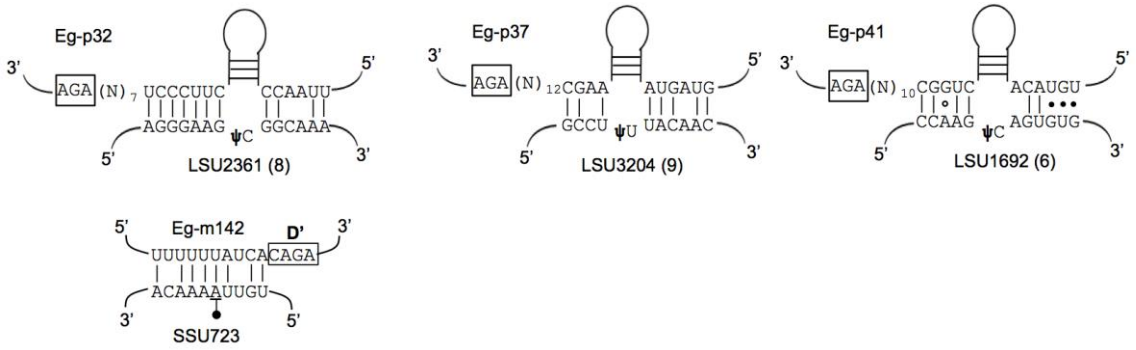


Figure 22. Potential modification-guide snoRNAs that show less than optimal base-pairing to modified rRNA sites. See Figures 2 and 3 for details on the snoRNA schematics.

In addition to snoRNAs predicted to target rRNA, numerous potential orphan snoRNAs were identified (Figure S8). These snoRNAs are similar in size to rRNA targeting snoRNAs, have canonical sequence box elements, and in the case of AGA box

RNAs, form 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 "snoRNAs". These orphan snoRNAs also do not appear to be involved in modifying snRNAs (based on predicted or experimentally mapped *Euglena* snRNA modification sites, data not shown). We currently do not know the function of these orphan snoRNAs; however, the recent discovery of pseudouridylation of mRNA species in yeast and humans (Carlile et al., 2014; Lovejoy et al., 2014; Schwartz et al., 2014) raises the exciting possibility that these orphan "snoRNAs" may be used to target modification sites in mRNA or other cellular RNA species.

3.8 Silencing RNAs in *E. gracilis*

To examine the function of predicted processing snoRNAs (such as U14 and Eg-h1) and to potentially determine the function of orphan snoRNAs in *Euglena gracilis*, an RNA-silencing approach was developed. Apparent silencing of snoRNAs using double-stranded RNA has been reported in trypanosomes (Liang et al., 2003; Gupta et al., 2010) and thus, we chose this method to attempt to silence individual snoRNA species in *Euglena.* It should be noted that in trypanosomes, the introduced double stranded RNA is produced from an *in vivo* plasmid construct. Previous studies demonstrated that RNAi is an effective method to silence the expression of some protein coding genes (such as calmodulin and photoactivated adenylyl cyclase) in *E. gracilis* (Iseki et al., 2002; Daiker et al., 2010). However, this method had never previously been utilized in our laboratory and therefore, the procedure and conditions were first optimized.

The conditions for introducing double-stranded RNA into *Euglena* by electroporation were optimized based on conditions described in previous studies (Iseki et al., 2002; Daiker et al., 2010) (Table 5). The expression of an essential *E. gracilis* protein was initially chosen to silence, as effective silencing should either be lethal or transiently inhibit growth and therefore, be easily observable. Fibrillarin is the catalytic component of box C/D snoRNP complexes and was the protein initially targeted for silencing. Anti-fibrillarin dsRNA was introduced into *Euglena* cells and the effectiveness of silencing was assessed based on the level of inhibition of cellular growth (Table S16 and Figure S9). The conditions that resulted in the least amount of growth (compared to the control cells which were treated with water) are indicated in Table 5. The growth inhibition phenotype was dependent on the sequence of the dsRNA, as *Euglena* growth was not affected by treatment with non-specific dsRNA (Figure S10).

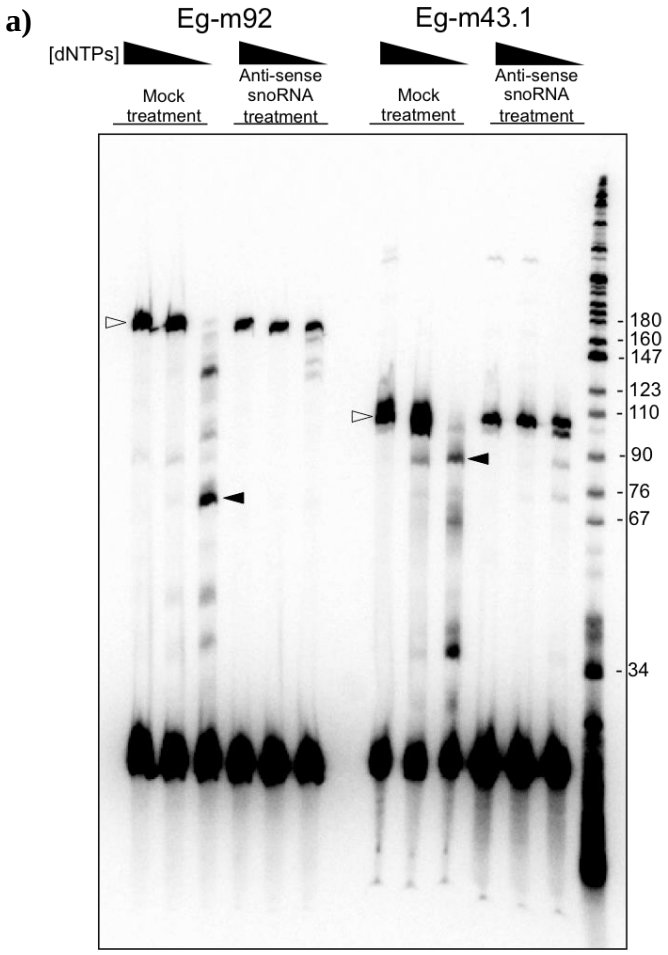
Table 5. Optimization of electroporation conditions for introducing double-stranded RNA into *Euglena gracilis*

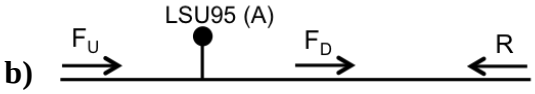
Treatment Number	Sample volume	Voltage	Incubation*
1	100 µL	1.2 kV	0 min
2	**100 µL**	**1.2 kV**	**10 min on ice**
3	100 µL	1.5 kV	0 min
4	100 µL	1.5 kV	10 min on ice
5	500 µL	1.2 kV	0 min
6	500 µL	1.2 kV	10 min on ice
7	500 µL	1.5 kV	0 min
8	500 µL	1.5 kV	10 min on ice

* Incubation refers to the time after electroporation that the samples were incubated prior to adding fresh media. The optimal conditions are boxed and indicated in bold text.

The same protocol was followed to silence snoRNAs, but because many single modification-guide snoRNAs are non-essential in other eukaryotes, effectiveness of silencing was instead assessed based on the status of the snoRNA's target 2′-O-

methylation site. We initially chose to silence snoRNAs targeting rRNA sites in regions with sparsely distributed modifications. Initial results suggested that Eg-m92 and Eg-m43.1 were successfully silenced, as their target rRNA sites were no longer 2′-O-methylated after treatment (Figure 23a). This effect was observed 9 and 25 days post-treatment (Figure S11). Methylation status was examined using the conventional reverse transcriptase dNTP concentration dependant detection of methylation sites and in addition, the successful silencing of snoRNA Eg-m92 was verified using an alternate technique for mapping 2′-O-methylated sites called reverse transcription at low dNTP concentrations followed by PCR (RTL-P) (Dong et al., 2012) (Figure 23b,c). Frustratingly, all subsequent attempts to reproduce these initial results, to silence other snoRNAs, or to silence snRNAs or other ncRNAs with antisense RNA were unsuccessful. Possible explanations for these inconsistent results are discussed later (see Discussion).


a)
Eg-m92
Eg-m43.1
[dNTPs]
Mock treatment
Anti-sense snoRNA treatment
Mock treatment
Anti-sense snoRNA treatment
180
160
147
123
110
90
76
67
34


b)
LSU95 (A)
F_U
F_D
R

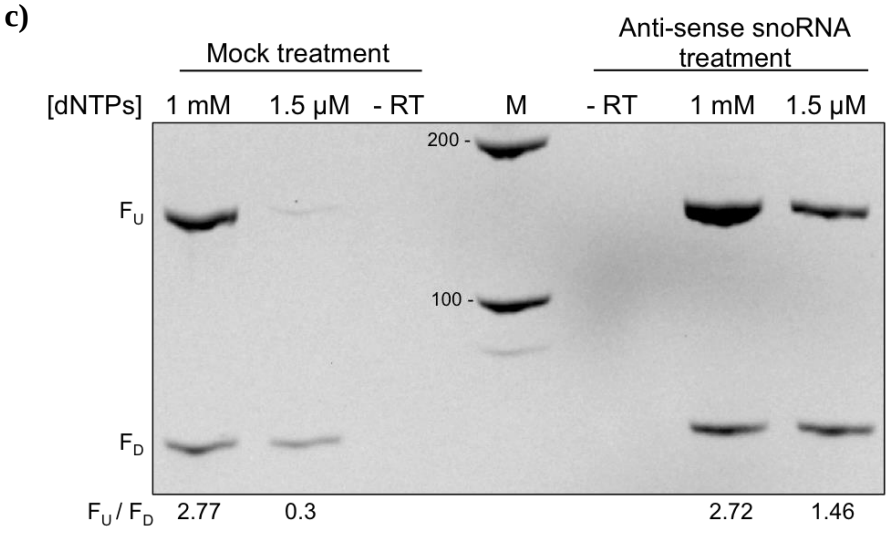

c)
Mock treatment
Anti-sense snoRNA treatment
[dNTPs] 1 mM 1.5 µM - RT M - RT 1 mM 1.5 µM
200 -
100 -
F_U
F_D
F_U / F_D 2.77 0.3 2.72 1.46

Figure 23. Methylation status of rRNA target sites after snoRNA silencing. To determine whether a specific snoRNA had been silenced in *E. gracilis* after treatment with dsRNA, the methylation status of its corresponding rRNA target site was assessed. **a)** Primer extension with decreasing concentrations of dNTPs. The snoRNA targeted for silencing is indicated. Mock treatment = RNA template isolated from control cells treated with water; anti-snoRNA treatment = RNA template isolated from cells treated with dsRNA targeting the snoRNA species indicated. The products indicated with black arrow heads are the 2′-O-methylation sites guided by the targeted snoRNAs. The white arrow heads indicate products that are the result of predicted reverse transcriptase stop sites: either the end of a LSU fragment or a mapped m^1A modified site. **b)** Oligonucleotide primer design to map 2′-O-methylated sites using RTL-P. The arrows indicate oligonucleotides. F_U = forward upstream, F_D = forward downstream, R = reverse. The filled circle indicates the 2′-O-methylation site guided by Eg-m92. **c)** Mapping the 2′-O-methylation site guided by Eg-m92 using RTL-P. The products have been separated on a 15% native polyacrylamide gel. F_U / F_D band intensity ratios were estimated using ImageJ software.

Chapter 4: Discussion

In this study, the organization of snoRNA genes in *E. gracilis* was characterized and through bioinformatic analysis, a significant number of new snoRNA sequences and isoforms were identified. Predicted rRNA targets for nucleotide modification were identified for the majority of the new snoRNAs, while some were classified as predicted (or 'orphan') snoRNAs. A prevalent gene organization mode in *Euglena* is tandemly repeated clusters encoding several different snoRNA species, and sometimes containing both box C/D and AGA (box H/ACA-like) snoRNAs. Many of these clusters are expressed initially as polycistronic RNAs with individual snoRNAs being released by an uncharacterized pathway.

The majority of the clusters identified thus far appear to contain only box C/D snoRNAs. The *Euglena* rRNA contains significantly more $O^{2'}$-methyl than Ψ sites (Schnare and Gray, 2011). Therefore, this organism should have a larger collection of box C/D than box AGA snoRNAs. Including those identified in this study, plus their associated isoforms and those previously identified (Russell et al., 2004, 2006), box C/D snoRNA sequences predicted to target 151 of the 211 total methylation sites have now been identified. On the other hand, only 48 AGA box RNA species (all single-guides) have been successfully identified thus far predicted to target 48 of 119 total Ψ sites. While our genomic analysis appears to support the prediction that *Euglena* will contain a larger set of box C/D RNAs, most oligonucleotide primers designed for the PCR-based genomic cluster amplification strategy were necessarily based on previously identified box C/D rather than box AGA RNA sequences (as only 4 were previously characterized). Since the majority of the newly identified AGA box RNAs were characterized from the

small RNA library, their genomic context is unknown. Subsequent PCR-based genomic cluster amplification experiments and BAC library analysis may reveal that some clusters contain exclusively box AGA snoRNAs.

The strategy used in this study to reduce the amplification of LSU rRNA fragments during small RNA library construction was extremely effective. When blocker oligonucleotides were designed to all 14 LSU fragments, only 3% of the sequence reads were matches to rRNA. The abundance of rRNA contamination (up to 90% of the sequence reads can be rRNA, often due to rRNA degradation products) in small RNA libraries is a common problem and many commercial kits have been developed to mitigate this issue (He et al., 2010; Chen and Duan, 2011; O'Neil et al., 2013; Peano et al., 2013). These kits are typically based on specific rRNA sequences and therefore are only available for the most frequently studied model organisms. For organisms such as *E. gracilis* where no such kits are available, an effective method to decrease the number of rRNA sequence reads in a library preparation was essential. Considering that in *Euglena* many mature, stable LSU fragments reside in the same size range as small RNAs, unwanted rRNA sequence reads would be even more problematic than for most other organisms.

The strategy used in this study was based on the previously described method to eliminate unwanted DNA sequences from environmental samples (Vestheim and Jarman, 2008). Other methods such as targeted RNase H digestion (Liu et al., 2009a) were tested; however, the primer blocking strategy proved the most effective. This approach could be used in any organism to limit the amplification of unwanted RNA species during library synthesis to increase the proportion of sequence reads that are novel and informative. This

would allow for RNomics in less studied species (such as protists) where there is currently a lack of information regarding the abundance and diversity of ncRNAs.

To identify novel snoRNA species in the rRNA-depleted small RNA library, sequences in the appropriate size range were scanned for snoRNA features using the pattern matching program 'Scan for Matches' (Dsouza et al., 1997). The 'hits' were manually inspected for their ability to form conserved secondary structures, maintain the sequence elements, and base-pair to the mapped rRNA modification sites. Programs such as snoScan (Lowe and Eddy, 1999), snoSeeker (Yang et al., 2006), snoGPS (Schattner et al., 2004) and Psiscan (Myslyuk et al., 2008) have been developed to search for snoRNAs in sequence data libraries. However, they are based on conserved yeast or mammalian snoRNA features, many of which are not present in *E. gracilis* snoRNAs. Furthermore, some of these programs also use complementarity to these organism's rRNA as a limiting factor during the search. Some of these programs were tested for their ability to successfully find already characterized *Euglena* snoRNA sequences in the library sequence data, but they were largely unsuccessful. Therefore, we found that manually inspecting the initial 'Scan for Matches' hits provided the best overall results. This strategy is based on the features of the previously identified snoRNAs in *E. gracilis*, hence, it is possible that the remaining "missing" snoRNAs may be structurally distinctive from those already characterized. For example, perhaps a set of *Euglena* Ψ guide RNAs will contain double stem-loop structures and thus appear more 'canonical'. Also, if a bioinformatic program could be designed that would scan the library sequence reads for conserved features of *Euglena* snoRNAs and also include sequence complementarity to mapped rRNA modification sites as an additional search parameter,

more rigorous and efficient screening of the library data would be possible. This could help identify the "missing" snoRNAs that guide the remaining rRNA modifications. It is also possible that some *Euglena* rRNA modifications are guided by stand-alone protein enzymes. A genome-wide search for snoRNAs in the protist *T. brucei* found some modification sites with no apparent corresponding snoRNA and it was proposed that these modifications might be carried out by protein enzymes alone (Liang et al., 2005).

A number of orphan snoRNAs that do not target the known rRNA modification sites were also identified in the library. As we learn more about the expanding roles of snoRNAs in eukaryotic cells, other targets for these RNAs may be revealed, and that they could be involved in other cellular functions remains an exciting possibility. With the recent findings that mRNA molecules are pseudouridylated in yeast and human cells, it is possible that these snoRNAs target other cellular RNAs, such as mRNA, for modification. Alternatively, they could be involved in *Euglena*'s unique rRNA processing/assembly pathway or may be processed into other types of ncRNAs. Experiments to determine what the orphan snoRNAs are binding to *in vivo* would be important for elucidating their cellular functions. For example, a novel method termed 'RNA Walk' was developed to identify RNA-RNA interactions *in vivo* that combines cross-linking, affinity selection, and quantitative real-time PCR (Lustig et al., 2010; Wachtel and Michaeli, 2011). In addition, silencing these RNAs and observing the resulting cellular effect may also help determine their function (see below for further discussion). As an example, silencing a snoRNA predicted to play a role in rRNA processing should result in the accumulation of specific rRNA precursors and provide evidence about which step of the pathway the snoRNA acts. Similar approaches could also be used to elucidate the functions of the candidate novel ncRNAs from the small

RNA library that show no obvious homology or secondary structural similarity to other characterized classes of eukaryotic RNAs.

Characterization of the genomic regions encoding snoRNAs revealed that tandemly repeated snoRNA gene arrays can be surprisingly large in *E. gracilis*. Repetitive snoRNA gene clusters have been characterized in other eukaryotic species. For example, in the protist *T. brucei*, a cluster is repeated 7.5 times in the genome (Liang et al., 2005) and in *O. sativa* one cluster of snoRNA genes is repeated 5 times, a surprising discovery at the time (Chen et al., 2003). We have shown that snoRNA gene arrays, composed of tandemly repeated gene clusters, can be unusually large in *Euglena* – one of the clusters is predicted to be repeated 50 times over a 40 kb region. Why *Euglena* has so many snoRNA genes is puzzling. One possibility is that gene copy number is related to RNA expression levels. snoRNA genes are highly transcribed in other eukaryotes, either by localizing to highly expressed genes or by being expressed independently under strong promoters (Dieci et al., 2009). Another mechanism to ensure high expression levels is multiple copies of each snoRNA gene. This may be especially important in *Euglena* as the LSU rRNA is fragmented into 14 individual pieces and successful modification of every targeted site may be necessary for sufficient stabilization during ribosome assembly. Mapping studies have shown that nearly every targeted rRNA site is completely modified in the rRNA pool in *Euglena* (Schnare and Gray, 2011), suggesting that the modification process is highly efficient. Having multiple copies of each snoRNA might be important to ensure an efficient modification pathway. In addition to the role of the nucleotide modifications in ribosome assembly, it has been proposed that the actual binding event between snoRNA and rRNA is also important for stabilization during *Euglena*'s unique rRNA processing pathway (Schnare and Gray, 2011). As evidenced by

my study, some modification-guide snoRNAs interact with the spacer sequences found between mature rRNA sequences in the precursor rRNA transcripts (Figure 21). This may contribute to this stabilization effect and possibly be important for rRNA processing. Multiple copies of each snoRNA may increase this effect making ribosome assembly more efficient. Efficient ribosome assembly under variable environmental conditions may be of particular importance for *Euglena gracilis*, as it is a highly resilient organism that can withstand unfavorable living conditions. *E. gracilis* can thrive under acidic conditions (Chae et al., 2006), can withstand high concentrations of pollutants such as chromium (dos Santos et al., 2007) and even survive doses of ionizing radiation (Hayashi et al., 2004). Highly modified rRNA may be important for ribosome assembly and function under these highly variable growth conditions.

An alternative explanation is that multiple snoRNA gene copies are simply the result of a highly repetitive genome. From *E. gracilis* EST data it is known that some protein-coding genes are multi-copy (O'Brien et al., 2007) as well as the genes encoding other ncRNAs such as snRNAs and the processing snoRNA U3 (Charette and Gray, 2009). Early DNA re-association kinetic studies also suggest that the genome is highly repetitive (Rawson et al., 1979). Therefore, it is possible that the high number of snoRNA gene repeats is the result of a highly repetitive genome, where gene duplication events are common (as appears to be the case in *E. gracilis*). If this is true, the numerous snoRNA gene copies may not necessarily provide a benefit but are also not a burden to the cell. However, this explanation seems less likely because the snoRNA gene isoforms are often highly conserved, especially in regions of the RNA most important for function, suggesting there is selective pressure to maintain sequence conservation and that the collection of isoforms of a snoRNA species are indeed functional and beneficial.

The genomic context of the large snoRNA gene arrays remains unknown. *In situ* hybridization experiments would help determine if these arrays are spread across multiple chromosomes or if they localize to a limited number of distinct loci. We initially speculated that the snoRNA genes may be encoded on extrachromosomal circular DNA molecules similar to the rDNA circle in *Euglena.* However, after specific nuclease digestion of linear DNA with Plasmid-Safe-ATP-Dependent DNase (Epicentre®), snoRNA (and protein-coding) genes could no longer be amplified by PCR, whereas rDNA was successfully amplified, which ruled out this possibility (data not shown). The sequences flanking the snoRNA arrays were searched for potential regulatory elements, but considering how very little is known about nuclear gene expression in *Euglena*, this proved difficult. BLAST searches of these regions revealed no significant similarities to available genomic and EST sequences. Additionally, the sequences were compared to one another to determine if any conserved sequence motifs were present. An 11 nt sequence motif (showing 80% sequence identity) was found downstream of the last snoRNA gene in three of the characterized arrays (data not shown). The distance from the most 3′ snoRNA gene in the array to the downstream motif was not conserved. Furthermore, the motif was also found in some of the intergenic regions between snoRNA genes making it unlikely that it is a transcription termination signal. It is possible this motif is important for pre-snoRNA transcript processing; however, it is not highly conserved amongst the numerous snoRNA gene clusters that have been identified thus far. Further investigation into the potential importance of this motif is required (see below).

In this study, it was determined that snoRNA genes of both classes can be expressed as polycistronic transcripts and be together on the same polycistronic transcript

in *Euglena gracilis*. While it is difficult to accurately predict how large these precursor transcripts may be, there is evidence they can be quite large and contain at least 4 unique snoRNA species (Figure 16). How the individual snoRNAs are processed from the precursor transcript has yet to be determined. Comparison of the intervening sequence between snoRNA genes reveals no sequence or structural conservation. The sequences downstream of those few box AGA RNAs whose genomic context is known (those that were identified in genomic repeats), have the ability to form stem-loop structures that range in size from 22-35 nt (data not shown). In yeast, polycistronic precursor transcripts are processed via the recognition and cleavage of conserved stem loop elements located in the intervening sequence between snoRNAs (Qu et al., 1999). The RNase III-like enzyme Rnt1p recognizes and cleaves a conserved 'AGNN' tetraloop structure (Ghazal et al., 2005). However, this element is not found in the predicted stem loop structures downstream of *Euglena* box AGA RNAs. The loop regions vary in size from 5-11 nt and show no sequence conservation. Therefore, if the stem loops are important for snoRNA processing, the pathway is likely different from that observed in yeast. Similarly, in plants the processing pathway of polycistronic snoRNA transcripts is not well understood but appears to be different than the pathway in yeast (Brown et al., 2008).

The sequence motif found downstream of three of the large snoRNA gene arrays may also play a role in snoRNA processing. If it is important for processing it may be unique to just some snoRNA genes. Studies that map the 3′ ends of precursor snoRNA transcripts may help determine if this motif (and/or the stem-loop structure found downstream of some AGA box RNAs) is part of the snoRNA processing pathway in *Euglena*. This type of analysis was recently performed in a study that characterized a

novel ncRNA 3′ end processing motif in the organism *G. lamblia* (Hudson et al., 2012). There is currently no developed system to manipulate *E. gracilis* gene expression *in vivo*. If an expression vector was developed, then snoRNA expression, including promoter and terminator characterization as well as snoRNA processing pathways, could be studied in more detail.

Sequencing of TMG-enriched RNAs provided additional information on the expression of snoRNA genes in *Euglena.* As the majority of identified snoRNAs were also found in the TMG-enriched library, it is likely that they are being transcribed by RNA pol II. However, no canonical eukaryotic pol II promoters could be identified in the genomic regions surrounding the snoRNA gene arrays. The presence of a TMG cap suggests that these snoRNAs are not intronic because if their maturation was dependent on the processing of intronic sequence, it is expected that the ends would have a 5′-monophosphate (Simoes-Barbosa et al., 2012). This has been validated in other eukaryotes where intronic snoRNAs do not contain TMG caps (Maxwell and Fournier, 1995). What is unusual, however, is that snoRNAs processed from polycistronic transcripts in other eukaryotes typically do not contain TMG caps (Maxwell and Fournier, 1995). Many of the snoRNAs found to be expressed as polycistronic transcripts in *Euglena* were also present in the TMG-enriched library. Again, this suggests a unique snoRNA processing pathway in this organism. Alternatively, due to the "redundancy" of snoRNAs in *Euglena*, it is possible that some isoforms of each snoRNA species are capped while others are not (which may reflect a difference in genomic organization and/or expression). The enzyme responsible for cap hypermethylation in other eukaryotes (Tgs1) has not been identified in *Euglena* and searches of the current EST database did

not yield any positive hits. It should be noted that this database may not be completely representative of all expressed mRNAs. A recent study in the divergent unicellular eukaryote *T. vaginalis* demonstrated that box H/ACA snoRNAs are a preferred substrate for Tgs1 in this organism (Simoes-Barbosa et al., 2012), while snRNAs do not contain the canonical 5′ cap (Simoes-Barbosa et al., 2008). Evidently in some protist organisms, the capping status of certain RNAs, and possibly the capping pathway itself, is unique compared to what is observed in model eukaryotes and such may also be the case in *E. gracilis*.

Interestingly, the processing snoRNA U3 was also found in the TMG-enriched library. A previous study hypothesized that U3 would be transcribed by RNA pol III in *Euglena* because of the relationship between kinetoplastids, where U3 is transcribed by RNA pol III, and euglenids (Charette and Gray, 2009). Our data instead suggests that it is being transcribed by RNA pol II, as is observed in vertebrates, invertebrates and yeast (Jawdekar and Henry, 2008). Other predicted processing snoRNAs, namely U14 (see below) and Eg-h1, were also found in the TMG library suggesting that both modification-guide and some processing snoRNAs are transcribed by the same RNA polymerase.

In this study, an apparent U14 snoRNA homolog was identified in *E. gracilis*. It was identified through analysis of the intervening sequence between tandem copies of the Ψ-guide RNA Eg-p3 (see Figure S2, Cluster 3). It is clustered together with both a box AGA and a box C/D snoRNA, and was found to exist in multiple isoforms (Figure S1). Overall, the *E. gracilis* U14 shows little sequence conservation with those characterized in other organisms, but still retains the ability to use Domain A to base-pair to the equivalent region of the 18S rRNA. However, it lacks Domain B base-pairing potential to

the 18S rRNA site known to be targeted for methylation by this domain of the U14 in *S. cerevisiae*, *H. sapiens* and *O. sativa* (Figure 14). Since *D. melanogaster* and *A. gambiae* instead employ a second distinct box C/D snoRNA for targeting this 18S rRNA site and like *Euglena* lack the U14 Domain B base-pairing potential, we considered the possibility that *Euglena* would also employ a distinct snoRNA for this function. However, complete rRNA modification mapping data detects no $O^{2'}$- methylation in *Euglena* at this 18S rRNA position (Schnare and Gray, 2011).

Recently, a bioinformatic search in *T. brucei* failed to identify a "canonical" U14 homolog containing Domain A (Gupta et al., 2010). Likewise, we were unable to identify U14-like sequences in the *T. brucei* genome when we used the *Euglena* U14 sequence to refine the search. The region of the 18S rRNA that normally base-pairs to Domain A has diverged significantly in *T. brucei*. It has been suggested that the snoRNA TB11Cs2C1, that has been shown to interact with a different region near the 3′ end of the *T. brucei* SSU rRNA, may be functionally analogous to U14 (Gupta et al., 2010). This RNA shows no structural or sequence resemblance to previously characterized U14 snoRNAs or the *Euglena* U14 identified in the current study.

The most parsimonious explanation for the evolution of U14 given its current distribution and functional divergence in eukaryotes is that the ancestral function of this RNA involved both specifying RNA cleavage events through its SSU rRNA interaction and likely also specifying $O^{2'}$-methylations using Domain B. Most eukaryotes, including *Euglena*, have maintained the cleavage-specifying function perhaps indicating its more critical biological function. Sporadic loss of Domain B functionally has occurred on several occasions sometimes being replaced by employing a distinct snoRNA molecule, as is the case in some insect lineages. In the most extreme case, seen in *T. brucei*, the U14

snoRNA has either been lost altogether or diverged so substantially as to be unrecognizable as a U14 ortholog.

The overall organization of the *Euglena* U14 and modification-guide snoRNA genes is significantly different from that of the U3 snoRNA genes (Charette and Gray, 2009). There is no indication of polycistronic U3 transcripts and those genes most closely-linked to U3, including both tRNAs and U5 snRNA, are encoded in the opposite transcriptional orientation (Charette and Gray, 2009). In this study, we found that U14 is expressed polycistronically with modification-guide RNAs belonging to both snoRNA classes (Figure 16). It is interesting that U14 and U3, both belonging to the sub-class of box C/D snoRNAs involved in pre-rRNA cleavage events, show such differences and we speculate that a different processing pathway may be employed for expression of these ncRNAs in *Euglena* (although it appears they are both transcribed by the same RNA polymerase).

Both large scale and localized genomic duplication events are an important factor influencing sequence and functional evolution of genes in all domains of life. Genome sequence analysis, between-species genome comparison, and experimentation have revealed mechanisms by which duplicated genes diverge in function (Conant and Wolfe, 2008). Due to the frequency of polyploidy in plants, gene duplications and potential redundancy are very common (Brown et al., 2003). Based on the prevalence of snoRNA isoforms, multiple genes encoding other non-coding RNAs such as snRNAs, as well as an observed redundancy in some protein-coding genes as detected through analysis of EST data, it appears that *Euglena gracilis* is similar to plants in this regard (ploidy state in *E. gracilis* is not known). This study highlights seven distinct examples of snoRNA and modification target site evolution being influenced through predicted gene duplication

events. Sequence drift that occurs between duplicated snoRNA genes that affects the specific region involved in base-pairing to rRNA and/or its distance to the adjacent conserved box element (D or D′) can alter modification site targeting. These examples demonstrate how duplicated snoRNA genes can evolve novel functions, an example of neofunctionalization. Comparable examples have also been observed in *Arabidopsis thaliana* (Barneche et al., 2001; Brown et al., 2001) and *C. elegans* (Zemann et al., 2006).

While *E. gracilis* rRNA has a large number of $O^{2'}$-methylations, many of these are clustered together and occur at positions not known to be modified in other organisms examined. This indicates that the emergence of new snoRNAs and novel modification sites is relatively "easy" for *Euglena* (Russell et al., 2006). Our characterization of the arrangement of snoRNA genes and more specifically the presence of many gene isoforms rationalizes this observation. We predict that frequent gene duplication is a common mechanism driving snoRNA emergence and evolution in *Euglena* that has resulted in both the large number and clustered patterning of rRNA modification sites. Gene duplication events also appear to be the source of the multiple copies of the U3 snoRNA gene characterized in *Euglena* (Charette and Gray, 2009), which is particularly evident when comparing sequences of the multiple closely-related U3 snoRNA-U5 snRNA linkage units. We also note that given the apparent rapid rate of snoRNA isoform evolution in this organism (see for example isoforms of Eg-m14 and Eg-m59 in Figure S1), many more adjacent snoRNA genes within the clusters that we have characterized may in fact be related through more ancient duplication events, which are no longer readily detected by simple nucleotide sequence comparisons.

Particularly striking and unexpected is the lack of double-guide RNAs in *E. gracilis*. Previous studies identified only two double-guide box C/D snoRNAs (Russell et

al., 2006) and no double ψ-guide RNAs have been identified in this organism to date. Despite identifying a large collection of new snoRNAs in my study, no additional double-guide box C/D snoRNAs were identified, only isoforms of previously characterized ones. Although this is consistent with what is observed in other eukaryotes, where these RNAs are less common than in archaea, the relative fraction of double-guide RNAs in *E. gracilis* is unusually low. One possibility is that as snoRNA genes duplicate and evolve (a common occurrence in *E. gracilis*), modification sites that were ancestrally targeted by double-guide snoRNAs instead become guided by two separate, but evolutionarily-related, snoRNAs (an example of subfunctionalization). If the methylation sites are far apart in the rRNA, using one double-guide RNA to target both sites may have been inefficient. By utilizing two unique snoRNAs, methylation efficiency may be improved and therefore, this method would be beneficially selected thus resulting in the observed scarcity of double-guide box C/D snoRNAs in *Euglena gracilis*.

The lack of double-stem ψ-guide RNAs in *E. gracilis* is unique among Euglenozoa (which includes the trypanosomes). In fact, single-stem ψ-guide RNAs containing an AGA box element are more similar to those observed in archaeal organisms (Rozhdestvensky et al., 2003; Tang et al., 2002). Interestingly, AGA box elements, as a sequence variant of the 3′ ACA box, are rarely found in eukaryotic double-stem ψ-guide RNAs, while the conversion of the ACA element to AAA and AUA has been observed. In fact, when the ACA box element has been artificially mutated to AGA in yeast, the mutated snoRNA species do not accumulate, whereas the presence of an AAA or AUA had little effect on accumulation (Balakin et al., 1996). This suggests that single-stem ψ-guide RNAs with an AGA box element may utilize a unique processing or assembly

pathway compared to double-stem RNAs. It has also been suggested that AGA box snoRNAs represent the functional equivalent of the 5′ end of canonical eukaryotic H/ACA RNAs, since the H-box element is often ‘AGANNA’ (Russell et al., 2004).

Some phylogenetic studies place *Euglena gracilis* as a very early-branching eukaryote (Cavalier-Smith, 2013) and therefore, the single-stem structure may represent the primordial ψ-guide RNA structure that was prevalent before the divergence of archaea and eukarya. The fusion of two single-stem RNAs to form the “canonical” double-stem structure observed in other eukaryotes may therefore have occurred after the divergence of Euglenozoa. Consistent with this model, ψ-guide snoRNAs in *T. brucei* (a distant but specific relative to *E. gracilis*) also contain a single-stem and AGA box element. Furthermore, the finding that two sets of core RNP proteins autonomously bind to each stem-loop element lends further support to such an evolutionary scenario. Characterization of snoRNAs in a wider range of Euglenozoa species, as well as other potential early-branching eukaryotes, could provide further support for this ‘fusion theory’.

The initial successful silencing of targeted snoRNAs by introducing double-stranded RNAs into *E. gracilis* cells appeared promising. The clear loss of the corresponding 2′-O-methylation site (Figure 23) suggests that the targeted snoRNAs were no longer functional. The snoRNAs initially chosen for silencing were ones that target rRNA sites in regions that are sparsely modified. This was to ensure we could clearly map the modified site and readily detect the loss of modification. Furthermore, because many rRNA modification sites are clustered together in *E. gracilis* and multiple, unique snoRNAs have overlapping rRNA base-pairing regions (Figure 7), it is hypothesized that

larger structurally dynamic complexes containing more than one snoRNP may form *in vivo*. If this is the case, the snoRNA being targeted for silencing might be temporally inaccessible and unable to base-pair with the introduced antisense RNA, preventing it from being silenced. If higher-order snoRNP complexes are indeed forming, silencing one snoRNA may also affect the modification of surrounding sites. Therefore, snoRNAs that target rRNA regions containing highly clustered modification sites were also targeted for silencing; however, no loss of modification was detected at any of the sites. While the initial results suggested the targeted snoRNAs were successfully silenced, multiple subsequent attempts to reproduce these results were unsuccessful. Numerous other ncRNAs were targeted for silencing including additional modification-guide snoRNAs, processing snoRNAs and snRNAs, none of which provided any indication that the knock-down strategy was successful.

As the results of this study demonstrate, snoRNA genes can exist in unusually large and repetitive arrays, which could explain the unsuccessful attempts and inconsistency in silencing snoRNAs in *Euglena*. It is likely that most of the snoRNA genes in these arrays are expressed, as the functional regions show a high level of conservation suggesting they are not pseudogenes. Furthermore, the majority of characterized snoRNAs are present as numerous, distinct isoforms. The interaction between antisense RNA and target RNA in the RNAi pathway is sequence specific; therefore, an introduced antisense snoRNA is only going to target one "sufficiently complementary" isoform species for silencing. In mammalian cells, the silencing of ncRNAs with antisense RNA is highly sequence specific – if even two mismatches are present in the RNA duplex, silencing does not occur (Liang et al., 2011). If this is also the case in *E. gracilis* cells, the targeted snoRNA isoform may be silenced but numerous

additional isoforms are not and therefore, the corresponding rRNA site would still be modified. Whether or not different isoforms are expressed at different levels dependent on growth conditions is also not known, and could potentially result in different impacts on rRNA modification levels at a particular site when targeting single isoforms for depletion. In addition to detecting snoRNA depletion by assessing the modification status of the corresponding rRNA site, Northern blotting was used in an attempt to monitor snoRNA levels; while the expression of other *E. gracilis* RNAs could be detected, snoRNAs could not be detected using this technique (data not shown).

We initially chose this method to silence snoRNAs based on studies done in trypanosome species. In those studies, instead of introducing exogenous RNA into the cells via electroporation, antisense or double-stranded RNA molecules were expressed in the cells utilizing an established trypanosome expression vector (Liang et al., 2003). The results demonstrated that mature snoRNA species, but not precursor molecules, could be silenced using an RNAi-like approach (a technique termed 'snoRNAi' by the authors) (Gupta et al., 2010). However, the efficiency of silencing varied significantly depending on the snoRNA targeted, suggesting different levels of accessibility between snoRNA species. Furthermore, two targeted snoRNAs could not be silenced at all (Liang et al., 2003). In contrast to what was reported in trypanosomes, studies in mammalian cells demonstrated that RNAi is not an effective tool to knock down snoRNAs (Ploner et al., 2009). However, a different nuclear ncRNA (7SK) has been successfully silenced using an RNAi approach in mammalian cells (Robb et al., 2005). It appears that the silencing of nuclear RNAs is challenging and the likelihood of successful knock down is dependent on numerous factors including sequence specificity, accessibility, and cellular localization.

Due to their small size, nucleolar localization and complex structure, it has been suggested that snoRNAs may be the most difficult eukaryotic knock-down targets (Ploner et al., 2009). Therefore, the efficacies of three different methods to silence human snoRNAs were tested, which included RNAi, a ribozyme-based strategy and locked nucleic acid (LNA) antisense oligonucleotides. The latter two methods produced the best knock-down effect while RNAi was found to be an unsuitable tool. Interestingly, while the cellular levels of the snoRNA decreased (by up to 60%), the enzymatic activity of the targeted snoRNP complex was unaffected - in other words, the level of modification did not change. This is likely due to the remaining 40% of uncleaved snoRNAs that were still capable of modifying the target site (Ploner et al., 2009). This may be similar to what I predict has been occurring in the RNAi experiments in *Euglena*. Similar studies in mice and human cell lines used chemically modified antisense oligonucleotides to silence snoRNAs and showed up to 90% reduction in RNA levels and resultant loss of the corresponding modification sites (Ideue et al., 2009; Liang et al., 2011). Collectively, these results suggest that alternative techniques for silencing snoRNAs in *E. gracilis*, such as ribozymes and modified antisense oligonucleotides, could be explored as more effective approaches to study the role of orphan and predicted processing snoRNAs, and to examine the complex patterns of modification targeting in the densely clustered regions of *E. gracilis* rRNA.

Chapter 5: Summary

The focus of this research project was to characterize snoRNAs in the protist organism *Euglena gracilis* to better understand the diversity of genomic organization and expression strategies observed in unicellular eukaryotic organisms and also gain insight into the evolution of this ancient class of RNA molecules. The main findings of this study include: i) the identification and characterization of more than 100 novel snoRNA species using both experimental and computational approaches; ii) characterization of the predominant snoRNA gene localization and expression strategies in *Euglena*; iii) a proposed mechanism of snoRNA evolution that also provides insight into the rRNA modification pattern observed in *E. gracilis*; iv) construction and initial characterization of a small *E. gracilis* RNA library containing a significantly reduced number of rRNA sequence reads by successfully inhibiting their amplification during the library synthesis procedure; and v) optimization of the conditions used to introduce exogenous nucleic acids into *Euglena gracilis* cells to potentially be used as a mechanism of ncRNA silencing.

Using a combined experimental and bioinformatic approach, a large collection of novel snoRNAs were identified in this study, more than doubling the number of known snoRNAs in *Euglena gracilis*. Features of the newly characterized snoRNAs are consistent with those previously identified. In particular, the number of known box AGA RNAs has been increased significantly, allowing for a more thorough examination of the features of *Euglena* ψ-guide RNAs. The snoRNAs that guide approximately 40% of the mapped rRNA pseudouridines have been identified and they all adopt a single-stem secondary structure and contain an AGA box element, unique features amongst eukaryotes and possibly representative of a more primitive form of this class of RNAs.

This study also more convincingly verifies that these single-stem RNAs are the predominant stable functional form of *Euglena* ψ-guide RNAs. Previous studies (Russell et al., 2004) used immunoprecipitation to isolate this class of RNAs and unwanted degradation (cleavage) of functional RNAs during the procedure was a possibility. The RNA library preparation procedure that I developed is faster and minimizes degradation issues.

This study also identified 78 novel box C/D snoRNAs, the majority of which have predicted rRNA methylation target sites. In addition, a U14 candidate was identified and characterized. It is unique from what is observed in most eukaryotes in that it lacks a modification-guide domain, but the essential rRNA processing domain is conserved. Based on its distribution and functional divergence in eukaryotes, a model for U14 snoRNA evolution is now proposed that suggests the ancestral function of this RNA included both specifying RNA cleavage and guiding 2′-O-methylation and that sporadic loss of modification-domain functionality has occurred in some lineages, including *Euglena*.

The genes encoding these snoRNAs are predominately arranged in clusters, often containing both classes of snoRNAs. Many snoRNA genes are transcribed by RNA pol II and are expressed as polycistronic transcripts which must undergo subsequent processing by a yet uncharacterized pathway to produce mature RNAs. It appears that numerous snoRNAs in *E. gracilis* contain hypermethylated cap structures and are likely processed via a unique mechanism, as conserved RNA features of polycistronic processing are not obvious. Studies to determine how snoRNAs are processed in *Euglena,* including mapping the ends of precursor molecules to potentially identify important sequence

motifs or through the development of an *in vivo* expression system, may reveal an interesting (and potentially novel) ncRNA processing pathway.

In *E. gracilis*, snoRNA gene clusters can be highly repetitive, forming large arrays of tandemly repeating units. Based on current knowledge, the snoRNA gene clusters identified in this study are some of the largest ever characterized. Examination of the chromosomal localization of these arrays using a technique such as *in situ* hybridization, will provide more genomic context about snoRNA genes in *Euglena gracilis*. The functional importance of having such large gene arrays is currently unknown; one explanation is multiple gene copies would ensure high levels of snoRNA expression, which ultimately results in the efficient modification of every targeted rRNA site. This may be of particular importance in *E. gracilis*, as a means of stabilizing the highly fragmented LSU rRNA during ribosome assembly.

The extensive characterization of snoRNA genes in *E. gracilis* revealed an interesting mechanism of generating novel snoRNAs and modification sites through snoRNA gene and cluster duplication events, which appear to be common genome evolution processes in this organism. Gene duplication followed by sequence divergence has generated numerous novel snoRNA species and can help explain the densely modified rRNA sites, many of which are unique to *Euglena*. Furthermore, the relative 'ease' at which gene duplications occur has led to the accumulation of numerous isoforms for the majority of the snoRNA species identified, exemplifying the extraordinary abundance of these molecules in *E. gracilis*.

In order to prepare a small RNA library for high-throughput sequencing, it was important to reduce the amount of rRNA fragments in the sample to prevent their overwhelming abundance in the sequenced library. The technique that proved most

successful was the use of blocker oligonucleotides specific to each fragment at the amplification step of library synthesis. This reduced the number of rRNA reads in the library to approximately 3%, a substantial decrease from not using any rRNA depletion strategy. This approach could be utilized to prevent the amplification of any unwanted RNA species during library preparation for any organism, as long as sequence information is available in order to design specific blocker oligonucleotides. This is especially useful in organisms for which no commercial rRNA depletion kits are available.

In addition to identifying novel modification-guide snoRNAs in the small RNA library, numerous orphans were identified that do not target any known rRNA site for modification and whose function is unknown. ncRNAs (including snoRNAs) are involved in a number of diverse cellular functions and unique roles are continually being identified. Therefore, it is possible the RNAs identified in this study play unique roles in *E. gracilis* cells. A potential approach to study their function *in vivo* is to utilize an RNAi-based method to specifically silence these RNAs and observe the resulting cellular effect. In order to use such an approach to study RNA function, the conditions required to introduce exogenous dsRNA into *Euglena* cells first needed to be optimized. Using these conditions, specific snoRNAs were then targeted for silencing and the efficacy of the treatment was assessed. While I appeared to have initial success at silencing snoRNAs in *Euglena* cells, unfortunately the results could not be replicated, and all subsequent attempts to silence these and other ncRNAs were also unsuccessful. Other nucleic acid molecules, such as LNAs, may prove a more viable option for nuclear RNA silencing and should be explored further, as this approach would be invaluable in determining the roles of newly identified ncRNAs in *Euglena gracilis*.

References

Adl, S.M., Simpson, A.G., Lane, C.E., Lukes, J., Bass, D., Bowser, S.S., Brown, M.W., Burki, F., Dunthorn, M., Hampl, V., *et al.* (2012). The revised classification of eukaryotes. J Eukaryot Microbiol *59*, 429-493.

Ahmadinejad, N., Dagan, T., and Martin, W. (2007). Genome history in the symbiotic hybrid *Euglena gracilis*. Gene *402*, 35-39.

Antal, M., Mougin, A., Kis, M., Boros, E., Steger, G., Jakab, G., Solymosy, F., and Branlant, C. (2000). Molecular characterization at the RNA and gene levels of U3 snoRNA from a unicellular green alga, *Chlamydomonas reinhardtii*. Nucleic Acids Res *28*, 2959-2968.

Arabidopsis Genome, I. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature *408*, 796-815.

Aspegren, A., Hinas, A., Larsson, P., Larsson, A., and Soderbom, F. (2004). Novel non-coding RNAs in *Dictyostelium discoideum* and their expression during development. Nucleic Acids Res *32*, 4646-4656.

Atzorn, V., Fragapane, P., and Kiss, T. (2004). U17/snR30 is a ubiquitous snoRNA with two conserved sequence motifs essential for 18S rRNA production. Mol Cell Biol *24*, 1769-1778.

Bachellerie, J.P., Cavaille, J., and Huttenhofer, A. (2002). The expanding snoRNA world. Biochimie *84*, 775-790.

Bachellerie, J.P., Michot, B., Nicoloso, M., Balakin, A., Ni, J., and Fournier, M.J. (1995). Antisense snoRNAs: a family of nucleolar RNAs with long complementarities to rRNA. Trends Biochem Sci *20*, 261-264.

Balakin, A.G., Smith, L., and Fournier, M.J. (1996). The RNA world of the nucleolus: two major families of small RNAs defined by different box elements with related functions. Cell *86*, 823-834.

Ban, N., Nissen, P., Hansen, J., Moore, P.B., and Steitz, T.A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science *289*, 905-920.

Barbezier, N., Canino, G., Rodor, J., Jobet, E., Saez-Vasquez, J., Marchfelder, A., and Echeverria, M. (2009). Processing of a dicistronic tRNA-snoRNA precursor: combined analysis in vitro and in vivo reveals alternate pathways and coupling to assembly of snoRNP. Plant Physiol *150*, 1598-1610.

Barneche, F., Gaspin, C., Guyot, R., and Echeverria, 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.

Baudin-Baillieu, A., Fabret, C., Liang, X.H., Piekna-Przybylska, D., Fournier, M.J., and Rousset, J.P. (2009). Nucleotide modifications in three functionally important regions of the *Saccharomyces cerevisiae* ribosome affect translation accuracy. Nucleic Acids Res *37*, 7665-7677.

Becker, H.F., Motorin, Y., Planta, R.J., and Grosjean, H. (1997). The yeast gene YNL292w encodes a pseudouridine synthase (Pus4) catalyzing the formation of psi55 in both mitochondrial and cytoplasmic tRNAs. Nucleic Acids Res *25*, 4493-4499.

Bennetzen, J.L. (2002). Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica *115*, 29-36.

Blackburn, E.H., and Collins, K. (2011). Telomerase: an RNP enzyme synthesizes DNA. Cold Spring Harb Perspect Biol *3*.

Bleichert, F., Gagnon, K.T., Brown, B.A., 2nd, Maxwell, E.S., Leschziner, A.E., Unger, V.M., and Baserga, S.J. (2009). A dimeric structure for archaeal box C/D small ribonucleoproteins. Science *325*, 1384-1387.

Borst, P., and Sabatini, R. (2008). Base J: discovery, biosynthesis, and possible functions. Annu Rev Microbiol *62*, 235-251.

Bortolin, M.L., Ganot, P., and Kiss, T. (1999). Elements essential for accumulation and function of small nucleolar RNAs directing site-specific pseudouridylation of ribosomal RNAs. EMBO J *18*, 457-469.

Brameier, M., Herwig, A., Reinhardt, R., Walter, L., and Gruber, J. (2011). Human box C/D snoRNAs with miRNA like functions: expanding the range of regulatory RNAs. Nucleic Acids Res *39*, 675-686.

Brandis, K.A., Gale, S., Jinn, S., Langmade, S.J., Dudley-Rucker, N., Jiang, H., Sidhu, R., Ren, A., Goldberg, A., Schaffer, J.E., *et al.* (2013). Box C/D small nucleolar RNA (snoRNA) U60 regulates intracellular cholesterol trafficking. J Biol Chem *288*, 35703-35713.

Brown, J.W., Clark, G.P., Leader, D.J., Simpson, C.G., and Lowe, T. (2001). Multiple snoRNA gene clusters from *Arabidopsis*. RNA *7*, 1817-1832.

Brown, J.W., Echeverria, M., and Qu, L.H. (2003). Plant snoRNAs: functional evolution and new modes of gene expression. Trends Plant Sci *8*, 42-49.

Brown, J.W., Marshall, D.F., and Echeverria, M. (2008). Intronic noncoding RNAs and splicing. Trends Plant Sci *13*, 335-342.

Carlile, T.M., Rojas-Duran, M.F., Zinshteyn, B., Shin, H., Bartoli, K.M., and Gilbert, W.V. (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature *515*, 143-146.

Cavaille, J., Hadjiolov, A.A., and Bachellerie, J.P. (1996). Processing of mammalian rRNA precursors at the 3' end of 18S rRNA. Identification of cis-acting signals suggests the involvement of U13 small nucleolar RNA. Eur J Biochem *242*, 206-213.

Cavalier-Smith, T. (2013). Early evolution of eukaryote feeding modes, cell structural diversity, and classification of the protozoan phyla Loukozoa, Sulcozoa, and Choanozoa. Eur J Protistol *49*, 115-178.

Chae, S.R., Hwang, E.J., and Shin, H.S. (2006). Single cell protein production of *Euglena gracilis* and carbon dioxide fixation in an innovative photo-bioreactor. Bioresource Technology *97*, 322-329.

Chakrabarti, K., Pearson, M., Grate, L., Sterne-Weiler, T., Deans, J., Donohue, J.P., and Ares, M., Jr. (2007). Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis. RNA *13*, 1923-1939.

Chanfreau, G., Legrain, P., and Jacquier, A. (1998). Yeast RNase III as a key processing enzyme in small nucleolar RNAs metabolism. J Mol Biol *284*, 975-988.

Charette, J.M., and Gray, M.W. (2009). U3 snoRNA genes are multi-copy and frequently linked to U5 snRNA genes in *Euglena gracilis*. BMC Genomics *10*, 528.

Chekanova, J.A., Gregory, B.D., Reverdatto, S.V., Chen, H., Kumar, R., Hooker, T., Yazaki, J., Li, P., Skiba, N., Peng, Q., *et al.* (2007). Genome-wide high-resolution mapping of exosome substrates reveals hidden features in the *Arabidopsis* transcriptome. Cell *131*, 1340-1353.

Chen, C.L., Chen, C.J., Vallon, O., Huang, Z.P., Zhou, H., and Qu, L.H. (2008). Genomewide analysis of box C/D and box H/ACA snoRNAs in *Chlamydomonas reinhardtii* reveals an extensive organization into intronic gene clusters. Genetics *179*, 21-30.

Chen, C.L., Liang, D., Zhou, H., Zhuo, M., Chen, Y.Q., and Qu, L.H. (2003). The high diversity of snoRNAs in plants: identification and comparative study of 120 snoRNA genes from *Oryza sativa*. Nucleic Acids Res *31*, 2601-2613.

Chen, J.L., Blasco, M.A., and Greider, C.W. (2000). Secondary structure of vertebrate telomerase RNA. Cell *100*, 503-514.

Chen, X.S., Penny, D., and Collins, L.J. (2011). Characterization of RNase MRP RNA and novel snoRNAs from *Giardia intestinalis* and *Trichomonas vaginalis*. BMC Genomics *12*, 550.

Chen, Z., and Duan, X. (2011). Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. Methods Mol Biol *733*, 93-103.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., *et al.* (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science *308*, 1149-1154.

Clark, M.B., Amaral, P.P., Schlesinger, F.J., Dinger, M.E., Taft, R.J., Rinn, J.L., Pontig, C.P., Stadler, P.F., Morris, K.V., Morillon, A., *et al.* (2011). The reality of pervasive transcription. PLoS Biol *9*, e1000625.

Comella, P., Pontvianne, F., Lahmy, S., Vignols, F., Barbezier, N., Debures, A., Jobet, E., Brugidou, E., Echeverria, M., and Saez-Vasquez, J. (2008). Characterization of a ribonuclease III-like protein required for cleavage of the pre-rRNA in the 3'ETS in *Arabidopsis*. Nucleic Acids Res *36*, 1163-1175.

Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet *9*, 938-950.

Costa, F.F. (2010). Non-coding RNAs: Meet thy masters. Bioessays *32*, 599-608.

Cougot, N., van Dijk, E., Babajko, S., and Seraphin, B. (2004). 'Cap-tabolism'. Trends Biochem Sci *29*, 436-444.

Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561-563.

Crick, F.H. (1958). On protein synthesis. Symp Soc Exp Biol *12*, 138-163.

Daiker, V., Lebert, M., Richter, P., and Hader, D.P. (2010). Molecular characterization of a calmodulin involved in the signal transduction chain of gravitaxis in *Euglena gracilis*. Planta *231*, 1229-1236.

Darzacq, X., Jady, B.E., Verheggen, C., Kiss, A.M., Bertrand, E., and Kiss, T. (2002). Cajal body-specific small nuclear RNAs: a novel class of 2'-O-methylation and pseudouridylation guide RNAs. EMBO J *21*, 2746-2756.

Decatur, W.A., and Fournier, M.J. (2002). rRNA modifications and ribosome function. Trends Biochem Sci *27*, 344-351.

Deng, W., Zhu, X., Skogerbo, G., Zhao, Y., Fu, Z., Wang, Y., He, H., Cai, L., Sun, H., Liu, C., *et al.* (2006). Organization of the *Caenorhabditis elegans* small non-coding transcriptome: genomic features, biogenesis, and expression. Genome Res *16*, 20-29.

Dennis, P.P., Omer, A., and Lowe, T. (2001). A guided tour: small RNA function in Archaea. Mol Microbiol *40*, 509-519.

Dez, C., Henras, A., Faucon, B., Lafontaine, D., Caizergues-Ferrer, M., and Henry, Y. (2001). Stable expression in yeast of the mature form of human telomerase RNA depends on its association with the box H/ACA small nucleolar RNP proteins Cbf5p, Nhp2p and Nop10p. Nucleic Acids Res *29*, 598-603.

Dieci, G., Preti, M., and Montanini, B. (2009). Eukaryotic snoRNAs: a paradigm for gene expression flexibility. Genomics *94*, 83-88.

Dong, Z.W., Shao, P., Diao, L.T., Zhou, H., Yu, C.H., and Qu, L.H. (2012). RTL-P: a sensitive approach for detecting sites of 2'-O-methylation in RNA molecules. Nucleic Acids Res *40*, e157.

Dooijes, D., Chaves, I., Kieft, R., Dirks-Mulder, A., Martin, W., and Borst, P. (2000). Base J originally found in kinetoplastida is also a minor constituent of nuclear DNA of *Euglena gracilis*. Nucleic Acids Res *28*, 3017-3021.

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

Dragon, F., Pogacic, V., and Filipowicz, W. (2000). In vitro assembly of human H/ACA small nucleolar RNPs reveals unique features of U17 and telomerase RNAs. Mol Cell Biol *20*, 3037-3048.

Dsouza, M., Larsen, N., and Overbeek, R. (1997). Searching for patterns in genomic data. Trends Genet *13*, 497-498.

Dutca, L.M., Gallagher, J.E., and Baserga, S.J. (2011). The initial U3 snoRNA:pre-rRNA base pairing interaction required for pre-18S rRNA folding revealed by in vivo chemical probing. Nucleic Acids Res *39*, 5164-5180.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460-2461.

Ender, C., Krek, A., Friedlander, M.R., Beitzinger, M., Weinmann, L., Chen, W., Pfeffer, S., Rajewsky, N., and Meister, G. (2008). A human snoRNA with microRNA-like functions. Mol Cell *32*, 519-528.

Esakova, O., and Krasilnikov, A.S. (2010). Of proteins and RNA: the RNase P/MRP family. RNA *16*, 1725-1747.

Fantoni, A., Dare, A.O., and 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.

Faust, T., Frankel, A., and D'Orso, I. (2012). Transcription control by long non-coding RNAs. Transcription *3*, 78-86.

Fayet-Lebaron, E., Atzorn, V., Henry, Y., and Kiss, T. (2009). 18S rRNA processing requires base pairings of snR30 H/ACA snoRNA to eukaryote-specific 18S sequences. EMBO J *28*, 1260-1270.

Feng, J., Bi, C., Clark, B.S., Mady, R., Shah, P., and Kohtz, J.D. (2006). The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. Genes Dev *20*, 1470-1484.

Fernandez, I.S., Ng, C.L., Kelley, A.C., Wu, G., Yu, Y.T., and Ramakrishnan, V. (2013). Unusual base pairing during the decoding of a stop codon by the ribosome. Nature *500*, 107-110.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature *391*, 806-811.

Frith, M.C., Pheasant, M., and Mattick, J.S. (2005). The amazing complexity of the human transcriptome. Eur J Hum Genet *13*, 894-897.

Gaillard, C., and Strauss, F. (1990). Ethanol precipitation of DNA with linear polyacrylamide as carrier. Nucleic Acids Res *18*, 378.

Ganot, P., Bortolin, M.L., and Kiss, T. (1997a). Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. Cell *89*, 799-809.

Ganot, P., Caizergues-Ferrer, M., and Kiss, T. (1997b). The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. Genes Dev *11*, 941-956.

Gaspin, C., Cavaille, J., Erauso, G., and Bachellerie, J.P. (2000). Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the *Pyrococcus* genomes. J Mol Biol *297*, 895-906.

Ge, J., and Yu, Y.T. (2013). RNA pseudouridylation: new insights into an old modification. Trends Biochem Sci *38*, 210-218.

Ghazal, G., Ge, D., Gervais-Bird, J., Gagnon, J., and Abou Elela, S. (2005). Genome-wide prediction and analysis of yeast RNase III-dependent snoRNA processing signals. Mol Cell Biol *25*, 2981-2994.

Gibbs, S.P. (1978). Chloroplasts of *Euglena* may have evolved from symbiotic green-algae. Can J Bot-Rev Can Bot *56*, 2883-2889.

Gilbert, W. (1986). Origin of life - the RNA world. Nature *319*, 618-618.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., *et al.* (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science *296*, 92-100.

Goodrich, J.A., and Kugel, J.F. (2009). 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 *44*, 3-15.

Gottesman, S., and Storz, G. (2011). Bacterial small RNA regulators: versatile roles and rapidly evolving variations. Cold Spring Harb Perspect Biol *3*.

Green, R., and Noller, H.F. (1996). In vitro complementation analysis localizes 23S rRNA posttranscriptional modifications that are required for *Escherichia coli* 50S ribosomal subunit assembly and function. RNA *2*, 1011-1021.

Greenwood, S.J., Schnare, M.N., Cook, J.R., and Gray, M.W. (2001). Analysis of intergenic spacer transcripts suggests 'read-around' transcription of the extrachromosomal circular rDNA in *Euglena gracilis*. Nucleic Acids Res *29*, 2191-2198.

Grzechnik, P., and Kufel, J. (2008). Polyadenylation linked to transcription termination directs the processing of snoRNA precursors in yeast. Mol Cell *32*, 247-258.

Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell *35*, 849-857.

Guffanti, E., Ferrari, R., Preti, M., Forloni, M., Harismendy, O., Lefebvre, O., and Dieci, G. (2006). A minimal promoter for TFIIIC-dependent in vitro transcription of snoRNA and tRNA genes by RNA polymerase III. J Biol Chem *281*, 23945-23957.

Gupta, S.K., Hury, A., Ziporen, Y., Shi, H., Ullu, E., and Michaeli, S. (2010). Small nucleolar RNA interference in *Trypanosoma brucei*: mechanism and utilization for elucidating the function of snoRNAs. Nucleic Acids Res *38*, 7236-7247.

Hamma, T., and Ferre-D'Amare, A.R. (2010). The box H/ACA ribonucleoprotein complex: interplay of RNA and protein structures in post-transcriptional RNA modification. J Biol Chem *285*, 805-809.

Hannon, G.J. (2002). RNA interference. Nature *418*, 244-251.

Hayashi, H., Narumi, I., Wada, S., Kikuchi, M., Furuta, M., Uehara, K., and Watanabe, H. (2004) Light dependency of resistance to ionizing radiation in *Euglena gracilis*. Journal of Plant Physiology *161*, 1101-1106.

He, S., Wurtzel, O., Singh, K., Froula, J.L., Yilmaz, S., Tringe, S.G., Wang, Z., Chen, F., Lindquist, E.A., Sorek, R., *et al.* (2010). Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. Nat Methods *7*, 807-812.

Helm, M. (2006). Post-transcriptional nucleotide modification and alternative folding of RNA. Nucleic Acids Res *34*, 721-733.

Henras, A.K., Soudet, J., Gerus, M., Lebaron, S., Caizergues-Ferrer, M., Mougin, A., and Henry, Y. (2008). The post-transcriptional steps of eukaryotic ribosome biogenesis. Cell Mol Life Sci *65*, 2334-2359.

Hoeppner, M.P., and Poole, A.M. (2012). Comparative genomics of eukaryotic small nucleolar RNAs reveals deep evolutionary ancestry amidst ongoing intragenomic mobility. BMC Evol Biol *12*, 183.

Huang, G.M., Jarmolowski, A., Struck, J.C., and Fournier, M.J. (1992). Accumulation of U14 small nuclear RNA in *Saccharomyces cerevisiae* requires box C, box D, and a 5', 3' terminal stem. Mol Cell Biol *12*, 4456-4463.

Huang, Z.P., Zhou, H., Liang, D., and Qu, L.H. (2004). Different expression strategy: multiple intronic gene clusters of box H/ACA snoRNA in *Drosophila melanogaster*. J Mol Biol *341*, 669-683.

Hudson, A.J., Moore, A.N., Elniski, D., Joseph, J., Yee, J., and Russell, A.G. (2012). Evolutionarily divergent spliceosomal snRNAs and a conserved non-coding RNA processing motif in *Giardia lamblia*. Nucleic Acids Res *40*, 10995-11008.

Hughes, J.M. (1996). Functional base-pairing interaction between highly conserved elements of U3 small nucleolar RNA and the small ribosomal subunit RNA. J Mol Biol *259*, 645-654.

Huntzinger, E., and Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat Rev Genet *12*, 99-110.

Huttenhofer, A., Brosius, J., and Bachellerie, J.P. (2002). RNomics: identification and function of small, non-messenger RNAs. Curr Opin Chem Biol *6*, 835-843.

Huttenhofer, A., Schattner, P., and Polacek, N. (2005). Non-coding RNAs: hope or hype? Trends Genet *21*, 289-297.

Ideue, T., Hino, K., Kitao, S., Yokoi, T., and Hirose, T. (2009). Efficient oligonucleotide-mediated degradation of nuclear noncoding RNAs in mammalian cultured cells. RNA *15*, 1578-1587.

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931-945.

Iseki, M., Matsunaga, S., Murakami, A., Ohno, K., Shiga, K., Yoshida, K., Sugai, M., Takahashi, T., Hori, T., and Watanabe, M. (2002). A blue-light-activated adenylyl cyclase mediates photoavoidance in *Euglena gracilis*. Nature *415*, 1047-1051.

Isogai, Y., Takada, S., Tjian, R., and Keles, S. (2007). Novel TRF1/BRF target genes revealed by genome-wide analysis of *Drosophila* Pol III transcription. EMBO J *26*, 79-89.

Jady, B.E., and Kiss, T. (2001). A small nucleolar guide RNA functions both in 2'-O-ribose methylation and pseudouridylation of the U5 spliceosomal RNA. EMBO J *20*, 541-551.

Jarmolowski, A., Zagorski, J., Li, H.V., and Fournier, M.J. (1990). Identification of essential elements in U14 RNA of *Saccharomyces cerevisiae*. EMBO J *9*, 4503-4509.

Jawdekar, G.W., and Henry, R.W. (2008). Transcriptional regulation of human small nuclear RNA genes. Biochim Biophys Acta *1779*, 295-305.

Jeffares, D.C., Poole, A.M., and Penny, D. (1995). Pre-rRNA processing and the path from the RNA world. Trends Biochem Sci *20*, 298-299.

Jiang, G., Chen, X., Li, W., Jin, Y., and Wang, D. (2002). Identification and characterization of a novel U14 small nucleolar RNA gene cluster in *Oryza sativa*. Gene *294*, 187-196.

Jongeneel, C.V., Delorenzi, M., Iseli, C., Zhou, D., Haudenschild, C.D., Khrebtukova, I., Kuznetsov, D., Stevenson, B.J., Strausberg, R.L., Simpson, A.J., *et al.* (2005). An atlas of human gene expression from massively parallel signature sequencing (MPSS). Genome Res *15*, 1007-1014.

Jurica, M.S., and Moore, M.J. (2003). Pre-mRNA splicing: awash in a sea of proteins. Mol Cell *12*, 5-14.

Karijolich, J., and Yu, Y.T. (2011). Converting nonsense codons into sense codons by targeted pseudouridylation. Nature *474*, 395-398.

Kehr, S., Bartschat, S., Tafer, H., Stadler, P.F., and Hertel, J. (2014). Matching of Soulmates: coevolution of snoRNAs and their targets. Mol Biol Evol *31*, 455-467.

Kenmochi, N., Higa, S., Yoshihama, M., and Tanaka, T. (1996). U14 snoRNAs are encoded in introns of human ribosomal protein S13 gene. Biochem Biophys Res Commun *228*, 371-374.

King, T.H., Liu, B., McCully, R.R., and Fournier, M.J. (2003). Ribosome structure and activity are altered in cells lacking snoRNPs that form pseudouridines in the peptidyl transferase center. Mol Cell *11*, 425-435.

Kishore, S., Khanna, A., Zhang, Z., Hui, J., Balwierz, P.J., Stefan, M., Beach, C., Nicholls, R.D., Zavolan, M., and Stamm, S. (2010). The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing. Hum Mol Genet *19*, 1153-1164.

Kishore, S., and Stamm, S. (2006). The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. Science *311*, 230-232.

Kiss, T. (2002). Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. Cell *109*, 145-148.

Kiss, T., Fayet-Lebaron, E., and Jady, B.E. (2010). Box H/ACA small ribonucleoproteins. Mol Cell *37*, 597-606.

Kiss, T., Marshallsay, C., and Filipowicz, W. (1991). Alteration of the RNA polymerase specificity of U3 snRNA genes during evolution and in vitro. Cell *65*, 517-526.

Kiss-Laszlo, Z., Henry, Y., Bachellerie, J.P., Caizergues-Ferrer, M., and Kiss, T. (1996). Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. Cell *85*, 1077-1088.

Kiss-Laszlo, Z., Henry, Y., and Kiss, T. (1998). Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. EMBO J *17*, 797-807.

Koonin, E.V. (1996). Pseudouridine synthases: four families of enzymes containing a putative uridine-binding motif also conserved in dUTPases and dCTP deaminases. Nucleic Acids Res *24*, 2411-2415.

Kos, M., and Tollervey, D. (2010). Yeast pre-rRNA processing and modification occur cotranscriptionally. Mol Cell *37*, 809-820.

Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., Gottschling, D.E., and Cech, T.R. (1982). Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. Cell *31*, 147-157.

Kruszka, K., Barneche, F., Guyot, R., Ailhas, J., Meneau, I., Schiffer, S., Marchfelder, A., and Echeverria, M. (2003). Plant dicistronic tRNA-snoRNA genes: a new mode of expression of the small nucleolar RNAs processed by RNase Z. EMBO J *22*, 621-632.

Kuhn, J.F., Tran, E.J., and Maxwell, E.S. (2002). Archaeal ribosomal protein L7 is a functional homolog of the eukaryotic 15.5kD/Snu13p snoRNP core protein. Nucleic Acids Res *30*, 931-941.

Lafontaine, D.L., and Tollervey, D. (1998). Birth of the snoRNPs: the evolution of the modification-guide snoRNAs. Trends Biochem Sci *23*, 383-388.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Leader, D.J., Clark, G.P., Watters, J., Beven, A.F., Shaw, P.J., and Brown, J.W. (1997). Clusters of multiple different small nucleolar RNA genes in plants are expressed as and processed from polycistronic pre-snoRNAs. EMBO J *16*, 5742-5751.

Leader, D.J., Clark, G.P., Watters, J., Beven, A.F., Shaw, P.J., and Brown, J.W. (1999). Splicing-independent processing of plant box C/D and box H/ACA small nucleolar RNAs. Plant Mol Biol *39*, 1091-1100.

Lecointe, F., Simos, G., Sauer, A., Hurt, E.C., Motorin, Y., and Grosjean, H. (1998). Characterization of yeast protein Deg1 as pseudouridine synthase (Pus3) catalyzing the formation of psi 38 and psi 39 in tRNA anticodon loop. J Biol Chem *273*, 1316-1323.

Lemay, J.F., D'Amours, A., Lemieux, C., Lackner, D.H., St-Sauveur, V.G., Bahler, J., and Bachand, F. (2010). The nuclear poly(A)-binding protein interacts with the exosome to promote synthesis of noncoding small nucleolar RNAs. Mol Cell *37*, 34-45.

Li, H.D., Zagorski, J., and Fournier, M.J. (1990). Depletion of U14 small nuclear RNA (snR128) disrupts production of 18S rRNA in *Saccharomyces cerevisiae*. Mol Cell Biol *10*, 1145-1152.

Li, L., and Ye, K. (2006). Crystal structure of an H/ACA box ribonucleoprotein particle. Nature *443*, 302-307.

Liang, D., Zhou, H., Zhang, P., Chen, Y.Q., Chen, X., Chen, C.L., and Qu, L.H. (2002a). A novel gene organization: intronic snoRNA gene clusters from *Oryza sativa*. Nucleic Acids Res *30*, 3262-3272.

Liang, W.Q., Clark, J.A., and Fournier, M.J. (1997). The rRNA-processing function of the yeast U14 small nucleolar RNA can be rescued by a conserved RNA helicase-like protein. Mol Cell Biol *17*, 4124-4132.

Liang, W.Q., and Fournier, M.J. (1995). U14 base-pairs with 18S rRNA: a novel snoRNA interaction required for rRNA processing. Genes Dev *9*, 2433-2443.

Liang, X.H., Liu, Q., and Fournier, M.J. (2007). rRNA modifications in an intersubunit bridge of the ribosome strongly affect both ribosome biogenesis and activity. Mol Cell *28*, 965-977.

Liang, X.H., Liu, Q., and Fournier, M.J. (2009). Loss of rRNA modifications in the decoding center of the ribosome impairs translation and strongly delays pre-rRNA processing. RNA *15*, 1716-1728.

Liang, X.H., Liu, Q., Liu, Q., King, T.H., and Fournier, M.J. (2010). Strong dependence between functional domains in a dual-function snoRNA infers coupling of rRNA processing and modification events. Nucleic Acids Res *38*, 3376-3387.

Liang, X.H., Liu, Q., and Michaeli, S. (2003). Small nucleolar RNA interference induced by antisense or double-stranded RNA in trypanosomatids. Proc Natl Acad Sci U S A *100*, 7521-7526.

Liang, X.H., Uliel, S., Hury, A., Barth, S., Doniger, T., Unger, R., and 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.

Liang, X.H., Vickers, T.A., Guo, S., and Crooke, S.T. (2011). Efficient and specific knockdown of small non-coding RNAs in mammalian cells and in mice. Nucleic Acids Res *39*, e13.

Liang, X.H., Xu, Y.X., and Michaeli, S. (2002b). The spliced leader-associated RNA is a trypanosome-specific sn(o) RNA that has the potential to guide pseudouridine formation on the SL RNA. RNA *8*, 237-246.

Lilley, D.M. (2012). The structure and folding of kink turns in RNA. Wiley Interdiscip Rev RNA *3*, 797-805.

Lin, J., Lai, S., Jia, R., Xu, A., Zhang, L., Lu, J., and Ye, K. (2011). Structural basis for site-specific ribose methylation by box C/D RNA protein complexes. Nature *469*, 559-563.

Liu, J.M., Livny, J., Lawrence, M.S., Kimball, M.D., Waldor, M.K., and Camilli, A. (2009a). Experimental discovery of sRNAs in *Vibrio cholerae* by direct cloning, 5S/tRNA depletion and parallel sequencing. Nucleic Acids Res *37*, e46.

Liu, N., Xiao, Z.D., Yu, C.H., Shao, P., Liang, Y.T., Guan, D.G., Yang, J.H., Chen, C.L., Qu, L.H., and Zhou, H. (2009b). SnoRNAs from the filamentous fungus *Neurospora crassa*: structural, functional and evolutionary insights. BMC Genomics *10*, 515.

Lovejoy, A.F., Riordan, D.P., and Brown, P.O. (2014). Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in *S. cerevisiae*. PLoS One *9*, e110799.

Lowe, T.M., and Eddy, S.R. (1999). A computational screen for methylation guide snoRNAs in yeast. Science *283*, 1168-1171.

Lukowiak, A.A., Narayanan, A., Li, Z.H., Terns, R.M., and Terns, M.P. (2001). The snoRNA domain of vertebrate telomerase RNA functions to localize the RNA within the nucleus. RNA *7*, 1833-1844.

Luo, Y., and Li, S. (2007). Genome-wide analyses of retrogenes derived from the human box H/ACA snoRNAs. Nucleic Acids Res *35*, 559-571.

Lustig, Y., Wachtel, C., Safro, M., Liu, L., and Michaeli, S. (2010). 'RNA walk' a novel approach to study RNA-RNA interactions between a small RNA and its target. Nucleic Acids Res *38*, e5.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science *290*, 1151-1155.

Machyna, M., Heyn, P., and Neugebauer, K.M. (2013). Cajal bodies: where form meets function. Wiley Interdiscip Rev RNA *4*, 17-34.

Makarova, J.A., and Kramerov, D.A. (2009). Analysis of C/D box snoRNA genes in vertebrates: The number of copies decreases in placental mammals. Genomics *94*, 11-19.

Mannoor, K., Liao, J., and Jiang, F. (2012). Small nucleolar RNAs in cancer. Biochim Biophys Acta *1826*, 121-128.

Marchfelder, A., Fischer, S., Brendel, J., Stoll, B., Maier, L.K., Jager, D., Prasse, D., Plagens, A., Schmitz, R.A., and Randau, L. (2012). Small RNAs for defence and regulation in archaea. Extremophiles *16*, 685-696.

Maruyama, S., Suzaki, T., Weber, A.P., Archibald, J.M., and Nozaki, H. (2011). Eukaryote-to-eukaryote gene transfer gives rise to genome mosaicism in euglenids. BMC Evol Biol *11*, 105.

Matera, A.G., Terns, R.M., and Terns, M.P. (2007). Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nat Rev Mol Cell Biol *8*, 209-220.

Mattick, J.S. (2004). RNA regulation: a new genetics? Nat Rev Genet *5*, 316-323.

Mattick, J.S., and Makunin, I.V. (2006). Non-coding RNA. Hum Mol Genet *15 Spec No 1*, R17-29.

Maxwell, E.S., and Fournier, M.J. (1995). The small nucleolar RNAs. Annu Rev Biochem *64*, 897-934.

Meister, G. (2008). Molecular biology. RNA interference in the nucleus. Science *321*, 496-497.

Meister, G., and Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA. Nature *431*, 343-349.

Mercer, T.R., Dinger, M.E., and Mattick, J.S. (2009). Long non-coding RNAs: insights into functions. Nat Rev Genet *10*, 155-159.

Michel, C.I., Holley, C.L., Scruggs, B.S., Sidhu, R., Brookheart, R.T., Listenberger, L.L., Behlke, M.A., Ory, D.S., and Schaffer, J.E. (2011). Small nucleolar RNAs U32a, U33, and U35a are critical mediators of metabolic stress. Cell Metab *14*, 33-44.

Mishra, P.C., Kumar, A., and Sharma, A. (2009). Analysis of small nucleolar RNAs reveals unique genetic features in malaria parasites. BMC Genomics *10*, 68.

Mitchell, J.R., Cheng, J., and Collins, K. (1999). A box H/ACA small nucleolar RNA-like domain at the human telomerase RNA 3' end. Mol Cell Biol *19*, 567-576.

Mo, D., Raabe, C.A., Reinhardt, R., Brosius, J., and Rozhdestvensky, T.S. (2013). Alternative processing as evolutionary mechanism for the origin of novel nonprotein coding RNAs. Genome Biol Evol *5*, 2061-2071.

Moore, A.N., and Russell, A.G. (2012). Clustered organization, polycistronic transcription, and evolution of modification-guide snoRNA genes in *Euglena gracilis.* Mol Genet Genomics *287*, 55-66.

Morrison, H.G., McArthur, A.G., Gillin, F.D., Aley, S.B., Adam, R.D., Olsen, G.J., Best, A.A., Cande, W.Z., Chen, F., Cipriano, M.J., *et al.* (2007). Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. Science *317*, 1921-1926.

Morrissey, J.P., and Tollervey, D. (1993). Yeast snR30 is a small nucleolar RNA required for 18S rRNA synthesis. Mol Cell Biol *13*, 2469-2477.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods *5*, 621-628.

Myslyuk, I., Doniger, T., Horesh, Y., Hury, A., Hoffer, R., Ziporen, Y., Michaeli, S., and Unger, R. (2008). Psiscan: a computational approach to identify H/ACA-like and AGA-like non-coding RNA in trypanosomatid genomes. BMC Bioinformatics *9*, 471.

Ni, J., Tien, A.L., and Fournier, M.J. (1997). Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA. Cell *89*, 565-573.

Nicoloso, M., Qu, L.H., Michot, B., and Bachellerie, J.P. (1996). Intron-encoded, antisense small nucleolar RNAs: the characterization of nine novel species points to their direct role as guides for the 2'-O-ribose methylation of rRNAs. J Mol Biol *260*, 178-195.

Nissen, P., Hansen, J., Ban, N., Moore, P.B., and Steitz, T.A. (2000). The structural basis of ribosome activity in peptide bond synthesis. Science *289*, 920-930.

Nixon, J.E., Wang, A., Morrison, H.G., McArthur, A.G., Sogin, M.L., Loftus, B.J., and Samuelson, J. (2002). A spliceosomal intron in *Giardia lamblia*. Proc Natl Acad Sci U S A *99*, 3701-3705.

O'Brien, E.A., Koski, L.B., Zhang, Y., Yang, L., Wang, E., Gray, M.W., Burger, G., and Lang, B.F. (2007). TBestDB: a taxonomically broad database of expressed sequence tags (ESTs). Nucleic Acids Res *35*, D445-451.

O'Neil, D., Glowatz, H., and Schlumpberger, M. (2013). Ribosomal RNA depletion for efficient use of RNA-seq capacity. Curr Protoc Mol Biol *Chapter 4*, Unit 4 19.

Ogbonna, J.C., Ichige, E., and Tanaka, H. (2002). Interactions between photoautotrophic and heterotrophic metabolism in photoheterotrophic cultures of *Euglena gracilis*. Appl Microbiol Biotechnol *58*, 532-538.

Omer, A.D., Lowe, T.M., Russell, A.G., Ebhardt, H., Eddy, S.R., and Dennis, P.P. (2000). Homologs of small nucleolar RNAs in Archaea. Science *288*, 517-522.

Omer, A.D., Ziesche, S., Decatur, W.A., Fournier, M.J., and Dennis, P.P. (2003). RNA-modifying machines in archaea. Mol Microbiol *48*, 617-629.

Ono, M., Scott, M.S., Yamada, K., Avolio, F., Barton, G.J., and Lamond, A.I. (2011). Identification of human miRNA precursors that resemble box C/D snoRNAs. Nucleic Acids Res *39*, 3879-3891.

Ono, M., Yamada, K., Avolio, F., Scott, M.S., van Koningsbruggen, S., Barton, G.J., and Lamond, A.I. (2010). Analysis of human small nucleolar RNAs (snoRNA) and the development of snoRNA modulator of gene expression vectors. Mol Biol Cell *21*, 1569-1584.

Orum, H., Nielsen, H., and Engberg, J. (1992). Structural organization of the genes encoding the small nuclear RNAs U1 to U6 of *Tetrahymena thermophila* is very similar to that of plant small nuclear RNA genes. J Mol Biol *227*, 114-121.

Peano, C., Pietrelli, A., Consolandi, C., Rossi, E., Petiti, L., Tagliabue, L., De Bellis, G., and Landini, P. (2013). An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. Microb Inform Exp *3*, 1.

Peculis, B.A. (1997). The sequence of the 5' end of the U8 small nucleolar RNA is critical for 5.8S and 28S rRNA maturation. Mol Cell Biol *17*, 3702-3713.

Peculis, B.A., and Steitz, J.A. (1993). Disruption of U8 nucleolar snRNA inhibits 5.8S and 28S rRNA processing in the *Xenopus* oocyte. Cell *73*, 1233-1245.

Penny, D., Hoeppner, M.P., Poole, A.M., and Jeffares, D.C. (2009). An overview of the introns-first theory. J Mol Evol *69*, 527-540.

Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S., and Brockdorff, N. (1996). Requirement for Xist in X chromosome inactivation. Nature *379*, 131-137.

Pertea, M. (2012). The human transcriptome: an unfinished story. Genes *3*, 344-360.

Peterlin, B.M., Brogie, J.E., and Price, D.H. (2012). 7SK snRNA: a noncoding RNA that plays a major role in regulating eukaryotic transcription. Wiley Interdiscip Rev RNA *3*, 92-103.

Phipps, K.R., Charette, J., and Baserga, S.J. (2011). The small subunit processome in ribosome biogenesis-progress and prospects. Wiley Interdiscip Rev RNA *2*, 1-21.

Piekna-Przybylska, D., Przybylski, P., Baudin-Baillieu, A., Rousset, J.P., and Fournier, M.J. (2008). Ribosome performance is enhanced by a rich cluster of pseudouridines in the A-site finger region of the large subunit. J Biol Chem *283*, 26026-26036.

Ploner, A., Ploner, C., Lukasser, M., Niederegger, H., and Huttenhofer, A. (2009). Methodological obstacles in knocking down small noncoding RNAs. RNA *15*, 1797-1804.

Pogacic, V., Dragon, F., and Filipowicz, W. (2000). Human H/ACA small nucleolar RNPs and telomerase share evolutionarily conserved proteins NHP2 and NOP10. Mol Cell Biol *20*, 9028-9040.

Qu, L.H., Henras, A., Lu, Y.J., Zhou, H., Zhou, W.X., Zhu, Y.Q., Zhao, J., Henry, Y., Caizergues-Ferrer, M., and Bachellerie, J.P. (1999). Seven novel methylation guide small nucleolar RNAs are processed from a common polycistronic transcript by Rat1p and RNase III in yeast. Mol Cell Biol *19*, 1144-1158.

Qu, L.H., Henry, Y., Nicoloso, M., Michot, B., Azum, M.C., Renalier, M.H., Caizergues-Ferrer, M., and Bachellerie, J.P. (1995). U24, a novel intron-encoded small nucleolar RNA with two 12 nt long, phylogenetically conserved complementarities to 28S rRNA. Nucleic Acids Res *23*, 2669-2676.

Rana, T.M. (2007). Illuminating the silence: understanding the structure and function of small RNAs. Nat Rev Mol Cell Biol *8*, 23-36.

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

Rawson, J.R., Eckenrode, V.K., Boerma, C.L., and Curtis, S. (1979). DNA sequence organization in the alga *Euglena gracilis*. Biochim Biophys Acta *563*, 1-16.

Rederstorff, M., and Huttenhofer, A. (2011). cDNA library generation from ribonucleoprotein particles. Nat Protoc *6*, 166-174.

Reichow, S.L., Hamma, T., Ferre-D'Amare, A.R., and Varani, G. (2007). The structure and function of small nucleolar ribonucleoproteins. Nucleic Acids Res *35*, 1452-1464.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., *et al.* (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311-1323.

Robb, G.B., Brown, K.M., Khurana, J., and Rana, T.M. (2005). Specific and potent RNAi in the nucleus of human cells. Nat Struct Mol Biol *12*, 133-137.

Rosenblad, M.A., Larsen, N., Samuelsson, T., and Zwieb, C. (2009). Kinship in the SRP RNA family. RNA Biol *6*, 508-516.

Rother, S., and Meister, G. (2011). Small RNAs derived from longer non-coding RNAs. Biochimie *93*, 1905-1915.

Roy, S.W., Hudson, A.J., Joseph, J., Yee, J., and Russell, A.G. (2012). Numerous fragmented spliceosomal introns, AT-AC splicing, and an unusual dynein gene expression pathway in *Giardia lamblia*. Mol Biol Evol *29*, 43-49.

Rozhdestvensky, T.S., Tang, T.H., Tchirkova, I.V., Brosius, J., Bachellerie, J.P., and Huttenhofer, A. (2003). Binding of L7Ae protein to the K-turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea. Nucleic Acids Res *31*, 869-877.

Russell, A.G., Schnare, M.N., and Gray, M.W. (2004). Pseudouridine-guide RNAs and other Cbf5p-associated RNAs in *Euglena gracilis*. RNA *10*, 1034-1046.

Russell, A.G., Schnare, M.N., and Gray, M.W. (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*, 1548-1565.

Russell, A.G., Shutt, T.E., Watkins, R.F., and Gray, M.W. (2005). An ancient spliceosomal intron in the ribosomal protein L7a gene (Rpl7a) of *Giardia lamblia*. BMC Evol Biol *5*, 45.

Sambrook, J., and Russell, D.W. (2001). Molecular cloning: A laboratory manual.

Schattner, P., Decatur, W.A., Davis, C.A., Ares, M., Jr., Fournier, M.J., and Lowe, T.M. (2004). Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. Nucleic Acids Res *32*, 4281-4296.

Schmitz, J., Zemann, A., Churakov, G., Kuhl, H., Grutzner, F., Reinhardt, R., and Brosius, J. (2008). Retroposed SNOfall--a mammalian-wide comparison of platypus snoRNAs. Genome Res *18*, 1005-1010.

Schnare, M.N., Cook, J.R., and Gray, M.W. (1990). Fourteen internal transcribed spacers in the circular ribosomal DNA of *Euglena gracilis*. J Mol Biol *215*, 85-91.

Schnare, M.N., and Gray, M.W. (1990). Sixteen discrete RNA components in the cytoplasmic ribosome of *Euglena gracilis*. J Mol Biol *215*, 73-83.

Schnare, M.N., and Gray, M.W. (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.

Schwartz, S., Bernstein, D.A., Mumbach, M.R., Jovanovic, M., Herbst, R.H., Leon-Ricardo, B.X., Engreitz, J.M., Guttman, M., Satija, R., Lander, E.S., *et al.* (2014). Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. Cell *159*, 148-162.

Shao, P., Yang, J.H., Zhou, H., Guan, D.G., and Qu, L.H. (2009). Genome-wide analysis of chicken snoRNAs provides unique implications for the evolution of vertebrate snoRNAs. BMC Genomics *10*, 86.

Siegel, T.N., Hekstra, D.R., Wang, X., Dewell, S., and Cross, G.A. (2010). Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. Nucleic Acids Res *38*, 4946-4957.

Simoes-Barbosa, A., Chakrabarti, K., Pearson, M., Benarroch, D., Shuman, S., and Johnson, P.J. (2012). Box H/ACA snoRNAs are preferred substrates for the trimethylguanosine synthase in the divergent unicellular eukaryote *Trichomonas vaginalis*. RNA *18*, 1656-1665.

Simoes-Barbosa, A., Meloni, D., Wohlschlegel, J.A., Konarska, M.M., and Johnson, P.J. (2008). Spliceosomal snRNAs in the unicellular eukaryote *Trichomonas vaginalis* are structurally conserved but lack a 5'-cap structure. RNA *14*, 1617-1631.

Simpson, A.G., and Roger, A.J. (2004). Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa. Mol Phylogenet Evol *30*, 201-212.

Sloan, K.E., Bohnsack, M.T., Schneider, C., and Watkins, N.J. (2014). The roles of SSU processome components and surveillance factors in the initial processing of human ribosomal RNA. RNA *20*, 540-550.

Steinmetz, E.J., Warren, C.L., Kuehner, J.N., Panbehi, B., Ansari, A.Z., and Brow, D.A. (2006). Genome-wide distribution of yeast RNA polymerase II and its control by Sen1 helicase. Mol Cell *24*, 735-746.

Taft, R.J., Glazov, E.A., Lassmann, T., Hayashizaki, Y., Carninci, P., and Mattick, J.S. (2009). Small RNAs derived from snoRNAs. RNA *15*, 1233-1240.

Tang, T.H., Bachellerie, J.P., Rozhdestvensky, T., Bortolin, M.L., Huber, H., Drungowski, M., Elge, T., Brosius, J., and Huttenhofer, A. (2002). Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. Proc Natl Acad Sci U S A *99*, 7536-7541.

Tijsterman, M., and Plasterk, R.H. (2004). Dicers at RISC; the mechanism of RNAi. Cell *117*, 1-3.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., and Martin, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet *5*, 123-135.

Tollervey, D., Lehtonen, H., Jansen, R., Kern, H., and Hurt, E.C. (1993). Temperature-sensitive mutations demonstrate roles for yeast fibrillarin in pre-rRNA processing, pre-rRNA methylation, and ribosome assembly. Cell *72*, 443-457.

Tran, E., Zhang, X., Lackey, L., and Maxwell, E.S. (2005). Conserved spacing between the box C/D and C'/D' RNPs of the archaeal box C/D sRNP complex is required for efficient 2'-O-methylation of target RNAs. RNA *11*, 285-293.

Turmel, M., Gagnon, M.C., O'Kelly, C.J., Otis, C., and Lemieux, C. (2009). The chloroplast genomes of the green algae *Pyramimonas*, *Monomastix*, and *Pycnococcus* shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. Mol Biol Evol *26*, 631-648.

Turowski, T.W., and Tollervey, D. (2015). Cotranscriptional events in eukaryotic ribosome synthesis. Wiley Interdiscip Rev RNA *6*, 129-139.

Tycowski, K.T., Aab, A., and Steitz, J.A. (2004). Guide RNAs with 5' caps and novel box C/D snoRNA-like domains for modification of snRNAs in metazoa. Curr Biol *14*, 1985-1995.

Tycowski, K.T., Smith, C.M., Shu, M.D., and Steitz, J.A. (1996). A small nucleolar RNA requirement for site-specific ribose methylation of rRNA in *Xenopus*. Proc Natl Acad Sci U S A *93*, 14480-14485.

Tycowski, K.T., and Steitz, J.A. (2001). Non-coding snoRNA host genes in *Drosophila*: expression strategies for modification guide snoRNAs. Eur J Cell Biol *80*, 119-125.

Tycowski, K.T., You, Z.H., Graham, P.J., and Steitz, J.A. (1998). Modification of U6 spliceosomal RNA is guided by other small RNAs. Mol Cell *2*, 629-638.

Uliel, S., Liang, X.H., Unger, R., and Michaeli, S. (2004). Small nucleolar RNAs that guide modification in trypanosomatids: repertoire, targets, genome organisation, and unique functions. Int J Parasitol *34*, 445-454.

Vestheim, H., and Jarman, S.N. (2008). Blocking primers to enhance PCR amplification of rare sequences in mixed samples - a case study on prey DNA in Antarctic krill stomachs. Front Zool *5*, 12.

Vitali, P., Basyuk, E., Le Meur, E., Bertrand, E., Muscatelli, F., Cavaille, J., and Huttenhofer, A. (2005). ADAR2-mediated editing of RNA substrates in the nucleolus is inhibited by C/D small nucleolar RNAs. J Cell Biol *169*, 745-753.

von der Heyden, S., Chao, E.E., Vickerman, K., and Cavalier-Smith, T. (2004). Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of euglenozoa. J Eukaryot Microbiol *51*, 402-416.

Wachtel, C., and Michaeli, S. (2011). Functional analysis of noncoding RNAs in trypanosomes: RNA walk, a novel approach to study RNA-RNA interactions between small RNA and its target. Methods Mol Biol *718*, 245-257.

Wang, Z., Gerstein, M., and Synder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet *10*, 57-63.

Wassarman, K.M. (2007). 6S RNA: a small RNA regulator of transcription. Curr Opin Microbiol *10*, 164-168.

Waters, L.S., and Storz, G. (2009). Regulatory RNAs in bacteria. Cell *136*, 615-628.

Watkins, N.J., and Bohnsack, M.T. (2012). 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 *3*, 397-414.

Watkins, N.J., Gottschalk, A., Neubauer, G., Kastner, B., Fabrizio, P., Mann, M., and Luhrmann, R. (1998). Cbf5p, a potential pseudouridine synthase, and Nhp2p, a putative RNA-binding protein, are present together with Gar1p in all H BOX/ACA-motif snoRNPs and constitute a common bipartite structure. RNA *4*, 1549-1568.

Watkins, N.J., Segault, V., Charpentier, B., Nottrott, S., Fabrizio, P., Bachi, A., Wilm, M., Rosbash, M., Branlant, C., and Luhrmann, R. (2000). A common core RNP structure shared between the small nucleoar box C/D RNPs and the spliceosomal U4 snRNP. Cell *103*, 457-466.

Weber, M.J. (2006). Mammalian small nucleolar RNAs are mobile genetic elements. PLoS Genet *2*, e205.

Williams, D.J., Boots, J.L., and Hall, K.B. (2001). Thermodynamics of 2'-ribose substitutions in UUCG tetraloops. RNA *7*, 44-53.

Williams, G.T., and Farzaneh, F. (2012). Are snoRNAs and snoRNA host genes new players in cancer? Nat Rev Cancer *12*, 84-88.

Yang, C.Y., Zhou, H., Luo, J., and Qu, L.H. (2005). Identification of 20 snoRNA-like RNAs from the primitive eukaryote, *Giardia lamblia*. Biochem Biophys Res Commun *328*, 1224-1231.

Yang, J.H., Zhang, X.C., Huang, Z.P., Zhou, H., Huang, M.B., Zhang, S., Chen, Y.Q., and Qu, L.H. (2006). snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. Nucleic Acids Res *34*, 5112-5123.

Yuan, G., Klambt, C., Bachellerie, J.P., Brosius, J., and Huttenhofer, A. (2003). RNomics in *Drosophila melanogaster*: identification of 66 candidates for novel non-messenger RNAs. Nucleic Acids Res *31*, 2495-2507.

Zamudio, J.R., Mittra, B., Chattopadhyay, A., Wohlschlegel, J.A., Sturm, N.R., and Campbell, D.A. (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.

Zebarjadian, Y., King, T., Fournier, M.J., Clarke, L., and Carbon, J. (1999). Point mutations in yeast CBF5 can abolish in vivo pseudouridylation of rRNA. Mol Cell Biol *19*, 7461-7472.

Zemann, A., op de Bekke, A., Kiefmann, M., Brosius, J., and Schmitz, J. (2006). Evolution of small nucleolar RNAs in nematodes. Nucleic Acids Res *34*, 2676-2685.

Zhang, Y., Liu, J., Jia, C., Li, T., Wu, R., Wang, J., Chen, Y., Zou, X., Chen, R., Wang, X.J., *et al.* (2010). Systematic identification and evolutionary features of rhesus monkey small nucleolar RNAs. BMC Genomics *11*, 61.

Zhao, J., Sun, B.K., Erwin, J.A., Song, J.J., and Lee, J.T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science *322*, 750-756.

Zhou, H., Chen, Y.Q., Du, Y.P., and Qu, L.H. (2002). The *Schizosaccharomyces pombe* mgU6-47 gene is required for 2'-O-methylation of U6 snRNA at A41. Nucleic Acids Res *30*, 894-902.

Appendix 1: Supplementary Tables

Table S1. Oligonucleotides used to amplify snoRNA gene repeats in *E. gracilis* and the snoRNA species to which they anneal. If the oligonucleotide pair successfully annealed to the target region and a PCR product was generated, the designated cluster (region amplified) is provided

Oligonucleotide combinations	Target region	Region amplified	Sequence
oAR54 (Fwd) + oAR55 (Rev)	snoRNA Eg-p1	Cluster 1	5’ CAACGGGTCGAGTCAGTGCCCGC 3’ (Fwd) + 5’ GCACCGATGGATCTTAAAGATGC 3’ (Rev)
oAR123 + oAR124	snoRNA Eg-m4	Cluster 1.1	5’ CACTTGCTTGATGTCATTTGCAC 3’ (Fwd) + 5’ GTCAGGCGGTAATTCCAAATATC 3’ (Rev)
oAR21 + oAR22	snoRNA Eg-p2	Cluster 2, 2.1	5’ TGGGCAGCGGCGTATACGCAGTG 3’ (Fwd) + 5’ CATGTCGTCCGAAAGGCATGGCAG 3’ (Rev)
oAR23 + oAR24	snoRNA Eg-p3	Cluster 3, 3.1	5’ GCAAGCGGGCCCTCTGCCCAGCC 3’ (Fwd) + 5’ GCAGCCCTCAGTCCATTTGCCCA 3’ (Rev)
oAR34 + oAR35	snoRNA Eg-m7		5’ GCTGATCCAATCTTCATCCTGA 3’ (Fwd) + 5’ GTCATCACACAGCCGTTGAATGA 3’ (Rev)
oAR117 + oAR118	snoRNA Eg-m40	Cluster 4, 4.1	5’ CAGCGTGAGAAGGCATACTGTTG 3’ (Fwd) + 5’ GCGACGTCAATTTGGGCTCATCA 3’ (Rev)
oAR70 + oAR71	snoRNA Eg-m19	Cluster 4.2	5’ CGACGTGATTTCCGCCGTAGTTG 3’ (Fwd) + 5’ GTGCGAATTTCCAACGTCATCC 3’ (Rev)
oAR115 + oAR116	snoRNA Eg-m64	Cluster 5	5’ CCCCTGCTGAAGCAACACTTTTC 3’ (Fwd) + 5’ GGATGGTGTTCCCAAATCCCATC3’ (Rev)
oAR166 + oAR167	snoRNA Eg-m26	Cluster 5.1	5’ GACGTGACCCATGCCTGCTTTTC 3’ (Fwd) + 5’ GTCGGTCGCAAATTGAACGCATC 3’ (Rev)
oAR150 + oAR151	snoRNA Eg-m28	Cluster 6, 6.1	5’ GAAGGCGTGATGCAACCCCTGTG 3’ (Fwd) + 5’ CACAAAGGATGTTGAATTGAAAC 3’ (Rev)
oAR148 + oAR149	snoRNA Eg-m29	Cluster 6.2	5’ GCTTGATGTGGCGTCTGATGCTG 3’ (Fwd) + 5’ GTCAGGGCAAGATTCTATTCCTC 3’ (Rev)
oAR45 + oAR46	snoRNA Eg-m11	Cluster 7, 7.1	5’ CTGCTGACAGCTCATCAAATCAC 3’ (Fwd) + 5’ GGAGTGTTTTGGATTCCAAGCAT 3’ (Rev)
oAR140 + oAR141	snoRNA Eg-m37	Cluster 7.2	5’ CGGACACTGATGATGCCATACTG 3’ (Fwd) + 5’ CTCTCACTACGAAACTAGCATCC 3’ (Rev)
oAR170 + oAR171	snoRNA Eg-m6	Cluster 7.3	5’ CTGCTTGGTGATGACGCTCAATC 3’ (Fwd) + 5’ GAAAGGTGGTAGTATTGAAACTC 3’ (Rev)
oAR134 + oAR135	snoRNA Eg-m59	Cluster 8	5’ GATAGCAAATCTTGTTTCGCCTG 3’ (Fwd) + 5’ CAAATACTGGGGATAACACATCA 3’ (Rev)
oAR160 + oAR161	snoRNA Eg-m23	Cluster 8.1, 8.2	5’ GATGCTCCAACCGTTATTTCTGA 3’ (Fwd) + 5’ CAGGCAGCAGACAGTGCTGGTGA 3’ (Rev)
oAM71 + oAM72	snoRNA Eg-p5	Cluster 8.3	5’ GCCTGTTCAATGCGGCTCAAAG 3’ (Fwd) + 5’ CCCAACCTGCCTTCGAAAGGC 3’ (Rev)
oAR214 + oAR215	snoRNA Eg-m15	Cluster 9, 9.1	5’ GTTTCTCGATGAGGAGCACTTTT 3’ (Fwd) + 5’ CATATGCTTTGTTCAATCAAAAC 3’ (Rev)
oAR224 + oAR225	snoRNA Eg-m30	Cluster 9.2	5’ GCCTGAAGGCCTGATGCGGTCAC 3’ (Fwd) + 5’ CTGAAACTATGAGGATGACATCA 3’ (Rev)
oAR190 + oAR191	snoRNA Eg-m21.2	Cluster 10.1	5’ GCTGATGCGTCCTCTTCTCTGA3’ (Fwd) + 5’ GTCAGGGTATACTAACATTCCTT 3’ (Rev)
oAR158 + oAR191	snoRNA Eg-m21	Cluster 10.2, 10.3	5’ CTGACTGCTGATGTGTCCTCTTC 3’ (Fwd) + 5’ GTCAGGGTATACTAACATTCCTT 3’ (Rev)
oAR162 + oAR163	snoRNA Eg-m53	Cluster 10	5’ GATCCGCTGATGATCGTGTGTCG 3’ (Fwd) + 5’ GTGGCTAATACAAATGTCATCCG 3’ (Rev)
oAR144 + oAR145	snoRNA Eg-m31.2	Cluster 11, 11.1, 11.2	5’ GATGCACCTGATGAACAAATCTG 3’ (Fwd) + 5’ GATGCACCTGATGAACAAATCTG 3’ (Rev)
oAM81 + oAM82	snoRNA Eg-p10	Clusters 11.3, 11.4, 11.5	5’ CTCTCCAAGTCAGAATCTGCGG 3’ (Fwd) + 5’ GCCTCCCAAGAACCAAATTGCG 3’ (Rev)
oAR125 + oAR126	snoRNA Eg-m38	Cluster 12, 12.1	5’ GATGCCAGATTACAACCAAGATC 3’ (Fwd) + 5’ GTCATCAACGAAGAGAGGGGTCA 3’ (Rev)

oAR168 + oAR169	snoRNA Eg-m55	Cluster 12.2	5’ CTGACTGGTGACGACCCGATCTT 3’ (Fwd) + 5’ CGGACTCAAGGAGAGATGGATCA 3’ (Rev)
oAR186 + oAR187	snoRNA Eg-m55.1	Cluster 12.3, 12.4, 12.5, 12.6	5’ CTGACTGGTCATGACCCGATCTT 3’ (Fwd) + 5’ CGGACTCAAAGAGAGATGGATCA 3’ (Rev)
oAR221 + oAR222	snoRNA Eg-m25	Cluster 13	5’ CTGCCATGTCAAACTCAACAGGG 3’ (Fwd) + 5’ GGCCGGTTCCTGGTGAAATGGTA 3’ (Rev)
oAR223 + oAR222	snoRNA Eg-m25	Cluster 13.1	5’ CTGCCATGTCAATCTCAACAGGG 3’ (Fwd) + 5’ CAACTCACCTGCACCGATTGGTG 3’ (Rev)
oAR119 + oAR120	snoRNA Eg-m39	Cluster 13.2	5’ GATATGTGACGACCATTTTGTTTG 3’ (Fwd) + 5’ GATACTGTGCAAATTGATGGGTC 3’(Rev)
oAR113 + oAR114	snoRNA Eg-m65	Cluster 14	5’ CTGATTATGACCGATTTGTCCCG 3’ (Fwd) + 5’ CAACTCACCTGCACCGATTGGTG 3’ (Rev)
oAR231 + oAR232	snoRNA Eg-m48	Cluster 14.1	5’ GAATCGGATGATGGCAGTAACCC 3’ (Fwd) + 5’ CACCTGCTTTGATCTGTCCATCA 3’ (Rev)
oAR38 + oAR39	snoRNA Eg-m10	Cluster 15	5’ CCAAATGATCACAGCTTTCTCA 3’ (Fwd) + 5’ CACGTGAGTATCAAACCAGCAT 3’ (Rev)
oAR42 + oAR43	snoRNA Eg-m8	Cluster 16, 16.1	5’ GCGGTTTTCTCATCGTGCCTGA 3’ (Fwd) + 5’ GCATCTCAGATTCGGCATAGGG 3’ (Rev)
oAR48 + oAR49	snoRNA Eg-m12	Cluster 17	5’ CGGAGCACGATTGATTGACCCTC 3’ (Fwd) + 5’ CTGGGGCGACATGCTAATCATCA 3’ (Rev)
oAR66 + oAR67	snoRNA Eg-m17		5’ GAAAAGATTATTGCAGGATCAAC 3’ (Fwd) + 5’ GTTCAATCTGCGAAGAATCATCC 3’ (Rev)
oAR74 + oAR75	snoRNA Eg-m52		5’ GATGCTATCACACTCCAGTCCTT 3’ (Fwd) + 5’ CAATCAGGATACAACAGTCATCA 3’ (Rev)
oAR109 + oAR110	snoRNA Eg-m66	Cluster 18	5’ GATTATTACATTTAACCCTCTTTC 3’ (Fwd) + 5’ CAATCCATGAGCCAGCTGGATCA 3’ (Rev)
oAR111 + oAR112	snoRNA Eg-m5.1	Cluster 19	5’ GATTCAAGCATTCATCTGGAGCT 3’ (Fwd) + 5’ GTGTCGGTCATCGAAAGAAGA 3’ (Rev)
oAR121 + oAR122	snoRNA Eg-m63	Cluster 20	5’ CACCCTTTGTTGTTCCTCCACTT 3’ (Fwd) + 5’ GGGACTGGGACGGCAATCGAATC 3’ (Rev)
oAR127 + oAR128	snoRNA Eg-m62		5’ GATAAAGAATGATGCTGACCGTT 3’(Fwd) + 5’ GGCAAATGCCTCGCACAGATCAT 3’ (Rev)
oAR129 + oAR220	snoRNA Eg-m22	Cluster 21, 21.1	5’ CGCATCTTTCTTTCCTCGCTGAT 3’ (Fwd) + 5’ CATGCCTTCGTCAGGTTTGAGGT 3’ (Rev)
oAR136 + oAR137	snoRNA Eg-m35		5’ GAGCGCATGACGAGCCTCAGCCT 3’ (Fwd) + 5’ CAAGGGCTAAAGGACAGCTCATC 3’ (Rev)
oAR138 + oAR139	snoRNA Eg-m58		5’ CCCTGATGATTGATGTCTTGCGG 3’ (Fwd) + 5’ GACCGCCCGTGATTTTGGTGTCA 3’ (Rev)
oAR142 + oAR143	snoRNA Eg-m33	Cluster 22, 22.1, 22.2	5’ CTTGATGAGTACCTCGCTTGCTG 3’ (Fwd) + 5’ CTGGTCATGTCCCCGATTCATGC 3’ (Rev)
oAR146 + oAR147	snoRNA Eg-m57		5’ GTTGATGTTGGTCCTTTTTCTGA 3’ (Fwd) + 5’ CATTCAGATTGCTGTGCAAGGA 3’ (Rev)
oAR152 + oAR153	snoRNA Eg-m56	Cluster 23	5’ GGCCCACTGATGTGGTATCGCCA 3’ (Fwd) + 5’ CAACGCATGAGAGTAACAATGTC 3’ (Rev)
oAM75 + oAM76	snoRNA Eg-p7	Cluster 23.1	5’ ATGGCTGCGTGGGTCAACACC 3’ (Fwd) + 5’ TGCTGGACGATCCCCGAAGC 3’ (Rev)
oAR154 + oAR155	snoRNA Eg-m27	Cluster 24	5’ CATTGCATGAGTGACGCCGTTTG 3’ (Fwd) + 5’ GTGGCTGAAACTTATGCGAACGA 3’ (Rev)
oAR156 + oAR157	snoRNA Eg-m18	Cluster 25	5’ CCACCTGATGTGTGCACTCACTG 3’ (Fwd) + 5’ GCGTTTAATTGAAATCGACATCC 3’ (Rev)
oAR164 + oAR165	snoRNA Eg-m54	Cluster 26	5’ GATCTCCTGCTTTCAAATCGTTT 3’ (Fwd) + 5’ CAATTGTGTAGCCATGTGTTCAT 3’ (Rev)
oAM73 + oAM74	snoRNA Eg-p6	Cluster 26.1	5’ CAGATGTAAGAAGGGGCCAGA 3’ (Fwd) + 5’ GTTAGCAGAAGCAGTCAGGGG 3’ (Rev)
oAR172 + oAR173	snoRNA Eg-m67		5’ GAAGAGTGACCCTTTCACCCCTA 3’(Fwd) + 5’ GATTCAGACCATTCATATCATCA 3’ (Rev)
oAR174 + oAR175	snoRNA Eg-m45		5’ CTGATGTGCATTTTTCCCTCTGA 3’ (Fwd) + 5’ CGATCAGCCAGCTCTTCCTAGAC 3’ (Rev)
oAR176 + oAR177	snoRNA Eg-m47.1		5’ CTCCGACTGCGTGATGCACATCCT 3’ (Fwd) + 5’ GTAGTAGCGAATCCAGAGCATCC 3’ (Rev)

oAR178 + oAR179	snoRNA Eg-m47.2		5’ CCTGATGCACATCCTCTGTCACC 3’ (Fwd) + 5’ GTCGGTGGTTGTAGTAGCGAAT 3’ (Rev)
oAR180 + oAR182	snoRNA Eg-m16	Cluster 27, 27.1, 27.2, 27.3	5’ CTTTGCAATGACACCCATCGGCT 3’ (Fwd) + 5’ CGGGCCAAACAGCAAAGCTAATC 3’ (Rev)
oAR184 + oAR185	snoRNA Eg-m70	Cluster 28	5’ GCACCGTGATGAACATTCCCAAG 3’ (Fwd) + 5’ GGAATTGAACAATCAGCGCATCA 3’ (Rev)
oAR188 + oAR189	snoRNA Eg-m50	Cluster 29	5’ CTGACGAAACATCCCGATTGA 3’ (Fwd) + 5’ CAGAGTTCAAGGGCTAAGACTCA 3’ (Rev)
oAR212 + oAR43	snoRNA Eg-m9		5’ CGACTGCCTGACGTGATTTGGTG 3’ (Fwd) + 5’ GCATCTCAGATTCGGCATAGGG 3’ (Rev)
oAR213 + oAR43	snoRNA Eg-m8		5’ CATCGTGCCTGACGTGATTTGGT 3’ (Fwd) + 5’ GCATCTCAGATTCGGCATAGGG 3’ (Rev)
oAR216 + oAR217	snoRNA Eg-m20		5’ GAGATACATTCCCACCACTGAC 3’ (Fwd) + 5’ CAATGTCAAACAAGTGGTAAGTT 3’ (Rev)
oAR218 + oAR219	snoRNA Eg-m20.1		5’ GAGATATATTCCCACCACTGAC3’ (Fwd) + 5’ CAGTGTCAAACAAGTGGTAAGTT 3’ (Rev)
oAR226 + oAR227	snoRNA Eg-m32		5’ CATGATTGACACACAATCGACAC 3’ (Fwd) + 5’ CTCTCCATCTGAAACGTGGAATG 3’ (Rev)
oAR228 + oAR137	snoRNA Eg-m35.1		5’ GAGCGCATGATGAGCATCAGCCT 3’ (Fwd) + 5’ CAAGGGCTAAAGGACAGCTCATC 3’ (Rev)
oAR229 + oAR230	snoRNA Eg-m60	Cluster 30	5’ GAGACATGGAGAAATTGCATCCT 3’ (Fwd) + 5’ GATAGGGATAACTGTGTCAAGCC 3’ (Rev)
oAR233 + oAR234	snoRNA Eg-m14		5’ GAAAATGATGTGAACCGATTGGC 3’ (Fwd) + 5’ GCATAAGGGTGATGTGTGGTTTC 3’ (Rev)
oAR240 + oAR241	snoRNA Eg-m65		5’ CCTGATTTGCTTTATTCCGTTCC 3’ (Fwd) + 5’ GACCTGGAACCTTCCATCCATCA 3’ (Rev)
oAR243 + oAR244	snoRNA Eg-m44		5’ GATTAGAACCAAACCAACGCTCT 3’ (Fwd) + 5’ CCGCAGAACCTGGAGAGTAGCGA 3’ (Rev)
oAR246 + oAR247	snoRNA Eg-m46		5’ GTTGAAAGATCTGCTCTGGGCT 3’ (Fwd) + 5’ CATCAGGATTCACATTGGACAGG 3’ (Rev)
oAR248 + oAR249	snoRNA Eg-m69	Cluster 31, 31.1	5’ GATGTCACTGACTGACGCCACCT 3’ (Fwd) + 5’ GCGGTTGAGGACAGTCTGGGTGT 3’ (Rev)
oAM79 + oAM80	snoRNA Eg-p9	Cluster 31.3	5’ GCTACCTACTCTTCCGTCCCA 3’ (Fwd) + 5’ CTGCCACCCGATTGAGGTCC 3’ (Rev)
oAR272 + oAR273	snoRNA Eg-m13	Cluster 32, 32.1	5’ CAGCCATGTGAGATGCAATGGAA 3’ (Fwd) + 5’ GATTCCTGCTAAAAGTGGGGGTC 3’ (Rev)
oAR132 + oAR133	snoRNA Eg-m61		5’ GACAGCGCAGTGATTCTGGTCTG 3’ (Fwd) + 5’ GACCGTCGTGAAAGGGTGTCATC 3’ (Rev)
oAM77 + oAM78	snoRNA Eg-p8		5’ CGAACAGTAGGAAGGTGCAAAG 3’ (Fwd) + 5’ GTGCACAACAACTCGAATGGTG 3’ (Rev)
oAM83 + oAM84	snoRNA Eg-p11		5’ GCTGGTGGGTAGAGGTGGAGAG 3’ (Fwd) 5’ CGGGAC TGGCAAAGACAACGG 3’ (Rev)
oAM85 + oAM86	snoRNA Eg-p12		5’ CCTCAGCTCTCAGGCCAAGATG 3’ (Fwd) 5’ CCCCTCTAACCTATTCAGGCC 3’ (Rev)
oAM87 + oAM88	snoRNA Eg-m83		5’ TGTCAGTGATGCGGC ATGAAGC 3’ (Fwd) 5’ TCAGTAGCTGGGATCCATGGG 3’ (Rev)
oAM278 + oAM279	snoRNA Eg-m100		5’ ACCAAGAAGGATTCATCATC 3’ (Fwd) 5’ GATGACCGTATTTCTGACTG 3’ (Rev)

The first oligonucleotide in each pair is the forward (fwd) primer and is sense to the 3’ end of the target snoRNA sequence (usually directly upstream of the D box element if the target gene is a C/D box snoRNA and is directly upstream of the AGA box element if the target gene is an AGA box snoRNA). The second oligonucleotide in each pair is the reverse (rev) primer and is antisense to the 5’ end of the target snoRNA gene (usually immediately downstream of the C box element if the target gene is a C/D box snoRNA and near the 5’ end if the target gene is an AGA box snoRNA). Also refer to Russell et al., 2004 for Eg-p1 – Eg-p4 and Eg-m1 – Eg-m4 sequences, Russell et al., 2006 for Eg-m5 – Eg-m70 sequences, and Supplementary Figure 1 (this study) for Eg-m71 – Eg-m100 and Eg-p5 – Eg-p12 sequences.

Table S2. Oligonucleotides used to amplify *E. gracilis* snoRNA genes or snoRNA gene clusters by PCR, using isolated BAC DNA as template

Oligonucleotide combinations	snoRNA(s) targeted	snoRNA cluster targeted	Radiolabelled probe synthesized?	Sequence
oAM130 + oAM132	Eg-h1	-	✓	5’ GGGTGCGCCCCTCCAATGGCG 3’ (Fwd) + 5’ GTGTTACGCACGTGAAGATCC 3’ (Rev)
oAM116 + oAM117	SL-RNA gene	-		5’ ACTTTCTGAGTGTCTATTTTTTTTCG 3’ (Fwd) + 5’ GTTGCCGCCTCCAAAAATTGGAAG 3’ (Rev)
oKV22 + oKV24	Fibrillarin gene	-		5’ GGGAAAGGAAAGGGTGGAAAGGG 3’ (Fwd) + 5’ GGGAGAATTCAACTGCGTACAC 3’ (Rev)
oAR54 + oAR55	Eg-p1	Cluster 1		5’ CAACGGGTCGAGTCAGTGCCCGC 3’ (Fwd) + 5’ GCACCGATGGATCTTAAAGATGC 3’ (Rev)
oAR21 + oAR22	Eg-p2	Cluster 2	✓	5’ TGGGCAGCGCCGTATACGCAGTG 3’ (Fwd) + 5’ CATGTCGTCCGAAAGGCATGGCAG 3’ (Rev)
oAR24 + oAM101	Eg-p3 Eg-m3	Cluster 3	✓	5’ GCAGCCCTCAGTCCATTTGCCCA 3’ (Fwd) + 5’ CAGATCCTCCTGATGCAACC 3’ (Rev)
oAM100 + oAM101	Eg-m3	-		5’ ATGCAGTTATTCATCTCCCCTG 3’ (Fwd) + 5’ CAGATCCTCCTGATGCAACC 3’ (Rev)
oAM144 + oAM146	U14	-		5’ GGCAATGATTGACACTGAAC 3’ (Fwd) + 5’ GGCTCAGACGGGCCAGGGCC 3’ (Rev)
oAM149 + oAM150	Eg-m26	-		5’ GGATGATGCGTTCAATTTGCG 3’ (Fwd) + 5’ GGTCAGAAAAGCAGGCATGGG 3’ (Rev)
oAR115 + oAR116	Eg-m64	Cluster 5		5’CCCTGCTGAAGCAACACTTTTC 3’ (Fwd) + 5’GGATGGTGTTCCCAAATCCCATC 3’ (Rev)
oAM209 + oAM211	Eg-m29	-		5’ CAGGATGGAGAGGAATAG 3’ (Fwd) + 5’ ATTCAGCATCAGACGCCAC 3’ (Rev)
oAM40 + oAM39	Eg-m88	-		5’ ACTGATGGTGTGATGCAACTCT 3’ (Fwd) + 5’ AGTCAGGCAAAAAGCACATCAT 3’ (Rev)
oAM38 + oAM37	Eg-m11	-		5’ TGTGTGATGCTTGGAATCCAAA 3’ (Fwd) + 5’ GTGTGATTTGATGAGCTGTCAG 3’ (Rev)
oAR140 + oAM37	Eg-m37 Eg-m11	Cluster 7.2		5’CGGACACTGATGATGCCATAC 3’ (Fwd) + 5’ GTGTGATTTGATGAGCTGTCAG 3’ (Rev)
oAR170 + oAR171	Eg-m6	Cluster 7.3		5’ CTGCTTGGTGATGACGCTCAATC 3’ (Fwd) + 5’ GAAAGGTGGTAGTATTGAAACTC 3’ (Rev)
oAM71 + oAM72	Eg-p5	Cluster 8.3		5’ GCCTGTTCAATGCGGCTCAAAG 3’ (Fwd) + 5’ CCCAACCTGCCTTCGAAAGGC 3’ (Rev)
oAR162 + oAR163	Eg-m53	Cluster 10		5’ GATCCGCTGATGATCGTGTGTCG 3’ (Fwd) + 5’ GTGGCTAATACAAATGTCATCCG 3’ (Rev)
oAR190 + oAR191	Eg-m21.2	Cluster 10.1		5’ GCTGATGCGTCCTCTTCTCTGA 3’ (Fwd) + 5’ GTCAGGGTATACTAACATTCCTT 3’ (Rev)
oAR158 + oAR191	Eg-m21	Cluster 10.2		5’ CTGACTGCTGATGTGTCCTCTTC 3’ (Fwd) + 5’ GTCAGGGTATACTAACATTCCTT 3’ (Rev)
oAM179 + oAM180	Eg-m14.3	-		5’ GGATGAGAAAACCTTACCC 3’ (Fwd) + 5’ CCTCAGTAGGCCAATCGC 3’ (Rev)
oAR144 + oAR145	Eg-m31.2	Cluster 11		5’ GATGCACCTGATGAACAAATCTG 3’ (Fwd) + 5’ GTAATACATGTTCCTCAGGCATC 3’ (Rev)
oAM81 + oAM82	Eg-p10	Cluster 11.3		5’ CTCTCCAAGTCAGAATCTGCGG 3’ (Fwd) + 5’ GCCTCCCAAGAACCAAATTGCG 3’ (Rev)
oAM171 + oAM172	Eg-m31	-		5’ TGATGCCTGAGGAACATGTA 3’ (Fwd) + 5’ CCTCAGGCAGATTTGTTC 3’ (Rev)
oAR125 + oAR126	Eg-m38	Cluster 12		5’ GATGCCAGATTACAACCAAGATC 3’ (Fwd) + 5’ GTCATCAACGAAGAGAGGGGTCA 3’ (Rev)
oAR168 + oAR169	Eg-m55	Cluster 12.2		5’ GTCACTGGTGACGACCCGATCTT 3’ (Fwd) + 5’ CGGACTCAAGGAGAGATGGATCA 3’ (Rev)
oAR186 + oAR187	Eg-m55.1	Cluster 12.5		5’ CTGACTGGTCATGACCCGATCTT 3’ (Fwd) + 5’ CGGACTCAAAGAGAGATGGATCA 3’ (Rev)
oAR221 + oAR222	Eg-m25	Cluster 13		5’ CTGCCATGTCAAACTCAACAGGG 3’ (Fwd) + 5’ GGCCGGTTCCTGGTGAAATGGTA 3’ (Rev)
oAR223 + oAR222	Eg-m25	Cluster 13.1		5’ CTGCCATGTCAATCTCAACAGGG 3’ (Fwd) + 5’ CAACTCACCTGCACCGATTGGTG 3’ (Rev)

oAR119 + oAR120	Eg-m39	Cluster 13.2	✓	5’ GATATGTGACGACCATTTTGTTTG 3’ (Fwd) + 5’ GATACTGTGCAAATTGATGGGTC 3’(Rev)
oAR113 + oAR114	Eg-m65	Cluster 14	✓	5’ CTGATTATGACCGATTTGTCCCG 3’ (Fwd) + 5’ CAACTCACCTGCACCGATTGGTG 3’ (Rev)
oAR231 + oAR232	Eg-m48	Cluster 14.1		5’ GAATCGGATGATGGCAGTAACCC 3’ (Fwd) + 5’ CACCTGCTTTGATCTGTCCATCA 3’ (Rev)
oAM164 + oAM166	Eg-m43.1	-		5’ GATGACCAACCTTTGCTCC 3’ (Fwd) + 5’ GTCAGACTCGACGGAACTGC 3’ (Rev)
oAR42 + oAR43	Eg-m8	Cluster 16, 16.1	✓	5’ GCGGTTTTCTCATCGTGCCTGA 3’ (Fwd) + 5’ GCATCTCAGATTCGGCATAGGG 3’ (Rev)
oAR48 + oAR49	Eg-m12	Cluster 17		5’ CGGAGCACGATTGATTGACCCTC 3’ (Fwd) + 5’ CTGGGGCGACATGCTAATCATCA 3’ (Rev)
oAM17 + oAM20	Eg-m91	-	✓	5’ ACCAAGATGAAGACAGTGCCCAG 3’ (Fwd) + 5’ GTAACTATGACTCAAAGAACCAT 3’ (Rev)
oAM22 + oAM23	Eg-m66	-	✓	5’ TGGGATGATCCAGCTGGCTG 3’ (Fwd) + 5’ GCTCAGAAAGAGGGTTAAATG 3’ (Rev)
oAR109 + oAR110	Eg-m66	Cluster 18		5’ GATTATTACATTTAACCCTCTTTC 3’ (Fwd) + 5’ CAATCCATGAGCCAGCTGGATCA 3’ (Rev)
oAR111 + oAR112	Eg-m5.1	Cluster 19		5’ GATTCAAGCATTCATCTGGAGCT 3’ (Fwd) + 5’ GTGTCGGTCATCGAAAGAAGA 3’ (Rev)
oAR129 + oAR220	Eg-m22	Cluster 21, 21.1		5’ CGCATCTTTCTTTCCTCGCTGAT 3’ (Fwd) + 5’ CATGCCTTCGTCAGGTTTGAGGT 3’ (Rev)
oAM213 + oAM215	Eg-m56	Cluster 23	✓	5’ GACGACATTGTTACTCTC 3’ (Fwd) + 5’ GGTCAGCTGGCGATACCAC 3’ (Rev)
oAR156 + oAR157	Eg-m18	Cluster 25	✓	5’ CCACCTGATGTGTGCACTCACTG 3’ (Fwd) + 5’ GCGTTTAATTGAAATCGACATCC 3’ (Rev)
oAR164 + oAR165	Eg-m54	Cluster 26		5’ GATCTCCTGCTTTCAAATCGTTT 3’ (Fwd) + 5’ CAATTGTGTAGCCATGTGTTCAT 3’ (Rev)
oAM73 + oAM74	Eg-p6	Cluster 26.1		5’ CAGATGTAAGAAGGGGCCAGA 3’ (Fwd) + 5’ GTTAGCAGAAGCAGTCAGGGG 3’ (Rev)
oAM90 + oAM60	Eg-p6	-		5’ CCGCCCCTGACTGCTTCTGC 3’ (Fwd) + 5’ GGCCCCTTCTTACATCTGATAACC 3’ (Rev)
oAR180 + oAR182	Eg-m16	Cluster 27 - 27.2	✓	5’ CTTTGCAATGACACCCATCGGCT 3’ (Fwd) + 5’ CGGGCCAAACAGCAAAGCTAATC 3’ (Rev)
oAR180 + oAM56	Eg-m16 Eg-m80	Cluster 27.3		5’ CTTTGCAATGACACCCATCGGCT 3’ (Fwd) + 5’ CAGACTCCGTACACGAATCAC 3’ (Rev)
oAM175 + oAM176	Eg-m16	-		5’ GGATTAGCTTTGCTGTTTG 3’ (Fwd) + 5’ GGTCATCCAATCGAGCCG 3’ (Rev)
oAM51 + oAM65	Eg-p11	-		5’ TCCACCGTTGTCTTTGCCAGTC 3’ (Fwd) + 5’ TCTCTCCACCTCTACCCACCAGCC 3’ (Rev)
oAR184 + oAM44	Eg-m70 Eg-m94	Cluster 28		5’ GCACCGTGATGAACATTCCCAAG 3’ (Fwd) + 5’ TCAGCCAGAATTATGGAGCCTC 3’ (Rev)
oAR184 + oAR185	Eg-m70	Cluster 28		5’ GCACCGTGATGAACATTCCCAAG 3’ (Fwd) + 5’ GGAATTGAACAATCAGCGCATCA 3’ (Rev)
oAM51 + oAM48	Eg-p11	-		5’ TCCACCGTTGTCTTTGCCAGTC 3’ (Fwd) + 5’ CTACCCACCAGCCCAGCCATG 3’ (Rev)
oAM153 + oAM154	Eg-m77	-		5’ GGATGTGCTCATTAACCTCTG 3’ (Fwd) + 5’ GGCATCAGAACTCGCAGCGC 3’ (Rev)
oAR188 + oAR189	Eg-m50	Cluster 29		5’ CTGACGAAACATCCCGATTGA 3’ (Fwd) + 5’ CAGAGTTCAAGGGCTAAGACTCA 3’ (Rev)
oAM151 + oAM152	Eg-m34	-		5’ GGTGATGACCCCCAAGTCCC 3’ (Fwd) + 5’ GGGACTTCTCTGAGGTTTGACC 3’ (Rev)
oAR248 + oAR249	Eg-m69	Cluster 31, 31.1		5’ GATGTCACTGACTGACGCCACCT 3’ (Fwd) + 5’ GCGGTTGAGGACAGTCTGGGTGT 3’ (Rev)
oAM79 + oAM80	Eg-p9	Cluster 31.3		5’ GCTACCTACTCTTCCGTCCCA 3’ (Fwd) + 5’ CTGCCACCCGATTGAGGTCC 3’ (Rev)
oAR272 + oAR273	Eg-m13	Cluster 32, 32.1	✓	5’ CAGCCATGTGAGATGCAATGGAA 3’ (Fwd) + 5’ GATTCCTGCTAAAAGTGGGGGTC 3’ (Rev)

The first oligonucleotide in each pair is the forward (fwd) primer and is sense to the target snoRNA sequence indicated. The second oligonucleotide in each pair is the reverse (rev) primer and is antisense to the target snoRNA sequence indicated. If a snoRNA gene cluster was amplified (as opposed to a single snoRNA gene), the cluster number is indicated. If the oligonucleotides were used to synthesize radiolabelled probes, it is indicated by a checkmark.

Table S3. Oligonucleotides used for sequencing *E. gracilis* genomic inserts cloned into BACs by primer walking

Oligonucleotide	BAC template	Sequence
pCC1 Fwd	pCC1BAC	5' GGATGTGCTGCAAGGCGATTAAGTTGG 3' (230 to 256)
pCC1 Rev	pCC1BAC	5' CTCGTATGTTGTGTGGAATTGTGAGC 3' (489 to 464)
oAM256	BAC Eg-h1 F	5' CTGGTTTAAGCGAGCGTTCAGG 3' (+523 to +544)
oAM258	BAC_Eg-h1 F	5' TTCTCAGGTTCGGGACCCGT 3' (+1020 to +1039)
oAM257	BAC Eg-h1 R	5' TGCAAATGTCCAAACCTTAAGTGCG 3' (+549 to +525)
oAM266	BAC Eg-h1 R	5' TGGCAAATATGGGTGAGCACA 3' (+1103 to +1083)
oAM260	BAC 23 F	5' ATTTCCTCGGGCACTGAAGTC 3' (+838 to +858)
oAM268	BAC 23 F	5' GTGCGCATTCTGATTTGTTC 3' (+1424 to +1443)
oAM261	BAC 23 R	5' ATCGGCCACTCTTCTCGAAAACC 3' (+611 to +589)
oAM269	BAC 23 R	5' CTGGGCGAAATTGGCAGGATG 3' (+975 to +955)
The oligonucleotides were designed to anneal to *Euglena gracilis* genomic DNA fragments ligated into the pCC1BAC vector EcoRI cloning site. The initial sequences were obtained using oligonucleotides that anneal to the pCC1BAC vector near the cloning site (pCCI Fwd and pCCI Rev). BAC#_F indicates the oligo is based on sequence originally obtained using pCC1 Fwd as the sequencing primer. BAC#_R indicates the oligo is based on sequence originally obtained using pCC1 Rev as the sequencing primer. The nucleotide position within the insert DNA (relative to the vector cloning site) to which each oligo anneals is indicated.		

Table S4. Oligonucleotides for PCR amplification of *Euglena* snoRNA coding regions found in the BAC library, to determine snoRNA gene orientation and the distance between cloning sites and snoRNA genes

Oligonucleotide combinations	Region targeted	Sequence
oAM127 + oAM257	Eg-h1 BAC Eg-h1	5' CTTCACGTGCGTAACACATCG 3' (Sense) + 5' TGCAAATGTCCAAACCTTAAGTGCG 3' (Antisense)
oAM260 + oAR152	BAC 23 Eg-m56	5' ATTTCCTCGGGCACTGAAGTC 3' (Sense) + 5' GGCCCACTGATGTGGTATCGCCA 3' (Sense)
oAM260 + oAM215	BAC 23 Eg-m56	5' ATTTCCTCGGGCACTGAAGTC 3' (Sense) + 5' GGTCAGCTGGCGATACCAC 3' (Antisense)
oAM268 + oAR152	BAC 23 Eg-m56	5' GTGCGCATTCTGATTTGTTC 3' (Sense) + 5' GGCCCACTGATGTGGTATCGCCA 3' (Sense)
oAM268 + oAM215	BAC 23 Eg-m56	5' GTGCGCATTCTGATTTGTTC 3' (Sense) + 5' GGTCAGCTGGCGATACCAC 3' (Antisense)
oAR152 + oAM261	Eg-m56 BAC 23	5' GGTCAGCTGGCGATACCAC 3' (Antisense) + 5' ATCGGCCACTCTTCTCGAAAACC 3' (Antisense)
oAM215 + oAM261	Eg-m56 BAC 23	5' GGTCAGCTGGCGATACCAC 3' (Antisense) + 5' ATCGGCCACTCTTCTCGAAAACC 3' (Antisense)
oAR152 + oAM269	Eg-m56 BAC 23	5' GGTCAGCTGGCGATACCAC 3' (Sense) + 5' CTGGGCGAAATTGGCAGGATG 3' (Antisense)
oAM215 + oAM269	Eg-m56 BAC 23	5' GGTCAGCTGGCGATACCAC 3' (Sense) + 5' CTGGGCGAAATTGGCAGGATG 3' (Antisense)
pCCI Fwd + oAR21	pCCIBAC Eg-p2	5' GGATGTGCTGCAAGGCGATTAAGTTGG 3' (Sense) + 5' TGGGCAGCGGCGTATACGCAGTG 3' (Antisense)
oAR22 + pCCI Rev	Eg-p2 pCCIBAC	5' CATGTCGTCCGAAAGGCATGGCAG 3' (Sense) + 5' CTCGTATGTTGTGTGGAATTGTGAGC 3' (Antisense)
oAR156 + pCCI Rev	Eg-m18 pCCIBAC	5' CCACCTGATGTGTGCACTCACTG 3' (Sense) + 5' CTCGTATGTTGTGTGGAATTGTGAGC 3' (Antisense)
One oligonucleotide in each pair anneals to the region surrounding the snoRNA gene (indicated by BAC number) and the other oligonucleotide anneals to the gene encoding the snoRNA indicated. The orientation of the primer		

relative to the targeted region is indicated. Because the orientation of the snoRNA genes relative to the vector is unknown, a combination of sense and antisense oligonucleotides were used.

Table S5. Oligonucleotides used to demonstrate polycistronic expression of snoRNA gene repeats by RT-PCR in *E. gracilis.* The location (cluster) of the targeted snoRNAs is indicated

Oligonucleotide combinations	snoRNA(s) targeted	Location of targeted snoRNAs	Sequence
oAR24 + oAM99	Eg-p3 U14	Cluster 3	5' GCAGCCCTCAGTCCATTTGCCCA 3' (Fwd +38 to +60) + 5' CTCAGACGGGCCAAGGCC 3' (Rev +98 to +81)
oAR24 + oAM101	Eg-p3 Eg-m3	Cluster 3	5' GCAGCCCTCAGTCCATTTGCCCA 3' (Fwd +38 to +60) + 5' CAGATCCTCCTGATGCAACC 3'(Rev +56 to +37)
oAM100 + oAM99	Eg-m3 U14	Cluster 3	5' ATGCAGTTATTCATCTCCCCTG 3' (Fwd +6 to +27) + 5' CTCAGACGGGCCAAGGCC 3' (Rev +98 to +81)
oAM98 + oAM101	U14 Eg-m3	Cluster 3	5' TGATTGACACTGAACCTCGTTC 3' (Fwd +6 to +27) + 5' CAGATCCTCCTGATGCAACC 3' (Rev +56 to +37)
oAR140 + oAM41	Eg-m37 Eg-m6	Cluster 7.2	5' CGGACACTGATGATGCCATACTG 3' (Fwd +34 to +56) + 5' GATCAGGGATTGAGCGTCATCA 3' (Rev +64 to +43)
oAR140 + oAM39	Eg-m37 Eg-m88	Cluster 7.2	5' CGGACACTGATGATGCCATACTG 3' (Fwd +34 to +56) + 5' AGTCAGGCAAAAAGCACATCAT 3' (Rev +68 to +47)
oAM42 + oAM39	Eg-m6 Eg-m88	Cluster 7.2	5' GCAGGATGAGTTTCAATACTAC 3' (Fwd +1 to +22)+ 5' AGTCAGGCAAAAAGCACATCAT 3' (Rev +68 to +47)
oAM42 + oAM37	Eg-m6 Eg-m11	Cluster 7.2	5' GCAGGATGAGTTTCAATACTAC 3' (Fwd +1 to +22) 5' GTGTGATTTGATGAGCTGTCAG 3' (Rev +62 to +41)
oAM71 + oAM72	Eg-p5	Cluster 8.3	5' GCCTGTTCAATGCGGCTCAAAG 3' (Fwd +41 to +62) 5' CCCAACCTGCCTTCGAAAGGC 3' (Rev +27 to +7)
oAR214+ oAR225	Eg-m15 Eg-m30	Cluster 9	5' GTTTCTCGATGAGGAGCACTTTT 3' (Fwd +37 to +59) 5' CTGAAACTATGAGGATGACATCA 3' (Rev +27 to +5)
oAR233 + oAR191	Eg-m14 Eg-m21.2	Cluster 10.1	5' GAAAATGATGTGAACCGATTGGC 3' (Fwd +35 to +57) 5' GTCAGGGTATACTAACATTCCTT 3' (Rev +37 to +15)
oAM28 + oAR234	intergenic region Eg-m14	Cluster 10	5' ATACGACAGTCTTCGACCTTGC 3' (Fwd +778 to +779) + 5' GCATAAGGGTGATGTGTGGTTTC 3' (Rev +1377 to +1355)
oAR144+ oAM52	Eg-m31.2 Eg-p10.1	Cluster 11.1	5' GATGCACCTGATGAACAAATCTG 3' (Fwd +27 to +49) + 5'CTGACTTGCGAAGGCACTTTG 3' (Rev +52 to +32)
oAM81 + oAM82	Eg-p10	Cluster 11.3	5' CTCTCCAAGTCAGAATCTGCGG 3' (Fwd +40 to +61) + 5' GCCTCCCAAGAACCAAATTGCG 3' (Rev + 29 to +8)
oAR186+ oAR126	Eg-m55.1 Eg-m38	Cluster 12	5' CTGACTGGTCATGACCCGATCTT 3' (Fwd +35 to +57) 5' GTCATCAACGAAGAGAGGGGTCA 3' (Rev + 29 to +7)
oAR231 + oAM49	Eg-m48 Eg-m98	Cluster 14.1	5' GAATCGGATGATGGCAGTAACCC 3' (Fwd +34 to +56) + 5' CTCAGATGCACGTTGTGATAG 3' (Rev +63 to +43)
oAR109 + oAR110	Eg-m66 Eg-m66.1	Cluster 18	5' GATTATTACATTTAACCCTCTTTC 3' (Fwd +31 to +54) + 5' CAATCCATGAGCCAGCTGGATCA 3' (Rev +28 to +6)
oAM77 + oAM78	Eg-p8	Cluster 21	5' CGAACAGTAGGAAGGTGCAAAG 3' (Fwd +35 to +56) + 5' GTGCACAACAACTCGAATGGTG 3' (Rev +25 to +4)
oAR142 + oAR143	Eg-m33	Cluster 22	5' CTTGATGAGTACCTCGCTTGCTG 3' (Fwd +35 to +57) + 5' CTGGTCATGTCCCCGATTCATGC 3' (Rev +33 to +1)
oAR142 + oAM43	Eg-m33 Eg-m92	Cluster 22	5' CTTGATGAGTACCTCGCTTGCTG 3' (Fwd +35 to +57) + 5' AGTCAGCACAGAGGTCATAAAC 3' (Rev +63 to +42)
oAM75 + oAM76	Eg-p7	Cluster 23.1	5' ATGGCTGCGTGGGTCAACACC 3' (Fwd +37 to +57) + 5' TGCTGGACGATCCCCGAAGC 3' (Rev +23 to +4)
oAM73 + oAM74	Eg-p6	Cluster 26.1	5' CAGATGTAAGAAGGGGCCAGA 3' (Fwd +41 to +61) + 5' GTTAGCAGAAGCAGTCAGGGG 3' (Rev +24 to +4)
oAR180 + oAM56	Eg-m16 Eg-m80	Cluster 27.3	5' CTTTGCAATGACACCCATCGGCT 3' (Fwd +28 to +50) + 5' CAGACTCCGTACACGAATCAC 3' (Rev +58 to +38)

oAR184 + oAM44	Eg-m70 Eg-m94	Cluster 28	5’ GCACCGTGATGAACATTCCCAAG 3’ (Fwd +38 to +60) + 5’ TCAGCCAGAATTATGGAGCCTC 3’ (Rev +58 to +36)
oAM83 + oAM84	Eg-p11	Cluster 28.1	5’ GCTGGTGGGTAGAGGTGGAGAG 3’ (Fwd +47 to +68) 5’ CGGGACTGGCAAAGACAACGG 3’ (Rev +28 to +8)
oAR229 + oAM47	Eg-m60 Eg-m96	Cluster 30	5’ GAGACATGGAGAAATTGCATCCT 3’ (Fwd +32 to +54) + 5’ TAATCACAATGTCAGTGCGTAT 3’ (Rev +43 to +22)
oAM79 + oAM80	Eg-p9	Cluster 31.3	5’ GCTACCTACTCTTCCGTCCCA 3’ (Fwd +45 to +65) + 5’ CTGCCACCCGATTGAGGTCC 3’(Rev +25 to +6)
The first oligonucleotide in each pair is the forward (fwd) primer and is sense to the 1st target snoRNA indicated (usually near the 3’ end). The second oligonucleotide is the reverse (rev) primer, which was used in the cDNA synthesis step and is antisense to the 2nd target snoRNA sequence indicated (usually near the 3’ end). The nucleotide positions within the target snoRNA/cluster sequences where the oligos anneal are indicated.			

Table S6. Oligonucleotides used to map the 3′ ends of box AGA snoRNAs identified in *E. gracilis* using 3′ RACE

Oligonucleotide	snoRNA targeted	Sequence
oAM89	Eg-p5	5’ CTTTCGAAGGCAGGTTGGGCTC 3’ (+9 to +30)
oAM90	Eg-p6	5’ CCGCCCCTGACTGCTTCTGC 3’ (+1 to +20)
oAM91	Eg-p7	5’ TCCGCTTCGGGGATCGTCCAGCAG 3’ (+1 to +24)
oAM92	Eg-p8	5’ ATGCACCATTCGAGTTGTTGTG 3’ (+1 to +22)
oAM93	Eg-p9	5’CTGGGACCTCAATCGGGTGGC 3’ (+1 to +21)
oAM50	Eg-p10	5’GAATCCGCAATTTGGTTCTTGG 3’ (+3 to +34)
oAM51	Eg-p11	5’ TCCACCGTTGTCTTTGCCAGTC 3’ (+1 to +22)
oAM94	Eg-p12	5’ GGCCTGAATAGGTTAGAGGGGAC 3’ (+5 to +27)
oAM122	Eg-p13	5’ CTTGTGGACTGGCCATCTTC 3’ (+10 to +29)
P-94	-	5’ AATAAAGCGGCCGCGGATCCAA(T_{17})C/A/G 3’
Each oligonucleotide listed is sense to the 5’ end of the target snoRNA. The nucleotide positions within the target snoRNA sequences where the oligonucleotides anneal are indicated.		

Table S7. Oligonucleotides used during synthesis of the *E. gracilis* small/capped RNA library

Oligonucleotide	Sequence	Description
5’ RNA linker	5’ GCUGAUGGCGAUGAAUGAACACUGCGUUUGCU GGCUUUGAUGAAA 3’	RNA linker ligated to the 5’ ends of size-selected or capped enriched RNA
oAR8	5’ CTCCCGCTTCCAGATCTCGAG($C)_{15}$G/A/T 3’	Poly-C oligo used during cDNA synthesis and subsequent PCR amplification (Rev)
oAM265	5’ GTTTGCTGGCTTTGATGAAA 3’	Anneals to the 5’ linker (+26 to +45) during PCR amplification (Fwd)

Table S8. Blocker oligonucleotides used during PCR amplification of cDNA synthesized from *E. gracilis* small RNA or cap-enriched RNA, to prevent amplification of *E. gracilis* LSU rRNA fragments

Oligonucleotide	LSU fragment targeted	Sequence
oAM264	14	5’ GGCTTTGATGAAAtgtccgatccgtt/3SpC3/ 3’
oAM284	11	5’ GGCTTTGATGAAAggagcatcgaggc/3SpC3/ 3’
oAM286	13	5’ GGCTTTGATGAAAccaacaccccgcc/3SpC3/ 3’
oAM288	12	5’ GGCTTTGATGAAAcggagtgattgcc/3SpC3/ 3’
oAM290	4	5’ GGCTTTGATGAAAtgaccaagcgtct/3SpC3/ 3’
oAM292	2	5’ GGCTTTGATGAAAgtgacctggcgca/3SpC3/ 3’
oAM294	1	5’ GGCTTTGATGAAAacctgttgtggtg/3SpC3/ 3’
oAM296	10	5’ GGCTTTGATGAAAtgggcgtgacaat/3SpC3/ 3’
oAM298	7	5’ GGCTTTGATGAAAcgggcaggaatgg/3SpC3/ 3’
oAM300	3	5’ GGCTTTGATGAAActcgatcggcttt/3SpC3/ 3’
oAM301	6	5’ GGCTTTGATGAAAcgttgaacaatgg/3SpC3/ 3’
oAM302	9	5’ GGCTTTGATGAAAggtgcgagcctgc/3SpC3/ 3’
oAM303	5	5’ GGCTTTGATGAAAtcgctggtggttg/3SpC3/ 3’
oAM304	8	5’ GGCTTTGATGAAAaagtggcagtcac/3SpC3/ 3’
Nucleotides 1-13 (capital letters) of each blocker oligonucleotide are sense to the 3’ end of the 5’ RNA linker. Nucleotides 14-26 (lowercase letters) of each blocker oligonucleotide are sense to the 5’ end of the LSU species indicated. /3SpC3/ = C3 spacer modification		

Table S9. Oligonucleotides used to synthesize templates for *in vitro* transcription of *Euglena* RNAs

Oligonucleotide combinations	Region targeted	Sequence
oKV21 + oKV24	Fibrillarin CDS	5’ *taatacgactcactataGG*GAAAGGAAAGGGTGGAAAGGG 3’ (Fwd + T7) + 5’ GGGAGAATTCAACTGCGTACAC 3’ (Rev)
oKV22 + oKV23	Fibrillarin CDS	5’ GGGAAAGGAAAGGGTGGAAAGGG 3’ (Fwd) + 5’ *taatacgactcactataGG*GAGAATTCAACTGCGTACAC 3’ (Rev + T7)
oAM137 + oAM140	pCR2.1-TOPO vector	5’ *taatacgactcactatagggA*GATATCCATCACACTGGCGG 3’ (Fwd + T7) + 5’ GTTTTCCCAGTCACGACGTTG 3’ (Rev)
oAM138 + oAM139	pCR2.1-TOPO vector	5’ AGATATCCATCACACTGGCGG 3’ (Fwd) + 5’ *taatacgactcactatagg*GTTTTCCCAGTCACGACGTTG 3’ (Rev + T7)
oMA1 + oMA4	Cbf5 CDS	5’ *taatacgactcactataGG*GCACCGCGGGACACATGGC 3’ (Fwd + T7) + 5’ GGGCGATGCCGAGGGCGATGG 3’ (Rev)
oMA2 + oMA3	Cbf5 CDS	5’ GGGCACCGCGGGACACATGGC (Fwd) + 5’ *taatacgactcactataGG*GCGATGCCGAGGGCGATGG 3’ (Rev + T7)
oAM133 + oAM136	U1	5’ *taatacgactcactataGG*GTAGCGACTGTGATCACGC 3’ (Fwd + T7) + 5’ GGTGGGTCGACGCGCAAGCGC 3’ (Rev)
oAM134 + oAM135	U1	5’ GGGTAGCGACTGTGATCACGC 3’ (Fwd) + 5’ *taatacgactcactatagG*GTGGGTCGACGCGCAAGCGC 3’ (Rev + T7)
oAM216 + oAM219	U3	5’ *taatacgactcactagggA*AGACTGTACTCCACAAGG 3’ (Fwd + T7) + 5’ CTCAAATTGCTGACCTCTC 3’ (Rev)
oAM217 + oAM218	U3	5’ AAGACTGTACTCCACAAGG 3’ (Fwd) + 5’ *taatacgactcactagggC*TCAAATTGCTGACCTCTC 3’ (Rev + T7)
oAM143 + oAM146	U14	5’ *taatacgactcactatagGG*CAATGATTGACACTGAAC 3’ (Fwd + T7) + 5’ GGCTCAGACGGGCCAGGGCC 3’ (Rev)

oAM144 + oAM145	U14	5’ GGCAATGATTGACACTGAAC 3’ (Fwd) + 5’ *taatacgactcactatagGG*CTCAGACGGGCCAGGGCC 3’ (Rev + T7)
oAM129 + oAM132	Eg-h1	5’ *taatacgactcactataGG*GTGCGCCCCTCCAATGGCG 3’ (Fwd + T7) + 5’ GTGTTACGCACGTGAAGATCC 3’ (Rev)
oAM130 + oAM131	Eg-h1	5’ GGGTGCGCCCCTCCAATGGCG 3’ (Fwd) + 5’ *taatacgactcactatagg*GTGTTACGCACGTGAAGATCC 3’ (Rev + T7)
oAM177 + oAM180	Eg-m14.3	5’ *taatacgactcactatagGG*ATGAGAAAACCTTACCC 3’ (Fwd + T7) 5’ CCTCAGTAGGCCAATCGC 3’ (Rev)
oAM179 + oAM178	Eg-m14.3	5’ GGATGAGAAAACCTTACCC 3’ (Fwd) 5’ *taatacgactcactatagggC*CTCAGTAGGCCAATCGC 3’ (Rev + T7)
oAM208 + oAM211	Eg-m29	5’ *taatacgactcactatagggC*AGGATGGAGAGGAATAG 3’ (Fwd + T7) 5’ ATTCAGCATCAGACGCCAC 3’ (Rev)
oAM209 + oAM210	Eg-m29	5’ CAGGATGGAGAGGAATAG 3’ (Fwd) 5’ *taatacgactcactatagggA*TTCAGCATCAGACGCCAC 3’ (Rev + T7)
oAM169 + oAM172	Eg-m31	5’ *taatacgactcactataggG*ATGCCTGAGGAACATG 3’ (Fwd + T7) 5’ CCTCAGGCAGATTTGTTC 3’ (Rev)
oAM171 + oAM170	Eg-m31	5’ TGATGCCTGAGGAACATGTA 3’ (Fwd) 5’ *taatacgactcactatagggC*CTCAGGCAGATTTGTTC 3’ (Rev + T7)
oAM163 + oAM166	Eg-m43	5’ *taatacgactcactataggG*ATGACCAACCTTTGCTCC 3’ (Fwd + T7) 5’ GTCAGACTCGACGGAACTGC 3’ (Rev)
oAM164 + oAM165	Eg-m43	5’ GATGACCAACCTTTGCTCC 3’ (Fwd) 5’ *taatacgactcactataggG*TCAGACTCGACGGAACTGC 3’ (Rev + T7)
oAM239 + oAM240	Eg-m49	5’ *taatacgactcactatagggC*AAATGATTACTGTTTACAAC 3’ (Fwd + T7) 5’ GCTCATTTTACGTTGGTAAC 3’ (Rev)
oAM238 + oAM241	Eg-m49	5’ CAAATGATTACTGTTTACAAC 3’ (Fwd) 5’ *taatacgactcactataggG*CTCATTTTACGTTGGTAAC 3’ (Rev + T7)
oAM212 + oAM215	Eg-m56	5’ *taatacgactcactataggG*ACGACATTGTTACTCTC 3’ (Fwd + T7) 5’ GGTCAGCTGGCGATACCAC 3’ (Rev)
oAM213 + oAM214	Eg-m56	5’ GACGACATTGTTACTCTC 3’ (Fwd) 5’ *taatacgactcactatagGG*TCAGCTGGCGATACCAC 3’ (Rev + T7)
oAM161 + oAM156	Eg-m92	5’ *taatacgactcactataggG*ATGATCTTTCTGCTGTTCC 3’(Fwd + T7) + 5’ ggTCAGCACAGAGGTCATAAAC 3’ (Rev)
oAM155 + oAM162	Eg-m92	5’ ggATGATCTTTCTGCTGTTCC 3’ (Fwd) + 5’ *taatacgactcactataggG*TCAGCACAGAGGTCATAAAC 3’ (Rev + T7)
The first oligonucleotide in each pair is the forward (Fwd) primer and is sense to the 5’ end of the target sequence. The second oligonucleotide in each pair is the reverse (Rev) primer and is antisense to the 3’ end of the target. The T7 RNA polymerase promoter sequence is italicized and in lower case, and gene specific nucleotides are capitalized.		

Table S10. Oligonucleotides used to detect 2′-O-methylated rRNA sites by primer extension in *Euglena gracilis*

Oligonucleotide	2′-O-methylation site	Guide snoRNA	Sequence
oAM158	SSU1371	Eg-m49	5’ GGCCGGCATCGTTTACAGTGG 3’ (SSU 1437-1417)
oAM159	LSU97	Eg-m92	5’ ggCAACACTGGCGCACAGGC 3’ (LSU 163-146)
oAM160	LSU949	Eg-m43	5’ gGACTAGCCCCATCTGGAAG 3’ (LSU 1033-1015)
oAM167	SSU186	Eg-m31	5’ GGTGAAGGGACTGTCCGGAG 3’ (SSU 252-233)
oAM168	LSU1218	Eg-m14.3	5’ gGCTATCCTGAGGGAAACTTC 3’ (LSU 1269-1250)
Each oligonucleotide binds downstream of a 2′-O-methylated rRNA site, which is guided by the snoRNA indicated. The region of rRNA (LSU = large subunit; SSU = small subunit) to which the oligonucleotides bind is indicated. Lower case letters indicate nucleotides which were added to increase radioactive 5′ end labelling efficiency.			

Table S11. Oligonucleotides used to detect 2′-O-methylated rRNA sites by RTL-P in *Euglena gracilis*

Oligonucleotide	2′-O-methylation site	Guide snoRNA	Oligonucleotide orientation	Sequence
oAM159	LSU 97	Eg-m92	Reverse	5'ggCAACACTGGCGCACAGGC 3' (LSU 163-146)
oAM188	LSU 97	Eg-m92	Forward Downstream (F_D)	5' GGAACAGCTTCGAACGCAAC 3' (LSU 98-117)
oAM187	LSU 97	Eg-m92	Forward Upstream (F_U)	5' GGCCCAGGTTCTGAGGAAG 3' (LSU 22-40)
oAM160	LSU 949	Eg-m43	Reverse	5' gGACTAGCCCCATCTGGAAG 3' (LSU 1033-1015)
oAM190	LSU 949	Eg-m43	Forward Downstream (F_D)	5' CAGTGAGAAATTGGCAGCAG 3' (LSU 868-887)
oAM189	LSU 949	Eg-m43	Forward Upstream (F_U)	5' TAAGGAGCTTGCCAGTCGG 3' (LSU 950-968)
oAM200	LSU 2022	Eg-m29	Reverse	5' CACAGGGGATTCACTGGTTC 3' R (LSU 2114-2095)
oAM202	LSU 2022	Eg-m29	Forward Downstream (F_D)	5' CTACTGTCATGGGCTAGAGG 3' (LSU 2027-2046)
oAM201	LSU 2022	Eg-m29	Forward Upstream (F_U)	5' CGGGCAGGAATGGCAAACAG 3' (LSU 1989-2008)
oAM203	LSU 1324	Eg-m56	Reverse	5' AAGCATGCCTGGGTTGTCCG 3' (LSU 1405-1386)
oAM205	LSU 1324	Eg-m56	Forward Downstream (F_D)	5' GCCCAGCTCCTCAACCTATC 3' (LSU 1347-1366)
oAM204	LSU 1324	Eg-m56	Forward Upstream (F_U)	5' GTATGTGACTGGTTGCATCC 3' (LSU 1293-1312)

Each oligonucleotide binds in the region surrounding a 2′-O-methylated rRNA site, which is guided by the snoRNA indicated. The region of rRNA (LSU = large subunit; SSU = small subunit) to which the oligonucleotides bind and the orientation of the oligonucleotide relative to the methylated site are indicated. Lower case letters indicate nucleotides which were added to increase radioactive 5′ end labelling efficiency.

Table S12. Characteristics of box C/D snoRNAs identified in *Euglena gracilis*

Box C/D RNA	Size (nt)	Target site	Target match*	Guide position
Eg-m14.3	65	LSU1204 (A) & LSU1218 (C)	12/0 & 11/0	D & D'
Eg-m71	58	LSU1315 (U)	14/0	D'
Eg-m72	74	LSU1826 (G)	14/0	D'
Eg-m73	76	-	-	-
Eg-m74	72	LSU2605 (U)	13/0	D'
Eg-m75	61	LSU2979 (G)	12/0	D'
Eg-m76	62	LSU206 (U)	12/0	D'
Eg-m77	60	LSU3957 (C)	12/0	D'
Eg-m78	67	LSU3718 (U)	12/0	D'
Eg-m79	61	LSU631 (G)	13/0	D'
Eg-m80	68	LSU628 (G)	12/0	D'
Eg-m81	68	LSU2610 (G)	10/0	D'
Eg-m82	62	LSU2809 (A)	12/0	D'
Eg-m83	58	LSU1270 (U)	13/0	D'
Eg-m84	65	LSU2870 (G)	12/0	D'
Eg-m85	62	LSU1655 (C)	11/0	D'
Eg-m86	60	LSU1553 (U)	12/0	D
Eg-m87	59	SSU621 (C)	9/0	D'
Eg-m88	68	LSU1106 (A)	14/0	D'
Eg-m89	65	LSU2996 (C)	11/0	D'
Eg-m90	56	LSU1824 (U)	13/0	D
Eg-m91	65	LSU2744 (A)	14/0	D
Eg-m92	63	LSU97 (A)	12/0	D'
Eg-m93	91	LSU577 (C)	12/1	D
Eg-m94	60	LSU1626 (C)	12/0	D'
Eg-m95	65	LSU1624 (A)	14/0	D'
Eg-m96	62	SSU702 (U)	14/0	D'
Eg-m97	59	LSU2009 (G)	12/0	D'
Eg-m98	64	SSU2123 (C)	12/0	D'
Eg-m99	62	SSU1681 (G)	12/0	D'
Eg-m100	58	LSU936 (C)	12/0	D'
Eg-m101	62	LSU2853 (C)	14/0	D
Eg-m102	62	LSU2806 (U)	13/0	D'
Eg-m103	66	LSU3187 (G)	12/0	D
Eg-m104	56	SSU1037 (C)	13/0	D'
Eg-m105	59	LSU1502 (A)	13/0	D′
Eg-m106	63	LSU39 (A)	12/0	D'
Eg-m107	67	LSU3315 (A)	11/0	D'
Eg-m108	57	LSU1039 (A)	13/0	D'
Eg-m109	74	LSU1608 (G)	13/0	D'

Eg-m110	60	LSU1819 (G)	14/0	D
Eg-m111	67	LSU2768 (A)	13/0	D'
Eg-m112	57	LSU1081 (U)	14/0	D'
Eg-m113	62	LSU2993 (C)	9/0	D'
Eg-m114	64	LSU2005 (A)	14/0	D'
Eg-m115	60	LSU3694 (U)	13/0	D'
Eg-m116	60	SSU2180 (A)	13/0	D'
Eg-m117	66	LSU2776 (G)	12/0	D'
Eg-m118	65	LSU2909 (U)	13/0	D'
Eg-m119	64	LSU3208 (C)	14/0	D'
Eg-m120	64	LSU3374 (C)	14/0	D'
Eg-m121	62	LSU3906 (A)	13/0	D'
Eg-m122	77	LSU2849 (A)	11/0	D
Eg-m123	63	LSU3688 (G)	12/0	D'
Eg-m124	63	LSU41 (G)	13/0	D'
Eg-m125	64	LSU3367 (A)	12/0	D'
Eg-m126	60	LSU1683 (G)	12/0	D'
Eg-m127	64	LSU1738 (A)	11/0	D'
Eg-m128	60	LSU1121 (A)	11/0	D'
Eg-m129	60	LSU2953 (G)	11/0	D'
Eg-m130	65	SSU1966 (U)	12/0	D
Eg-m131	62	SSU1736 (G)	13/0	D'
Eg-m132	65	LSU2685 (C)	11/0	D'
Eg-m133	62	SSU110 (A)	11/0	D'
Eg-m134	63	LSU1189 (U)	13/0	D'
Eg-m135	61	LSU1452 (U)	10/0	D'
Eg-m136	60	LSU491 (U)	15/0	D'
Eg-m137	59	LSU2146 (G)	10/0	D'
Eg-m138	61	LSU1888 (A)	12/0	D
Eg-m139	62	LSU483 (C)	10/0	D'
Eg-m140	60	SSU57 (G)	13/0	D
Eg-m141	62	LSU1312 (C)	14/0	D'
Eg-m142	61	SSU723 (A)	10/2	D'
Eg-m143	71	SSU485 (G)	10/0	D'
Eg-m144	58	LSU2982 (G)	11/0	D'
Eg-m145	70	LSU2713 (U)	12/0	D'
Eg-m146	65	LSU1898 (U)	15/0	D
Eg-m147	61	LSU74 (G)	13/0	D'
Eg-m148	57	LSU2981 (A)	11/1	D'
* Target match indicates the number of canonical & G•U base-pairs /mismatches				

Table S13. Screening an *E. gracilis* BAC genomic DNA library for snoRNA gene clusters

Cluster	snoRNA(s) amplified by PCR	Present in library (based on PCR screening)?	Used in hybridization screening?	BAC isolated and characterized?
N/A	Eg-h1	yes	yes	yes
N/A	Eg-m34	yes	no	
1	**Eg-p1**	no		
2	**Eg-m2, Eg-m61, Eg-m1**	yes	yes	yes
3	Eg-p3, U14, Eg-m3	yes	yes	yes
3	Eg-m3	yes	no	
3	U14	yes	no	
5	Eg-m26	yes	no	
6	**Eg-m28**	no		
7	Eg-m88	no		
7	Eg-m11	no		
7.2	Eg-m37, Eg-m6, Eg- m88, Eg-m11	no		
7.3	Eg-m6, Eg-m88, Eg-m11, Eg-m37	no		
8.3	Eg-p5, Eg-m59.2, Eg-m23.5, Eg-p5	no		
10	Eg-m53, Eg-m21.3, Eg-m89, **Eg-m14.2, Eg-m14, Eg-m14.3**	no		
10	**Eg-m14.3**	yes	no	
10.1	Eg-m21.2, Eg-m89, Eg-m**14.2, Eg-m14, Eg-m14.4**, Eg-m53	no		
10.2	Eg-m21, Eg-m89, **Eg-m14.1**	no		
11-11.2	Eg-m31.2, **Eg-p10** & their isoforms	no		
11.3	**Eg-p10**, Eg-m31 & their isoforms	no		
11	Eg-m31	no		
12.5	Eg-m55.1, Eg-m72.1, Eg-m90	no		
12	**Eg-m38, Eg-m90, Eg-m72**	no		
12.2	Eg-m55, Eg-m72, **Eg-m38.1**	no		
12.3-12.6	Eg-m55.1, Eg-m72.2, **Eg-m90.1**	no		
13	**Eg-m25, Eg-m39.2**	yes	no	
13.1	**Eg-m25, Eg-m39.3**	yes	no	
13.2	**Eg-m39**, Eg-m93, Eg-m42.2, **Eg-m25**	yes	yes	no
14	**Eg-m65, Eg-m48**	yes	yes	no
14.1	**Eg-m48, Eg-m65.1**, Eg-m98	yes	no	
15	Eg-m43.1	no		
16, 16.1	**Eg-m8, Eg-m75.1**	yes	yes	no
17	Eg-m12, **Eg-m85**	no		
18	**Eg-m91**	yes	yes	no
18	Eg-m66	yes	no	
18	**Eg-m91**	no		
19	**Eg-m86**	no		
21	**Eg-m22**	no		

23	Eg-m56	yes	yes	yes
25	**Eg-m18**	yes	yes	yes
26	Eg-p6	yes	no	
26	**Eg-m54**, Eg-p6, Eg-m24	no		
26.1	Eg-p6, Eg-m24, **Eg-m54**	no		
27-27.2	**Eg-m16**, Eg-m79, Eg-m80, Eg-m81	yes	yes	no
27.3	**Eg-m16**, Eg-m80, Eg-m82	yes	no	
27	**Eg-m16**	no		
28	Eg-p11	no		
28	Eg-m70, **Eg-m94**	no		
28	Eg-m70, Eg-m94, Eg-p11, **Eg-m95**	no		
28	Eg-p11	no		
29	Eg-m77	no		
29	Eg-m50, Eg-m77, Eg-m50.2, Eg-m35	no		
31	**Eg-p9**	no		
31	**Eg-p9**	no		
32	**Eg-m13, Eg-m84**	yes	yes	no
The names of snoRNAs that target densely modified rRNA sites are indicated in bold.				

Table S14. Bioinformatic "clean-up" procedure for *E. gracilis* small RNA library Illumina MiSeq data

Clean up step	Number of sequence reads remaining
Remove 5′ & 3′ adaptors	1,876,362
Remove 2 highly abundant sequences	1,244,712
Cluster related sequences	69,173
Remove known RNA species	66,393
Extract sequences between 50 – 80 nt	29,230
snoRNA pattern 'hits'	193 (box C/D) & 895 (box AGA)

Table S15. Characteristics of box AGA snoRNAs identified in *Euglena gracilis*

Box AGA RNA	Size (nt)	Target site	Target match*	Distance to Ψ pocket (nt)	Basal Stem (nt)	Apical Stem (nt)	Stem-loop (nt)
Eg-p5	66	LSU3451	4/0 + 4/0	15	6	10; 0+1 nt bulge	8
Eg-p6	64	SSU105	5/0 + 8/1	15	5	11; 1 m.m.	7
Eg-p7	70	LSU3332	5/0 + 6/0	18	8	9; 1 m.m.	10-12
Eg-p8	60	LSU1365	4/0 + 6/0	14	6	11; 0+1 nt bulge	6
Eg-p9	69	LSU3503	5/0 + 5/0	15	6	15; two 0+1 nt bulges	7
Eg-p10	68	LSU2842	4/0 + 6/0	15	9; 1 nt bulge	11; 0+1 nt bulge	8
Eg-p11	72	SSU544	4/0 + 9/0	16	6	14; 1 m.m. 0+1 nt bulge	5
Eg-p12	66	LSU2904	4/0 + 4/0	13	7	10; 0+1 nt bulge	12
Eg-p13	67	LSU1568	4/0 + 4/0	15	9; 1 nt bulge	13; 0+2 nt bulge	7
Eg-p14	71	LSU3235	5/0 + 5/0	15	7	14-16; 2 m.m. 0+1 nt bulge	4 - 6
Eg-p15	72	LSU2752	5/0 + 5/0	15	8	12; 0+1 nt bulge & 0+4 nt bulge	4
Eg-p16	67	SSU1592	4/0 + 7/0	15	6	11; 2+3 nt bulge	10
Eg-p17	66	LSU3701	5/0 + 5/0	16	6	12; 1 m.m.	6
Eg-p18	67	SSU1393	5/0 + 6/0	15	7	11	8
Eg-p19	63	LSU3953	4/0 + 8/0	14	6	12; 1+0 nt bulge	6
Eg-p20	70	LSU2915	4/0 + 8/0	15	6	14; 1+0 nt bulge	8
Eg-p21	67	LSU2914	6/0 + 4/0	15	7	12; 1+0 nt bulge	7
Eg-p22	65	LSU1023	3/0 + 8/0	16	6	13; 1+2 nt bulge	4
Eg-p23	62	LSU1266	6/0 + 6/0	15	6	13; 1+2 nt bulge	3
Eg-p24	64	LSU280	4/0 + 9/0	16	6	10; 2+2 nt bulge	6
Eg-p25	64	LSU1235	4/0 + 9/0	15	7	11; 0+1 nt bulge	5
Eg-p26	66	LSU3697	6/0 + 6/1	15	6	15; 1 m.m. 0+2 nt bulge	4
Eg-p27	66	LSU3568	7/0 + 4/1	15	7	11	10
Eg-p28	65	LSU3042	4/0 + 8/2	16	8; 0+1 nt bulge	8; 1 m.m.	12

Eg-p29	66	SU1660	3/0 + 7/0	14	6	10; 2 m.m.	12
Eg-p30	68	SSU2065	4/0 + 5/0	17	7; 0+1 nt bulge	14; 0+1 nt bulge, 1 m.m.	5
Eg-p31	67	LSU2119	4/0 + 7/0	14	5	10; 1 m.m.	12
Eg-p32	64	LSU2361	7/0 + 6/2	15	6	10; 1+0 nt bulge	10
Eg-p33	70	LSU2943	5/0 + 5/0	15	8	10	14
Eg-p34	64	SSU2116	4/0 + 6/0	16	6	13; 0+1 nt bulge	4
Eg-p35	67	LSU3542	8/1 + 7/0	14	4	11	10
Eg-p36	68	SSU1068	3/0 + 5/0	15	6	7; 0+1 nt bulge	17
Eg-p37	63	LSU3204	4/1 + 6/1	17	8	10; 2 m.m.	5
Eg-p38	65	SSU2305	7/0 + 4/0	15	6	12; 0+1 nt bulge	5
Eg-p39	64	LSU1407	7/0 + 6/0	14	5	12; 1+0 & 0+1 nt bulges	7
Eg-p40	62	LSU567	4/0 + 6/1	15	8; 1+0 nt bulge	8	10
Eg-p41	65	LSU1692	5/1 + 6/1	16	8	11; 1 m.m.	7
Eg-p42	66	SSU89	4/0 + 6/0	15	7	12	7
Eg-p43	67	LSU2746	5/0 + 7/0	15	8	12; 2 0+1 nt bulges	7
Eg-p44	60	SSU1624	3/0 + 4/0	16	6	10; 2+1 nt bulge	5
Eg-p45	68	LSU2837	4/0 + 5/0	14	8; 2 m.m.	10; 0+1 nt bulge	15
Eg-p46	63	LSU1926	7/1 + 7/0	14	5	11; 2 1+1 nt bulges	7
Eg-p47	65	LSU480	7/2 + 6/0	15	6	12; 0+1 nt bulge	5
Eg-p48	69	LSU3591	5/0 + 5/1	16	6	13; 1+0 nt bulge	8

* Target match indicates the number of canonical & G•U base-pairs /mismatches. The two sets of numbers for each snoRNA represent the regions of base-pair interactions upstream and downstream of the uridine targeted for modification. See Figure 10 for further clarification.
m.m. = mismatch

Table S16. Monitoring growth of *E. gracilis* cultures after electroporation with fibrillarin antisense dsRNA. For visual examples of the cultures, see Figure S9

	Density of *Euglena* cultures (OD_{600})		
Treatment	**5 days post-treatment**	**7 days post-treatment**	**10 days post-treatment**
1	0.029	0.35	1.08
1-C	0.041	0.73	1.42
2	**0.018**	**0.20**	**0.37**
2-C	**0.034**	**0.61**	**1.33**
3	0.026	0.21	0.41
3-C	0.031	0.24	1.25
4	0.027	0.02	0.55
4-C	0.031	0.11	1.23
5	0.082	0.43	1.22
5-C	0.145	1.21	1.45
6	0.045	0.38	0.79
6-C	0.099	1.18	1.43
7	0.031	0.44	1.34
7-C	0.030	0.72	1.47
8	0.027	0.54	1.38
8-C	0.041	0.90	1.45

Treatment numbers correspond to the conditions indicated in Table 5. Control treatments where no dsRNA was added are indicated with a 'C'. Treatment 2 showed the greatest decrease in growth compared to the control (highlighted with a box and bold text).

Appendix 2: Supplementary Figures

Figure S1. Newly identified *Euglena gracilis* snoRNA sequences.

Box C/D snoRNAs

The $O^{2'}$-methylated rRNA nucleotide targeted by each snoRNA is indicated in square brackets and, when applicable, the species of the large subunit rRNA where the modification resides is shown in parentheses. The nucleotides predicted to pair to the rRNA are underlined and the predicted box elements are shown in the following colors: C box, D′ box, C′ box and D box (shown in 5′ to 3′ direction). When one or more isoforms of a snoRNA species has been identified, sequences are aligned and asterisks show positions of nucleotide identity. Isoforms identified through bioinformatic analysis of a ncRNA library are boxed in pink.

Newly identified box C/D snoRNA species and their isoforms identified in genomic DNA amplified by PCR

```
Eg-m71      [LSU1315 (5)]
Eg-m71      ATGTGACTTTTCGTTTGGATTCCCCGTGACGCTGATTCCGTTGCTTTACCGGCTGAGA
Eg-m71.1    ATGTCACTTTTCGTTTGGATTCCCCGTGACGCTGATTCCGTTGCTTTACCGGCTGAGA
            **** ************************************************

Eg-m72      [LSU1826 (6)]
Eg-m72      TGATGATTCCAGACATCTGTTGCCGAGCTTCTGCCCTGAGGAAAACTTTTCTCGCTACTACAACCAATATGAAG
Eg-m72.1    TGATGATTCCAGACATCTGTTGCCGAGCTTCTGCCCTGAGGAAAACTTCTCTCGCTACTACAACCAATATGAAG
Eg-m72.2    TGATGTTTCCAGACGTCTGTTGCCGAGCTTCTGCCCTGAGGAAAACTTCTCTCGCTACTACAACCAATATGAAG
Eg-m72.3    TGATGATTCCAGACATCTGTTGCCGAGATTCTGCCCTGAGGAAAACTTCTCTCGCTACTACAACCAATATGAAA
            ***** ******** ************ ******************** ***********************

Eg-m73
Eg-m73      GGATGGCATGCCAAACGTTCTCCGTT-AGCTGACTCCAGAAAGGGTGGCACTTCGTGCGCGCGGTGGGTTTCTGAGA
Eg-m73.1    GGATGGCCTGCCAAACGTTCTCCGTT-AGCTGACTCCAGAAAGGGTGGCATTTTGTGCGCGCGGTGGGTTTCTGAGG
Eg-m73.2    GGATGGCATGCCAAACGTTCTCCGTTCAGCTGACTCCAGAAAGGGTGGCACTTCGTGCGCGCGGTGGGTTTCTGAGA
            ******* ****************** ********************** ** **********************
```

```
Eg-m74      [LSU2605 (8)]
Eg-m74      TGATGCGGCGGCCCAGCCGCTCTGTTTCCCTGCACAGACTGATCTCTGCGATGCGAAGGTAACGCTCTGATG
Eg-m74.1    TGATGCGGCGGCCCAGCCGCTCTGTTTCCCTGCACAGACTGATCTCAGCGATGCGAAGGTAACGCTCTGATG
Eg-m74.2    TGATGCGGCGGCCCAGCCGCTCTGTTTCCCTGCACAGACTGATCTCTGCGATGCGAAGGTAACGCTCCGATG
            ********************************************** ******************** ****

Eg-m75      [LSU2979 (9)]
Eg-m75      ----TGATGGTTGATTTTCACTGCAGGGCTGACGTCCATTGATGTCCATGGACCCCAAACTGAGT
Eg-m75      GCGGTGATGGTTGATTTTCACTGCAGGGCTGACGTCCATTGATGTCCATGGACCCCAAACTGAGT
Eg-m75.1    ----TGATGGTTGATTTTCACTGCAGGGCTGACATCCATTGATGTCCATGGACCCCAAACTGAGT
                ***************************** ********************************

Eg-m76      [LSU206 (2)]
Eg-m76      TGATTAACAACAATACTGAGTGATATTTCGATGCAACCCGTGAGACGATGTGAACCATGATG
Eg-m76.1    TGATTA-CGAGAGTACTGAGTGATATTTCGATGCA---TGTGAGACGATGTGAGCCATGATG
            ****** * * * **********************    ************** ******

Eg-m77      [LSU3957 (13)]
            GGATGTGCTCATTAACCTCTGAACTCTGAAGGCCTGATGACAGCGCTGCGAGTTCTGATG

Eg-m78      [LSU3718 (10)]
Eg-m78      GGATGAAACCTGTAACCCGGATTCAGGCTCCGACGACGGGTGATGCGGAAAGAATTGCCGGCTGAGT
Eg-m78.1    GGATGAAACCTGTAACCCGGATTCAGGCTCCGACGACGGGTGATGCGGAAAGAGTTGCCGGCTCAGT
            ********************************************************* ********* ***

Eg-m79      [LSU631 (3)]
Eg-m79      TGATGATCCCCTTACTGCACACCACAATGAGCCTCGCGTTGATTCTTTGTCCGCCCTGATT
Eg-m79.1    TGATGATCCCCTTACTGCACACCAGAATGAGCCTCGCGTTGATTCTTTGTCCGCCCTGATT
            ************************ ************************************
```

```
Eg-m80      [LSU628 (3)]
Eg-m80      ----TGATGACCACCACCGTTGCACACCACCGAAATGAATTGTGATTCGTGTACGGAGTCTGGAGTCTGATG
Eg-m80      CTTGTGATGACCACCACCGTTGCACACCACCGAAATGAATTGTGATTCGTGTACGGAGTCTGGAGTCTGATG
                ************************************************************

Eg-m81      [LSU2610 (8)]
Eg-m81      ----TGATGCACATCATATTTCTCAACACCAGAGTTCACACTGTGATGGCACAATGATTCCCCTGGCTGATC
Eg-m81      GACATGATGCACATCATATTTCTCAACACCAGAGTTCACACTGTGATGGCACAATGATTCCCCTGGCTGATC
                ************************************************************

Eg-m82      [LSU2908 (8)]
            GGATGGATGTGAATCTCGTTGAACCGAACGCATTGTGAAGGAGATGGCTTGGATTGCTGACC

Eg-m83      [LSU1270 (5)]
            TGAGGATAACCCATGGATCCCAGCTACTGACTGTCAGTGATGCGGCATGAAGCTGAGT

Eg-m84      [LSU2870 (8)]
Eg-m84      TGATGGCATTTTCATGACACATCCAATCGGACGCTCATCGTGACGTGCACCCAACATGGCTGAGC
Eg-m84.1    TGATGGCATTTTCATGACACATCCAATCGGACGCTCATCATGACGTGCACCCAACATGGCTGAGC
            ************************************ ************************

Eg-m85      [LSU1655 (6)]
Eg-m85      TGATGAATGTTACATCACTTCGGCAGTCGGATGTCAGCGTGACGCAGTATGAGTTTCTGAGA
Eg-m85.1    TGATGAATGTTTCATCACTTCGGCAGTCGGATGTCAGCGTGACGCAGTATGTGTTTCTGAGA
            *********** ************************************** **********

Eg-m86      [LSU1553 (6)]
Eg-m86      ----GGATGAGTTTTCGTTGTTTCCCTGACCCTCTGATGCAACATCATGAGCCAGCATCTGAAC
Eg-m86      CACAGGATGAGTTTTCGTTGTTTCCCTGACCCTCTGATGCAACATCATGAGCCAGCATCTGAAC
                ************************************************************
```

```
Eg-m87      [SSU621]
Eg-m87       AGATGACATCCATGGTATAGCGTCCTCTGAGCATGCTGATACCACATCGACCGCTGAAT
Eg-m87.1     AGATGACATCCATGGTTTAGCGTCCTCTGAGCATGCTGATACCATATCGACCGCTGAAT
             ********************************************************

Eg-m88      [LSU1106 (5)]
            TGATGGTGTGATGCAACTCTTTACCACCTTTCGCCTGACTGCTTGGATGATGTGCTTTTTGCCTGACT

Eg-m89      [LSU2996 (9)]
Eg-m89      TGATGACATTCATGTGGTGAAGGTAGGCTGACCTTTTAGTGAAGCATTTCTTGAAACTCCTGATG
Eg-m89.1    TGATAACATTCATGTGGTGAAGGTAGGCTGACCTTTTAGTGAAGCATTTCTTGAAACTCCTGATG
            **** ************************** ******************************

Eg-m90      [LSU1824 (6)]
Eg-m90      ----TGATGACCC-TCTCTTCGTTGATGACCCTGATGCCAGATTACAACCAAGACCTGAGT
Eg-m90.1    ACGATGATGACCCCTCTCTTCGTTGATGACCCTGATGCCAGATTACAACCAAGACCTGAGT
                ********* ************************************************

Eg-m91      [LSU2744 (8)]
Eg-m91      TGATTAAACCAAGATGAAGACAGTGCCCAGCACCGATGGTTCTTTGAGTCATAGTTACTATGAGC
Eg-m91.1    TGATTAAATTGAAATGAAGACAGTGCCCAGCACCGATGGTTCTTTGAGTCATAGTTACTATGAGC
            ********   * ************************************************

Eg-m92      [LSU97 (1)]
Eg-m92      TGATGATCT--TTCTGCTGTTCCTTTGAATGAATGTCATCTGTGTTTATGACCTCTGTGCTGACT
Eg-m92.1    TGATGATCTCTTTCTGCTGTTCCTTTGAATGAAT----TCTGTGTTTATGACCTCTGTGCTGACT
Eg-m92.2    GGCTGATCTCTTTCTGCTGTTCCTTTGAATGAAT----TCTGTGTTTATGACCTCTGTGCTGACT
Eg-m92.3    -GATGATCT--TTCTGCTGTTCCTTTGAATGAAT----TCTGTGTTTATGACCTCTGTGCTGAC-
             * ******  ***********************    *************************

Eg-m93      [LSU577 (3)]
TGAGGACGATGCGATAACCCCTATTGCATTTTGATGCCTTTTGTTGCAGTTTGGAGTGGCGCATTAAACAAGAGGTCTTTGCATGATGACT
```

```
Eg-m94      [LSU1626 (6)]
Eg-m94      ----TGATGATTCCTCTCAGCGGGTTTCTGAGACGATGCTGAGGCTCCATAATTCTGGCTGATA
Eg-m94.1    ----TGATGATTCTTCTCAGCGGGTTTCTGAGACGATGCTGAGGCTCCATAATTCTGGCTGATA
Eg-m94.1    TGAGTGATGATTCTTCTCAGCGGGTTTCTGAGACGATGCTGAGGCTCCATAATTCTGGCTGATA
                ********* **************************************************

Eg-m95      [LSU1624 (6)]
            TGATGACCCCCTTCTCAGCGGGTTCCACTGATTAGTGATCGAGAGTGACAACCGAGCGACTGAGA

Eg-m96      [SSU702]
Eg-m96      ----TGATGTGCAGAACATGTTAGTATACGCACTGACATTGTGATTACACCGACAACCCCCTGATT
Eg-m96.1    GGCATGATGTG-AGAAGATGTCAGTATACGCACTGACACAGTGATTACCTCAGCAAACC-CTGAAA
                ******* **** **** ****************  ********  *  *** ** ****

Eg-m97      [LSU2009 (7)]
            TGATGACAAGTGCAATTTGCTGATTACTCTGAGGCTTCACAGCACTTGCCTGTCTGAT

Eg-m98      [SSU2123]
            TGATGATCACCAATGGCAGGGATATGATTGCCAATGATTCGGCTATCACAACGTGCATCTGAGG

Eg-m99      [SSU1681]
Eg-m99      ----TGATGACTGTATCCATGCACCACTCTGAAGTCATTGTGGAGGTGATTCCCAACGATCTGATG
Eg-m99.1    ----GGATGATTGCTTCCATGCACCACTCCGAAGTTGCTGTGGAGCTAGATCCTTCCAACCTGATT
Eg-m99.1    AAGAGGATGATTGCTTCCATGCACCACTCCGAAGTTGCTGTGGAGCTAGATCCTTCCAACCTGATT
Eg-m99.2    ----GGATGATTGCCTCCATGCACCACTCCGAAGTTGCTGTGGAGCTAGATCCTTCCAACCTGATT
Eg-m99.3    ----GGATGATTGCTTCCATGCACCACTCCGAAGTTGCTGTGGAGCTAGATCCTTC----CTGCAC
Eg-m99.4    ----TGATGACTGTATCCATGCACCACTCTGCAGTCATTGTGGAGGTGACTCCCAACGATCTGATG
                 ***** **  ************** * ***   ******* *   ***       ***
```

```
U14    [SSU83-95]
U14.3  GGCAATGATTGACACCGAACCTCGTTCACAGATCAATCCGAGCCACC-AGTGAGAACCCATTT-GGGGTGGCGCTGGCGGATGGC
U14.4  GGCAATGATTGACACCGAACCTCGTTCACAGATCAATCCGAGCCACC-AGTGAGAACCCATTT-GGGGTGGCGCAGGCGGATGGC
U14.5  GGCAATGATTGACACCGAACCTCGTTCACAGATCAATCCGAGCCACC-AGTGAGAACCCATTTTGGGGTGGCGCAGGCGGAAGGC
U14.6  GGCAATGATTGACACCGAACCTCGTTCACAGATCAATCCGAGCCACC-AGTGAGAACCCGTTT-GGGGTGGCGCAGGCGGAAGGC
U14.7  GGCAATGATTGACACCGAACCTCGTTCACAGATCAATCCGAGCCACC-AGTGAGAACCCATTT-GGGGTGGCGCAGGCGGAAGGC
U14.7  GGCAATGATTGACACCGAACCTCGTTCACAGATCAATCCGAGCCACC-AGTGAGAACCCATTT-GGGGTGGCGCAGGCGGAAGGC
U14.8  GGCAATGATTGACACCGAACCTCGTTCACAGATCAATCCGAGCCACC-AGTGAGAACCCATTT-GGGGTGGCGCAGGCGGAAGGC
U14.9  -----TGATTGACACTGAACCTCGTTCACAGATCAATCCGAGCCACC-AGTGAGAACCCATTT-GGGGTGGCGCAGGCGGAAGGC
U14.9  -----TGATTGACACTGAACCTCGTTCACAGATCAATCCGAGCCACC-AGTGAGAACCCATTT-GGGGTGGCGCAGGCGGAAGGC
U14.10 -----TGATTGACACTGAACCTCGTTCACAGATCAATCCGAGCCACC-AGTGAGAACCCATTT-GGGGTGGCACAGGCGGAAGGC
U14.11 -----TGATTGACACTGAACCTCGTTCACAGATCAATCCGAGCCACCCAGTGAGAACCCATTT-GGGGTGGCACAGGCGGAAGGC
U14.12 GGCAATGATTGACACCGAACCTCGTTCACAGATCAATCCGAGCCACCCAGTGAGAACCCATTT-GGGGTGGCACAGGTGGAAGGC
U14    GGCAATGATTGACACTGAACCTCGTTCACAGATCAATCCGAGCCACC-AGTGAGAACCCATTT-GGGGTGGCGGAGGCGGAAGGC
U14.1  GGCAATGATTGACACCGAACCTCGTTCACAGATCAATCCGAGCCACC-AGTGAGAACCCATTT-GGGGTGGCGCAGGCGGATGGC
U14.2  GGCAATGATTGACACCGAACCTCGTTCACAGATCAATCCGAGCCACCCAGTGAGAACCCATTT-GGGGTGGCGCAGGCGGATGGC
            ********** ******************************** *************** ********   ** *** ***

U14.3  CTTGGCCCGTCTGAG--
U14.4  CTTGGCCCGTCTGAG--
U14.5  CTTGGCCCGTCTGAG--
U14.6  CTTGGCCCGTCTGAGA-
U14.7  CTTGGCCCGTCTGAGA-
U14.7  CTTGGCCCGTCTGAG--
U14.8  CTTGGCCAGTCTGAGCC
U14.9  CTTGGCCAGTCTGAGCC
U14.9  CTTGGCCAGTCTGAGC-
U14.10 CCTGGCCCGTCTGAGCC
U14.11 CTTGGCCCGTCTGAGCC
U14.12 CTTGGCCCGTCTGAGCC
U14    CTTGGCCCGTCTGAGCC
U14.1  CCTGGCCCGTCTGAGCC
U14.2  CCTGGCCCGTCTGAGCC
       * ***** *******
```

Newly identified box C/D snoRNA species and their isoforms identified through bioinformatic analysis of a ncRNA library

```
Eg-m100       [LSU936 (5)]
Eg-m100       TGGATGATGAATCCTTCTTGGTCCGTGTACTGATGTTTGATGACCGTATTTCTGACTG
Eg-m100.1     --GATGATGAATCCTTCTTGGTCCGTGTACTGATGATTGATGACCGTATTTCTGACTG   3'RACE
Eg-m100.2     --GATGATGAATCCTTCTTGGTCCGTGTACTGATGTTTGATGACCGTATTTCTGATT-   3'RACE
                ********************************* ****************** *

Eg-m101       [LSU2853 (8)]
              TGGATGATGACTCCTTTGGATGACGTGCCGTGATCTGATTCTTGTTCCCTTGGTTGTCTGAC

Eg-m102       [LSU2806 (8)]
              CAGATGATGTTTTGTTCTCGTTGATCCACTGACGCCCTGCTGAGGCTACCATTTTCCTGACC

Eg-m103       [LSU3187 (9)]
              TTGATGATGAGCCCATTTTCTTTGCTGATCGCCAATATGCTGACAAATGTCCCAGCCAATCTGAGG

Eg-m104       [SSU1037]
Eg-m104       GCCATGATGCACT-GAACACTGAGTTTTCTGAAT--CTTGAAGTCAATCCTCTCCTGAG
Eg-m104.1     TTCATGATGCTTTCAAACACTGAGTTTTCTGATTGTCATGAAGTCAAAGCTTTCCTGA-
                ********  *  ****************** *  * *********  ** ******

Eg-m105       [LSU1502 (6)]
              AACGTGATGAAGTCTCAATCAGCACTGCTTCTGACATTTGATGATTCAAACAGCTGACA

Eg-m106       [LSU39 (1)] Eg-m106.2 D box [LSU2981 (9)]?
Eg-m106       CAAGTGATGTCATGCTTTGTGTCCTTCCCCTGACCACTGATGCCGCTCTGTCTGTAGCTGATT
Eg-m106.1     -CCGTGATGTCATGTTTTGTGTCCTTCCCCTGACCGATGATGTTGCTCTTTCTGCAGCTGATC
Eg-m106.2     GGCATGATGGCATGCTTTGTGTCCTTCCACTGACTGGTGATGTTGTTTTTTCTGCAGCTGATT
                  ***** **** ************* *****   *****  * * * **** *******
```

```
Eg-m107       [LSU3315 (9)]
Eg-m107       ACAAGGATGTCTGGATTAAGCACGTTTCGGCTGACGGCCTTGGGTGATTTCAGCTCCACCGCTGAGT
Eg-m107       ----GGATGTCTGGATTAAGCACGTTTCGGCTGACGGCCTTGGGTGATTTCAGCTCCACCGCTGA--
                  ************************************************************

Eg-m108       [LSU1039 (5)]
Eg-m108       GCTGTGATGC--GTTGTTCCATGCACCTGACACTGCGTGATGCGCTTTTCTGGTCTGAT
Eg-m108       GCTGTGATGC--GTTGTTCCATGCACCTGACACTGCGTGATGCGCTTTTCTGGTCTGA-
Eg-m108.1     GCTATGATGTTTGTTGTTCCATGCAACTGATATTG-GTGATGCGGTTTTCTGGTCTGA-
              *** *****   ************* **** * ** ******** *************

Eg-m109       [LSU1608 (6)]
Eg-m109       TGCAGGATGACACCATTTCATGGCCACCTCTGATTGCCATGCTGAGGCCCCTTTCTGCTATCCCAACTCTGAGG
Eg-m109.1     TGCAGGATGACACCATTTCATGGCCACCTCTGATTGCCATGCTGAGGCCCCTTTCTGCTATCCCAACGCTGAGG
              ************************************************************** ******

Eg-m110       [LSU1819 (6)]
Eg-m110       CGCCTGATGACCATTTGCTTTTGATT-CCTGATCATTCCAAACCAAGATCTGTTCTGATG-
Eg-m110.1     CGCCTGATGACTATTTGCTTTTGATT-CCTGATCATTCCAAACCAAGATCTGTTCTGATCG
Eg-m110.2     AGATTGATGACC-CGTGCTGTTGATTGCCTGATTGTCTGAAACCAAGATCTGTTCTGACT-
               *  *******    **** ****** ******  *   ******************

Eg-m111       [LSU2768 (8)]
              GCAGGGATGACCCCTTTGAGGCATTTGGCCTGACCGTTCCGCATGATTTCACCCGAAGCGGCTGAGC

Eg-m112       [LSU1081 (5)]
Eg-m112       GGGATGATGAACCATTTGCAAGCACTCGCTGACCACTGATGCCGAGTTTCATCTGAG-
Eg-m112.1     ATGATGATGAACCATTTGCAAGCACTCGCTGACCACTGATGCAGAGTTTCTTCTGATG
                ************************************* ******* *****

Eg-m113       [LSU2993 (9)]
              CAAATGATGCTGTGTTAGGTAGTGGACTGATCCGATTATGCTGCAGCATTTGTACGGACTGA
```

Eg-m114 **[LSU2005 (7)]**

TGGATGATGATTTGAAACTTGCCTGTTTGGCTGATGCGTTGATGCTTGCTTGTGTATTGCTGAC

Eg-m115 **[LSU3694 (10)]**

CGCATGATGTTATTCTCAGTCAGTATCTTCTGACCTCGGATGCTGTGCCGCTTTGGCTGA

Eg-m116 **[SSU2180]**

```
Eg-m116      TGAATGATGTTGTATCAACCTCCTGATTCTGAGTTCGCTTTGAGGAGCCTCTT---TTCTGAT-
Eg-m116.1    CATGTGATGCCTGA--AACCTCCTGATTCTGATT-CTTATTGATGAGCCTTTT--TCTCTGACT
Eg-m116.2    TTCATGATGTCCAA--AACCTCCTGATGCTGAA--CTCCTTGATGTGCTTTTT-CCTCCTGACC
Eg-m116.3    ----TGATGTCCAA--AACCTCCTGATGCTGAA--CTCCTTGATGTGCTTTTTTCCTCCTGA--
                 *****    *  *********** ****   *   **** * ** * **     ****
```

Eg-m117 **[LSU2776 (8)]**

ATGATGATGACAAGCATTTGGTTGACGAGTCTGAGCCAATCTGAAGCACAACGATGGTGTCTGATT

Eg-m118 **[LSU2909 (8)]**

TTCATGATGTAGCTTTCAAACTAGAGTCTCTGACGAGGCGTGATTACAGCATGCAGCTCTGATCT

Eg-m119 **[LSU3208 (9)]**

```
Eg-m119      CAAATGATGTGATCAACGTTGTGTTGTAACCTGACGTCTTTTTGATGACCATGCATCCCTGATT
Eg-m119.1    CAAATGATGTGATCAATGTTGTGTTGTAACCTGACGTCTTTTTGATGAACATGCATCCCTGAAT
             ********************************** ************************** *
```

Eg-m120 **[LSU3374 (9)]**

CCTTGGATGAGACACCTCTGGCTTGCTGACCTGACGTGACTGAAGGCATTAATGCTTTCTGAGG

Eg-m121 **[LSU3906 (12)] *binds in intergenic spacer region**

CCTGTGATGAGTTAAGGTCTCTGTTGCTGAACGTCATGACTGCGAACCCTGTACCGTCTGAT

Eg-m122 **[LSU2849 (8)]**

CAGATGATGCTCAATGTGTAGTCTGAGATTTAATGACTGCTGCCTGGCCCTGATGTCGCTGGTGGTTGGGCTGGATT

Eg-m123 **[LSU3688 (10)]**

TTTTTGATGAATTGCTTGTATCCGTCACTGCTGACCCATATGATTTCTCTTGCCCTGCTGACC

Eg-m124 **[LSU41 (1)]**

Eg-m124 ----------ATCGTGATGATTTCCTATGCTGTGTCCTTTCTGATCTTTTTGACTGATATCATTCTCTTCTGA

Eg-m124 TTGCTGCTGTGAATCGTGATGATTTCCTATGCTGTGTCCTTTCTGATCTTTTTGACTGATATCATTCTCTTCTGA

Eg-m125 **[LSU3367 (9)]**

Eg-m125 CAAGTGATGTATCTCAAGGCTTCGATACCACTGACTGAGATCTGAAGGCTATTTCCTCCCTGA--

Eg-m125.1 CAAGTGATGCATCTCAAGGCTTCGATATCACTGACCGAGATCTGAAGGCTATT-CTTCCCTGATG

********* ********** ******* ********** * ********

Eg-m126 **[LSU1683 (6)]**

TGCATGATGAGCATTTTTGGGCACCATTCTGATTTATTTTGATGTGCTTGCACTCTGAAC

Eg-m127 **[LSU1738 (6)]**

TGTGGGATGACTCCTGTGATGCTCTGCCACTGAGACTTAGGATGGAATGGACAGTTGTCTGACA

Eg-m128 **[LSU1121 (5)]**

CATATGATGTATGAATCAGAACGGTATATCTGAAAGGATGAAGTGTCTTGTGAGGCTGAC

Eg-m129 **[LSU2953 (8)]**

Eg-m129 CAGAGGATGTACACACCTCCCACTTTTGACGGATTTGATGCAACCCTGTTTCCACTGACT

Eg-m129.1 CAGAGGATGTGAACACCTCCCACCTTTGACGGATTTGATGCAACCCTGTTTCCACTGATT

Eg-m129.2 CAGAGGATGTAAACACCTCCCACCTTTGACGGATTTGATGCAACCCTGTTTCCACTGATT

Eg-m129.3 CAGAGGATGTAAACACCTCCCACTTTTGACGGATTTGATGCAACCCTGTTTCCACTGACT

********** ********** ********** *

Eg-m130 **[SSU1966]**

TTTATGATGACCAGCTCCATCTTGAAAAGACATGATGCTTGTTGATACACCTTCACTGCCTGATC

Eg-m131 **[SSU1736]**

TGTGTGATGGATGGCAATGTCTCACTCGTCTGACTCCTGCTGCTGTCACATGCTCTGACTGA

Eg-m132 **[LSU2685 (8)]**

CAGATGATGATAACATTTGAGCACTGGGTTCTGATGTTTTGATGTTGAATTGATTGAACACTGAT

Eg-m133 **[SSU110]**

CAAATGATGACCTCCAGATGGATGACTGCTTCTGAGATTCTTGAAGATACATCTATCTGAAC

Eg-m134 **[LSU1189 (5)]**

ACCGTGATGAAACACTCTCAGGCTACACACTGAAACCACTGATGCATTTCTGTAATACTGATT

Eg-m135 **[LSU1452 (5)**

```
Eg-m135      GCAATGATGTTTCATACATGCTCAGAATGTGATGACTGCATGATCC-ACTTCTTT-GCTGACG
Eg-m135.1    ACAATGATGTT-CA---ATGCTCAGAATGTGATAATTTCATGATCTTACCTCTTTTGCTGA--
              ********** **   ***************** * * *******  ** ***** ****
```

Eg-m136 **[LSU491 (3)]**

CGAATGATGACTTTGTTCAAACTGCAGTCTCTGATACTTGATGAACCAATTTCTCTGAAG

Eg-m137 **[LSU2146 (7)]**

```
Eg-m137      ---CCAAAGATGAAACACAGCCCGCGTGTCTGAAGGTTTTGATGCTTCCTTCCTGGCTGAGG
Eg-m137.1    GATGGACGGATGAAATACAGCCCGCGTGTCTGAA--TTTTGATGCTCTCCTCTCGGCTGATTGCGGGATTTGAAGAA
                  *  ******* ******************  **********  * **  ******
```

Eg-m138 **[LSU1888 (6)]**

```
Eg-m138      CCCATGATGAAAAATCTGTTCAGACCTTGCTGATGAGTTTCATCCAGTCCTCAAACTGATG
Eg-m138.1    CAAATGATGAAACATTTGCACAGATCAGAGTGATGAGTTTCATCCAGTCCTCAAACTGATG
             *  ********* ** **  **** *    ******************************
```

Eg-m139 **[LSU483 (3)]**

TATGTGATGACATTCAAAATCCCAAGCCATCTGAATGCAGTGATTGATGCATCTGGCTGAAG

Eg-m140 **[SSU57]**

TCTGTGATGACGGTCTGAAGCAGGTGATTCCTGATCAACTGCAGCGCTGAGACGTGTCTGA

Eg-m141 **[LSU1312 (5)]**

CAAATGATGTACCGTTTTTGCTTTACCGGATTCTGCCCTGTGATGATGCTCTCCCACTGACC

Eg-m142 **[SSU723]**

TCCATGATGAATTTTTTATCACAGACCTGTCCTGATTATCTGATGATCATTGCCACTGAAG

Eg-m143 **[SSU485]**

CATTTGATGACACCTGGCTCCCTCTACTGAGGTGATGAGGAGATTCTGCATCGAACCCATTTTCTCTGAAG

Eg-m144 **[LSU2982 (9)]**

```
Eg-m144      CACATGATGGTTACATTTCACTGCACTGACGTCGGCTGAAGA-GCGAGTTCTGCTGAGT----
Eg-m144.1    CATATGATGGTTTCATTTCACTGCACTGACGCCAGTTGATGACACTCTTCCCTCTGATTTAGC
             ** ********* ****************** * * *** **  *   * *  **** *
```

Eg-m145 **[LSU2713 (8)]**

ACACTGATGCAGAGCAAATGGCTGAATCTTCCGACGGCAGACTTGAGGTTCACTGATGAGATTCTGAGGG

Eg-m146 **[LSU1898 (6)]**

```
Eg-m146      TGGATGATTCAATTTCTATCGCAGACCAATGACGACTTATTGACAACCTGCTTCACTCTATGACT
Eg-m146.1    TGGATGATTCGATTTCTATCGCAGACCAATGACGACTTATTGACAACCTGCTTCACTCTATGAC-
             ********** ************************************************
```

Eg-m147 **[LSU74 (1)]**

ACGATGATGAGACCATAGACTCTGCGTATCTGACTGCTGGATGAAACACAATCTGCTGAAC

Eg-m148 **[LSU2981 (9)]**

CCCATGATGGTGAACTTCACTGCACTTGACCCAAGCTGATGAGCACGCCCCAACTGA

Isoforms of previously identified box C/D snoRNA species

New isoforms identified through PCR amplification of genomic DNA are boxed in yellow. Isoforms boxed in green were identified through analysis of snoRNA repeats from the BAC genomic DNA library. Isoforms boxed in pink were identified through bioinformatic analysis of a ncRNA library and some have extended 5′ or 3′ ends. In some alignments, previously identified snoRNA sequences were obtained biochemically and these RNAs were 5′ or 3′ truncated (Russell et al., 2006).

```
Eg-m1       [LSU3535 (9), LSU3521 (9)]
Eg_m1       ACATGATGACCCCATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTCCTTTTGTTCCCTGTTGGGCTGA--
Eg_m1.1     ACATGATGACCCCATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGACG
Eg_m1.2     ACATGATGACCCCGTTCTAACCCAGCTCGATGACGCCGGTGTGATTATTCTTTTTGTTCCCTGTTGGGCTGACG
Eg_m1.3     ACAGGATGACCCCATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGGCG
Eg_m1.4     ACAGGATGACCCCATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGACG
Eg_m1.5     ACATGATGACCCCATTCTAACCCAGCTCGATGACGCTGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGACG
Eg_m1.6     ACATGATGACCCCATTCTAACCCAGCTCGATGACGCCGGTGTGGTTATTCCTTTTGTTCCCTGTTGGGCTGACG
Eg_m1.7     ACATGATGACCCCATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTCTTTTTGTTCCCTGTTGGGCTGACG
Eg_m1.8     ACATGATGACCCCATTCTAACCCAGCTCGATGACGCCGGTGTAATTATTCCTTTTGTTCCCTGTTGGGCTGACG
            *** ********* ********************** *****  *****  ******************** **

Eg-m2       [LSU1856 (6)]
Eg_m2       ATGATGAGCCCC-CGCTTGCCACTCCTCTTTGATGCCACAGATGAGATTCGTAGCGCGCCTGA--
Eg_m2.1     ATGATGAGCCCC-CGCTTGCCACTCCTCTTTGATGCCGCAAATGAGATTCGTAGCGCGCCTGATA
Eg_m2.2     ATGATGAGCCCC-CGCTTGCCACTCCTCTCTGATGCCGCAAATGAGATTCGTAGCGCGCCTGATA
Eg_m2.3     ATGATGAGCCCC-CGCTTGCCACTCCTCTTTGATGCCGCAGATGAGATTCGTAGCGCGCCTGATA
Eg_m2.4     ATGATGAGCCCC-TGCTTGCCACTCCTCTTTGATGCCACAGATGAGATTCGTAGCGCGCCTGATA
Eg_m2.5     ATGATGAGCCCCCCGCTTGCCACTCCTCTTTGATGCCGCAAATGAGATCCGTAGCGCGCCTGATA
Eg_m2.6     ATGATAAGCCCC-TGCTTGCCACTCCTCTTTGATGCCGCAAATGAGATTCGTAGCGCGCCTGATA
Eg_m2.7     ATGATGAGCCCC-TGCTTGCCACTCCTCTTTGATGCCGCAAATGAGATTCGTAGCGCGCCTGATA
Eg-m2.8     ATGATGAGCCCC-TGCTTGCCACTTCTCTTTGATGCCGCAAATGAGATTCGTAGCGCGCCTGATA
            ***** ******  ********** **** * ***** ** ******* ***************
```

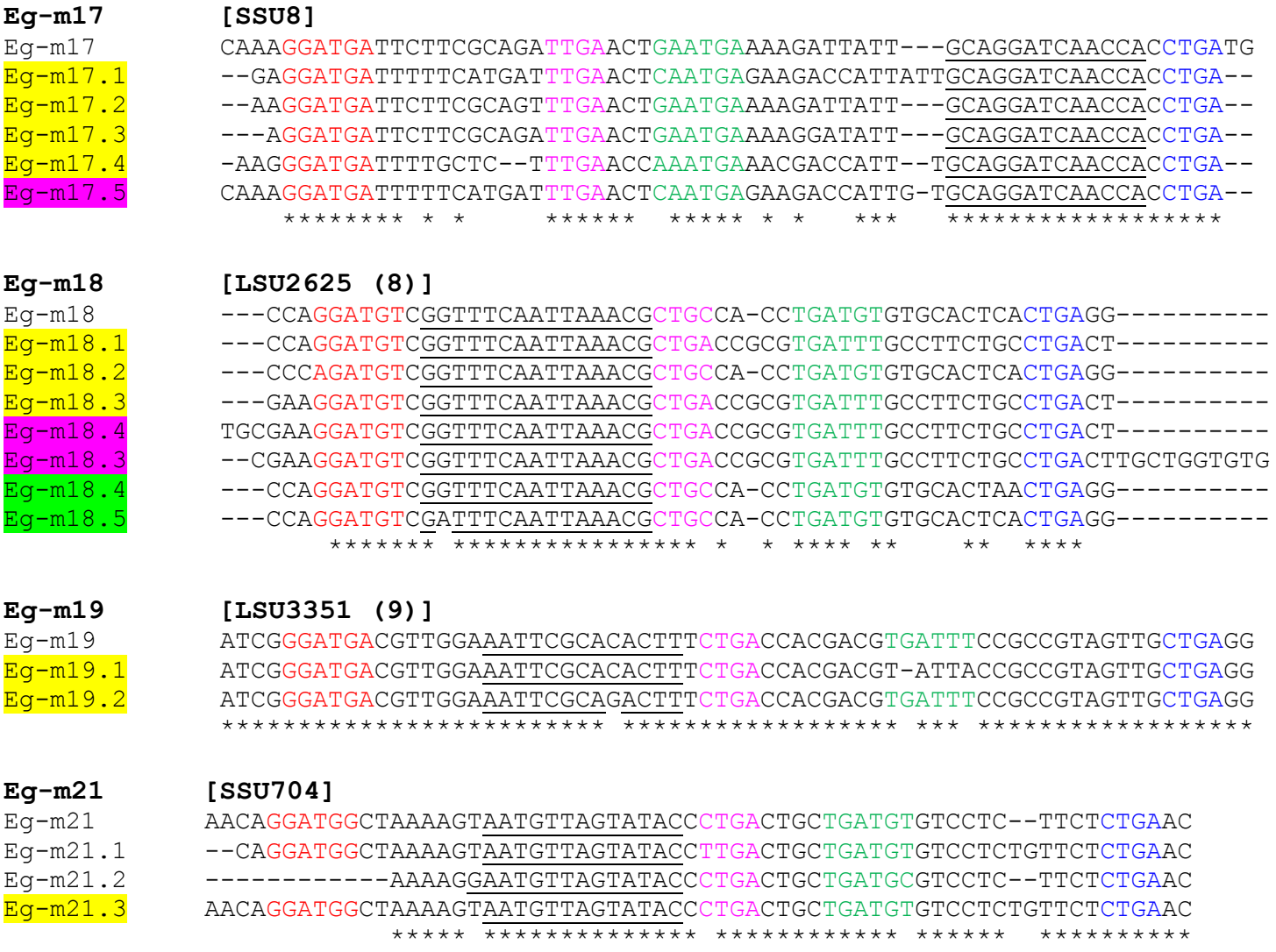
```
Eg-m3       [LSU2129 (7)]
Eg-m3       -TCAGGATGCAGTTATTCATCTCCCCTGAATTCTGATGGTTGCATCAGGAGGATCTGA--
Eg-m3       GTCAGGATGCAGTTATTCATCTCCCCTGAATTCTGATGGTTGCATCAGGAGGATCTGAAT
             ************************************************************

Eg-m4          [SSU682]
Eg-m4          -ACATGATGCGATACATGGAATTACCGCCTGACACACTTGCTTGATGTCATTTGCACCCTGATT
Eg-m4          GACATGATGCGATACATGGAATTACCGCCTGACACACTTGCTTGATGTCATTTGCACCCTGATT
Eg-m4.1        GACATGATGCGATATTTGGAATTACCGCCTGACACACTTGCTTGATGTCATTTGCACCCTGATT
                *************  ************************************************

Eg-m8          [LSU1201 (5)]
Eg-m8          --CGTGATGCTAATTACCCTATGCCGAATCTGAGATGCTTTGATGCGGTTTTCTCATCGTGCCTGACG
Eg-m8          -CCGTGATGCTAATTACCCTATGCCGAATCTGAGATGCTTTGATGCGGTTTTCTCATCGTGCCTGACG
Eg-m8          ----TGATGCTAATTACCCTATGCCGAATCTGAGATGCTTTGATGCGGTTTTCTCATCGTGCCTGACG
Eg-m8          GCCGTGATGCTAATTACCCTATGCCGAATCTGAGATGCTTTGATGCGGTTTTCTCATCGTGCCTGAC-
                   ************************************************************

Eg-m14      [LSU1204 (5), LSU1217 (5)] [Eg-m14.3 & 14.4 LSU1204 (5), LSU1218 (5)]
Eg-m14         GGATGAAACCACACATCACCCTTATGCCTCTGAGAA--AATGATGTGAACCGATTGGCCTTGCTGAGA
Eg-m14.1       TGATGAAAGTTCACATCCCCCT-ATGCCTCTGAG----AATGATGTGGAATGATTGGCCTTGCTGACA
Eg-m14.2       TGATGAAAGTTCACATCACCCTTATGCCTCTGAG----AATGATGTGGAATGATTGGCCTTGCTGACA
Eg-m14.3       GGATGAGAAAAC-CTTACCCCT-ATGCCCCTGACAACAGATGACGGGAAGCGATTGGCCTA-CTGAGG
Eg-m14.4       GGATGAGAAAAT-CTTACCCCT-ATGCCCCTGACAACAGATGACGGGAAGCGATTGGCCTA-CTGAGG
                ***** *     * *  **** ***** ****      **** * * *  *********  ****

Eg-m15         [SSU28]
Eg-m15         CACATGATGTTTGATTGAACAAAGCATATGGCGGAGTTTCTCGATGAGGAGCACTTTTTCGCTGAGG
Eg-m15.1       CGCATGATGTTTGATTGAACAAAGCATATGGCGGAGTTTCTCGATGAGGAGCACTTTTTCGCTGAGG
               * ************************************************************
```


Eg-m17 [SSU8]
Eg-m17 CAAAGGATGATTCTTCGCAGATTGAACTGAATGAAAAGATTATT---GCAGGATCAACCACCTGATG
Eg-m17.1 --GAGGATGATTTTTCATGATTTGAACTCAATGAGAAGACCATTATTGCAGGATCAACCACCTGA--
Eg-m17.2 --AAGGATGATTCTTCGCAGTTTGAACTGAATGAAAAGATTATT---GCAGGATCAACCACCTGA--
Eg-m17.3 ---AGGATGATTCTTCGCAGATTGAACTGAATGAAAAGGATATT---GCAGGATCAACCACCTGA--
Eg-m17.4 -AAGGGATGATTTTGCTC--TTTGAACCAAATGAAACGACCATT--TGCAGGATCAACCACCTGA--
Eg-m17.5 CAAAGGATGATTTTTCATGATTTGAACTCAATGAGAAGACCATTG-TGCAGGATCAACCACCTGA--
Eg-m18 [LSU2625 (8)]
Eg-m18 ---CCAGGATGTCGGTTTCAATTAAACGCTGCCA-CCTGATGTGTGCACTCACTGAGG----------
Eg-m18.1 ---CCAGGATGTCGGTTTCAATTAAACGCTGACCGCGTGATTTGCCTTCTGCCTGACT----------
Eg-m18.2 ---CCCAGATGTCGGTTTCAATTAAACGCTGCCA-CCTGATGTGTGCACTCACTGAGG----------
Eg-m18.3 ---GAAGGATGTCGGTTTCAATTAAACGCTGACCGCGTGATTTGCCTTCTGCCTGACT----------
Eg-m18.4 TGCGAAGGATGTCGGTTTCAATTAAACGCTGACCGCGTGATTTGCCTTCTGCCTGACT----------
Eg-m18.3 --CGAAGGATGTCGGTTTCAATTAAACGCTGACCGCGTGATTTGCCTTCTGCCTGACTTGCTGGTGTG
Eg-m18.4 ---CCAGGATGTCGGTTTCAATTAAACGCTGCCA-CCTGATGTGTGCACTAACTGAGG----------
Eg-m18.5 ---CCAGGATGTCGATTTCAATTAAACGCTGCCA-CCTGATGTGTGCACTCACTGAGG----------
Eg-m19 [LSU3351 (9)]
Eg-m19 ATCGGGATGACGTTGGAAATTCGCACACTTTCTGACCACGACGTGATTTCCGCCGTAGTTGCTGAGG
Eg-m19.1 ATCGGGATGACGTTGGAAATTCGCACACTTTCTGACCACGACGT-ATTACCGCCGTAGTTGCTGAGG
Eg-m19.2 ATCGGGATGACGTTGGAAATTCGCAGACTTTCTGACCACGACGTGATTTCCGCCGTAGTTGCTGAGG
Eg-m21 [SSU704]
Eg-m21 AACAGGATGGCTAAAAGTAATGTTAGTATACCCTGACTGCTGATGTGTCCTC--TTCTCTGAAC
Eg-m21.1 --CAGGATGGCTAAAAGTAATGTTAGTATACCTTGACTGCTGATGTGTCCTCTGTTCTCTGAAC
Eg-m21.2 ------------AAAAGGAATGTTAGTATACCCTGACTGCTGATGCGTCCTC--TTCTCTGAAC
Eg-m21.3 AACAGGATGGCTAAAAGTAATGTTAGTATACCCTGACTGCTGATGTGTCCTCTGTTCTCTGAAC

```
Eg-m23          [LSU1504 (6)]
Eg-m23          TGATGCGAACTATTTCATCACCAGCACTGTCTGCTGCCTGATGATGCTCCAACCGTTATTTCTGAGA
Eg-m23.1        TGATGCAAACTATTCCATCACCAGCACTGTCTGCTGCCTGATGATGCTCCAACCGTTATTTCTGAGA
Eg-m23.2        TGATCCAAACTATTCCATCACCAGCACTGTCTGCTGCCTGATGATGCTCCGACCATTATTTCTGAGA
Eg-m23.3        TGATGCAAACTATTCCATCACCAGCACTGTCTGCTGCCTGATGATGCTCCGACCATTATTTCTGAGA
Eg-m23.4        TGATGCAAACTATTTCATCACCAGCACTGTCTGCTGCCTGATGATGCTCCAACCATTTTTTCTGCGA
Eg-m23.5        TGATGCAAACCATTCCATCACCAGCACTGTCTGCTGCCTGATGATGCTCCGACCATTATTTCTGAGA
                **** * *** *** ************************************ *** ** ****** **

Eg-m26          [SSU533]
Eg-m26          -----GGATGATGCGTTCAATTTGCGACCGACTCTGTGACGTGACCC-ATGCCTGCTTTT---CTGACC-
Eg-m26          ----CGGATGATGCGTTCAATTTGCGACCGACTCTGTGACGTGACCC-ATGCCTGCTTTT---CTGACC-
Eg-m26          GCACCGGATGATGCGTTCAATTTGCGACCGACTCTGTGACGTGACCC-ATGCCTGCTTTT---CTGACCG
Eg-m26.1        -------ATGATGCGTTCAATTTGCGGCCGACTCTGTGACGTGACCC-ATGCCTGCTTTT---CTGA---
Eg-m26.2        -----GGATGATGTGTTCAATTTGCGGCCGACTCTGTGACGCGATTTTATGCCTGCTTGG---CTGACT-
Eg-m26.2        ----TGGATGATGTGTTCAATTTGCGGCCGACTCTGTGACGCGATTTTATGCCTGCTTGG---CTGAC--
Eg-m26.3        -----AGATGATGCGTTCAATTTGCGGCCGACTCTGTGACGTGATTTTAAGCCTGCTTTTTTCCTGATC-
Eg-m26.4        -----AGATGATGCGTTCAATTTGCGGCCGACTCTGTGACGTGATTTTATGCCTGCTTTTTT-CTGACC-
                      ****** ************ ************** **    * ********     ****

Eg-m27          [SSU1539]
Eg-m27          ----GGATGCTCGTTCGCATAAGTTTCAGCCACCGACATTGCATGAGTGACGCCGTTTGGCACTGAGG
Eg-m27.1        -----GATGCTCGTTCGCATAAGTTTCAGCCACCGACATTGCATGAGTGACGCCGTTTGGCGCTGAGG
Eg-m27.2        ----GGATGCTTGCTGCAAAATGTTTCAGCCACTGACACCCATGAGTGATACCGTCTAGCACTGATG
Eg-m27.3        CCCAGGATGCTCGTTTGCATAAGTTTCAGCCACCGACATTGCATGAGTGACGCCGTTTGGCGCT----
                     ****** * *   * * *********** ****   *********  **** * *  * **

Eg-m28          [LSU1573 (6)]
Eg-m28          CAATGATGTGTTTCAATTCAACATCCTTTGTGGACTGAGAAGGCGTGATGCAACCCCTGTGGGAACTGATC
Eg-m28.1        CAATGATGTGTTTCAATTCAACATCCTTTGTG-ATTGTGAAGGCGTGATGCAACCCCTGTGGGAACTGATC
                ******************************* * ** ********************************
```

```
Eg-m30      [SSU1536]
Eg-m30      CTGGTGATGTCATCCTCATAGTTTCAGCCTTGCCTGAAGGCCTGATGCGGTCACTGGTC-----
Eg-m30.1    CTGGTGATGTCATCCTCATAGTTTCAGCCTTGCCTGAAGGCCTGATGCGGTCACTGGCCTGATA
            ********************************************************** *

Eg-m31      [SSU186]
Eg-m31      ----TGATGCCTGAGGAACGTGTATTACCTGATGCACCTGATGAACAAATCTGCCTGA--
Eg-m31      CCCATGATGCCTGAGGAACGTGTATTACCTGATGCACCTGATGAACAAATCTGCCTGAGG
Eg-m31.1    -----GATGCCTGAGGAACATGTATTACCTGATGCAACTGATGAACAAATCTGCCTGAGG
Eg-m31.2    ----TGATGCCTGAGGAACATGTATTACCTGATGCACCTGATGAACAAATCTGCCTGAGG
Eg-m31.3    ----TGATGCCTGAGGAACGTGTATTACCTGATGAACCTGATGAGCAAATTTGTCTGAGG
Eg-m31.4    ----TGATGCCTGAGGAACGTGTATTGCCTGATGAACCTGATGAGCAAATTTGTCTGAGG
Eg-m31.5    ----TGATGCCTGAGGAACGTGTATTACCTGATGCAACTGATGAACAAATCTGCCTGAGG
                 ************** ****** ******* * ******* ***** ** *****

Eg-m35      [SSU40]
Eg-m35      GCTTGATGAGCTGTCCTTTAGCCCTTGACCTGAGCGCATGACGAGCCTCAGCCTGATC
Eg-m35.1    GCTTGATGAGCTGTCCTTTAGCCCTTGACCTGAGCGCATGATGAGCATCAGCCTGACC
Eg-m35.2    GCTTGATGAGCTGTCTTTTAGCCCTTGACCTGAGCGCATGACGAGCCTGAGCCTGATC
            *************** ************************* **** * ******* *

Eg-m36      [SSU1641, SSU1625]
Eg-m36      ------------------AATCTGTCAATCCCCTGAACCGGATGATGAACCGTCCTGACCTCTGAC-
Eg-m36      AGCGGAGGATGCAGCGGTAATCTGTCAATCCCCTGAACCGGATGATGAACCGTCCTGACCTCTGACA
Eg-m36.1    --TGGATGATGCTCGTATTATCTGTCAATCCCCTGATCTGGATGATGAATCGTCCTGACCTCTGAC-
               *** *****     * ***************** * ********** ***************

Eg-m38      [LSU1822 (6)
Eg-m38      -CGATGATGACCCCTCTCTTCGTTGATGACCCTGATGCCAGATTACAACCAAGATCGCTGAGT
Eg-m38      ACGATGATGACCCCTCTCTTCGTTGATGACCCTGATGCCAGATTACAACCAAGATCGCTGAGT
```

```
Eg-m39        [LSU2920 (8)]
Eg-m39        GGGATGAGGACCCATCAATTTGCACAGTATCTCCGATGGATATGTGACGACCATTTTGTTTGCTGATC
Eg-m39.1      GGGATGAGGACCCATCAATTTGCACAGTATCTCCGATGGATATGTGACGACCGTTTTGTTTGCTGATC
Eg-m39.2      GGGATGAGGACCCATCAATTTGCACAGTATCTCTGATGGATACACGACGACCGTTCTGTTTGCTGATT
Eg-m39.3      GAGATGAGGACCCATCAATTTGCACAGTATCTCCGATGGATATGTGACGATCGTTTTGTTTGCTGATC
              * ******************************** ********   ***** * ** ***********

Eg-m42        [SSU407]
Eg-m42        ---ATGATGACGCATTAATTTCTGTGAATCGCCGTGTGCGCTGAGAATTGATAGGTCAGTCTGAGA
Eg-m42.1      ---ATGATGACTCATTAATTTCTGTGAATCGCCATGTGCGCTGAGAATTGATAGGTCAGTCTGA--
Eg-m42.1      TGCATGATGACTCATTAATTTCTGTGAATCGCCATGTGCGCTGAGAATTGATAGGTCAGTCTGAGA
Eg-m42.2      ---ATGATGACTCATTAATTTCTGTGAATCGTCGTGTGCGCTGACAATTGATAGGTCAGTCTGAGA
                 ******** ******************* * ********** ******************

Eg-m43        [LSU949 (5)]
Eg-m43        --------------TGCTCCTTAGACTGCTGATTCCTGTGATGCAGTTCCGTTGAGTCTGACT
Eg-m43.1      CGATGACCAACCTTTGCTCCTTAGACTGCTGATTCCTGTGATGCAGTTCCGTCGAGTCTGACT
                            ************************************* **********

Eg-m48         [LSU1649 (6)]
Eg-m48         ---TGATGATGGACAGATCAAAGCAGGTGAGTTCTGAATCGGATGATGGCAGTAACCCTTGGCTGACG---------
Eg-m48         --CTGATGATGGACAGATCAAAGCAGGTGAGTTCTGAATCGGATGATGGCAGTAACCCTTGGCTGA-----------
Eg-m48         GGCTGATGATGGACAGATCAAAGCAGGTGAGTTCTGAATCGGATGATGGCAGTAACCCTTGGCTGACGCCGGGTCCC
Eg-m48.1       ---TGATGATGACCA--TCAAAGCAGGTGAGTTCTGAATCGAATGATGGCAGTAACTCTTGGCTGACG---------
Eg-m48.1       GGCTGATGATGACCA--TCAAAGCAGGTGAGTTCTGAATCGAATGATGGCAGTAACTCTTGGCTGACGCCGGGTCCC
                  ********  **  ************************ ************** ***********

Eg-m50         [SSU38]
Eg-m50         ---GTGATGA-GT----CTTAGCCCTTGAACTCTGAAGGCCTGACGAAACATCCCGATTGACCTGATG
Eg-m50.1       ---GTGATGA-CT----CTTAGCCCTTGAACTCCGAAGGCCTGACGAAACATCCCGATTGGCCTGATG
Eg-m50.2       ---CTGATGATCTCTGTTTTAGCCCTTGAACTCCGAAGGCCTGACCAATCAGTGCAGTGCTGCTGATG
Eg-m50.2       TTGCTGATGATCTCTGTTTTAGCCCTTGAACTCCGAAGGCCTGACCAATCAGTGCAGTGCTGCTGATG
                   ******  ***    ************** ************** **   *  *    ******
```

```
Eg-m52        [LSU1891 (6)]
Eg-m52        --TATGATGACTGTTGTATCCTGATTGCTGATGCTATCACACTCCAGTCCTTCTGACC
Eg-m52.1      TGCATGATGATTGTTGAATCCTGATTGCTGAGGCTATCACACTCCAGTCCTTCTGACC
                 ******* ***** ************** *************************

Eg-m54         [LSU1185 (5)]
Eg-m54           --GATGATGAACACATGGCTACACAATTGCTGATCTCCTGCTTTCAAATCGTTTGCTGATT
Eg-m54.1         CGAATGATGAGTCTATGGCTACACAATTGCTGACCTCGTGATTTCAAATCCTTGTCTGATC
                    *******    ******************* *** ** ********* **  *****

Eg-m55        [LSU935 (5)]
Eg-m55        -CGATGATGATCCATCTCTCCTTGAGTCCGTGTTGCTGACTGGTGACGACCCGATCTTCTGACG
Eg-m55        ACGATGATGATCCATCTCTCCTTGAGTCCGTGTTGCTGACTGGTGACGACCCGATCTTCTGACG
Eg-m55.1      -CGATGATGATCCATCTCTCTTTGAGTCCGTGTTGCTGACTGGTGATGACCCGATCTTCTGACG
Eg-m55.2      -CGATGATGATCCATCTCTCCTTGAGTCCGTGTTGCTGACTGGTGACGACCCCATCTTCTGACG
Eg-m55.3      -CGATGATGATCCATCTCTCCTTGAGTCCGTGTTGCTGACTGGTGACGACCCCATCTCCTGACG
Eg-m55.4      GCACTGATGATCTGTCTTTCCTTGAGTCCGTGTTGCTGACTGGTGACGATCTTGCCTCCTGA--
               *  ********  *** ** ************************* ** *    ** ****

Eg-m56         [LSU1324 (5)]
Egm56          GGGATGACGACATTGTTACTCTCATGCGTTGCTGAGGCCCACTGATGTGGTATCGCCAGCTGACC
Egm56.1        GGGATGACGACAATGTTACTCTCATGCGTTGCTGAGGCCCACTGATGTGGTATCACCAGCTGACC
Egm56.2        GGGATGACGACATTGTTACTCTCATGCGGTGCTGAGGCCCACTGATGTGGTATCGCCAGCTGACC
Egm56.3        GGGATGACGACATTGTTACTCTCATGCGTTGCTGAGGCCCACTGATGTGGTATCTCCAGCTGACC
Egm56.4        GGGATGACGACATTGTTACTCTCATGCGTTGCTGAAGCCCACTGATGTGGTATCGCCAGCTGACC
               ************ *************** ****** ****************** **********

Eg-m57         [LSU2602 (8)]
Eg-m57         ----------------ATCCTTGCACAGCAATCTGAATGGTTGATGTTGGTCCTTTTTCTGAAG
Eg-m57.1       ACTGTGATGATGTTTAATCCTTGCACAGCAATCTGAATGGTTGATGTTGTTTCGTTTTCTGA--
                                ********************************* * * ********
```

```
Eg-m59        [SSU2041]
Eg-m59        -GCAGGATGATGTGTTATCC--------CCAGTATTTGAGCACTGATAGCAAATCTTGTTTCGCCTGATC-
Eg-m59        GGCAGGATGATGTGTTATCC--------CCAGTATTTGAGCACTGATAGCAAATCTTGTTTCGCCTGATCT
Eg-m59.1      -GCAGGATGATGTGTTATCC--------CCAGTATTTGAGCACTGACAGCCAATCTTGTTTCGCCTGATC-
Eg-m59.2      -GCAGGATGATGTGTTATCC--------CCAGTATTTGAGCACTGATAGCCAATCTTGTTTCGCCTGATC-
Eg-m59.3      -GCAGGATGATGTGTTATCCAGTGTTATCCAGTATTTGAGCACTGATAGCAAATCTTGTTTCGCCTGATC-
               ******************         ****************** *** ******************

Eg-m61     [LSU3546 (9)]
Eg-m61     -CCGTGATGACACCCTTTCACGACGGTCTTATGACAGCGCAGTGATTCTGGTCTGTTCTGAC
Eg-m61     ACCGTGATGACACCCTTTCACGACGGTCTTATGACAGCGCAGTGATTCTGGTCTGTTCTGAC

Eg-m62     [LSU2769 (8)]
Eg-m62     --GGGCTGATGAT-CTGTGCGAGGCATTTGCCTGATAAAGAATGATGC----TGACCGTTTGACCTGAT---------
Eg-m62.1   --TTGCTGATGAG-CTGTACGAGGCATTTGCCTGACGAA-AATGATGCCAACTCACCGCTTGGCCTGATAGTGGGTGA
Eg-m62.2   GCTTGATGATGACGATGTGCGAGGCATTTGTCTGACGGA-CCTGAGGATATTTCACCAATTGATCTGATG--------
                ******   *** ********** ****   *   *** *  *  * ***  ***  *****

Eg-m65     [LSU1647 (6)]
Eg-m65     TGAGGACACCAATCGGTGCAGGTGAGTTGGATGACCACTGATTATGACCGATTTGTCCCGCTGAGG
Eg-m65.1   TGAGGACACCAATCGGTGCAGGTGAGTTGGATGACCACTGATTATTACCGATTTGTCCCGCTGAGG
           ******************************************** ******************

Eg-m66     [LSU594 (3)]
Eg-m66     ----TGGGATGATCCAGCTGGCTGATGGATTGCTGATTATTACA-TTTAACCCTCTTTCTGAGC
Eg-m66.1   GGCATGAGCTGACCCACCTCACTGATGGATTGCTGATTTTAGCATTTTAACCCTCTTTCTGAGC
               ** * *** *** **  ***************** *  ** ******************

Eg-m68     [LSU537 (3)]
Eg-m68     CCATGATGGATGGAAGGTTCCAGGTCAAATGAATTGCCTGATTTGCTTTATTCCGTTCCTGCTGAGT
Eg-m68.1   CCATGATGGATGGAAGGTTCCAGGTCAAATGAATTGCTTGATTTGCTTTATTCCGTTCTTGCTGAGT
           ************************************ ******************** ********
```

```
Eg-m69      [LSU1883 (6)]
Eg-m69      CGCGGATGACACCCAGACTGTCCTCAACCGCCTCTGATGTCACTGACTGACGCCACCTGTCTGAGC
Eg-m69.1    CGCGGATGACACCCAGACTGTCCTCAACCGCATCTGATGCCACTGACTGACGCCACCTGTCTGAGC
            ******************************** ******* *************************

Eg-m70      [SSU1549]
Eg-m70      ACATGATGCGCTGATTGTTCAATTCCTTTACCTGCACCGTGATGAACATTCCCAAGCTGATGAA
Eg-m70.1    ACATGATGTGCTGATTGTTCAATTCCTTTACCTGCACCGTGATGAACATTCCCAA-CTGATGGG
            ******** ******************************************** ******
```

Box AGA snoRNAs

The pseudouridylated rRNA nucleotide targeted by each snoRNA is indicated in square brackets and, when applicable, the species of the large subunit rRNA where the modification site resides is shown in parentheses. The nucleotides that pair to the rRNA are underlined and the predicted AGA box elements are shown in red. When isoforms have been identified, sequences are aligned and asterisks show nucleotide identity. Isoforms where RNA 3′ ends have been mapped by 3′ RACE techniques are indicated. Isoforms indicated with green highlighting were identified in the BAC genomic DNA library. Ellipses at the 3′ end indicate high-throughput sequence reads with extended 3′ ends.

Newly identified box AGA snoRNA species and their isoforms identified in genomic DNA amplified by PCR

```
Eg-p2       [LSU68 (5.8S)]
Eg_p2       ATGCCACTGCGTATACGCCGCTGCCCAGCCTTGGCAGCATGTCGTCCGAAAGGCATGGCAGACGC
Eg_p2.1     ATGTCACTGCGTATACGCCGCTGCCCAGCCTTGGCAGCATGTCGTCCGAAAGGCATGGCAGACGC
Eg_p2.2     ATGTCACTGCGTATACGCCGCTGCCCAGCCTTGGCAGCATGTCGTCCAAAAGGCATGACAGACGC
Eg_p2.3     ATGTCACTGCGTATACGCCGCTGCCCAGCCTTGGCAGCATGTCGTCCGAAAAGCATGGCAGACGC
Eg_p2.4     ATGCCACTGCGTATATGCCGCTGCCCAGCCTTGGCAGCATGTCGTCCGAAAGGCATGGCAGACGC
Eg_p2.5     -----ACTGCGTATACGCCGCTGCCCAGCCTTGGCAGCATGTCGTCCGAAAAGCATGGCAGACGC
Eg-p2.6     CATTCACTGCGTATACGCCGCTGCCCAGCCTTGGCAGCATGTCGTCCGAAAGGCATGGCAGACGC
                ********** **** ************************* *** ***** *******

Eg-p5       [LSU3451 (9)]
            AAGTGAGCCTTTCGAAGGCAGGTTGGGCTCCATTGTGCCCGCCTGTTCAATGCGGCTCAAAGATGT 3′RACE
```

```
Eg-p6       [SSU105]
Eg-p6       CCGCCCCTGACTGCTTCTGCTAACTGTTCGTCTTGGTTATCAGATGTAAGAAGGGGCCAGATGA 3'RACE
Eg-p6.1     CCGCCCCTGACTGCTTCTGCTAACTGTTCGTCTTGGTTATCAGATGTAAGAAGGGGCAAGGTGA 3'RACE
            ********************************************************** ** ***

Eg-p7       [LSU3332 (9)]
Eg-p7       TCCGCTTCGGGGATCGTCCAGCAGCCCCTCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAGATGG
Eg-p7.1     TCCGCTTCGGGGATCGTCCAGCAGCCCCTCTGTGGCATGGCTGCGTGGGTCAACACCCGGAGCCAGATGG
Eg-p7.2     TCCACTTCGGGGATCGTCCTGCAGCC--TCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAGATGG
Eg-p7.3     TCCGCTTCGGGGATCGTCCAGCAGCCCCTCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAAAGGG
Eg-p7.4     TCCGCTTCGGGGATCGTCCAGCAGCCCCTCTGTGACATGGCT---TGGGTCAACACCCGGAGCCAGATGG
Eg-p7.5     TCCGCTTCGGGGATCGTCCAGCAGCCCCTCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAAATGG
Eg-p7.6     TCCGCTTCGGGGATCGTCCAGCAACCCCTCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAGATGG
            *** *************** *** **  ****** *******   ********************* * **

Eg-p8       [LSU1365 (5)]
Eg-p8       ATGCACCATTCGAGTTGTTGTGCACAACATTGCACGAACAGTAGGAAAGGTGCAAGATTG
Eg-p8.1     ATGCACCAT-CGAGCTGTTGTGCACAACATTGCACGAACAGTAGGAAGGTGCAAAGATTG
Eg-p8.2     ATGCACCAT-CGAGCTGTTGTGCACAACATTGCATGAACAGTAGGAAGGTGCAAAGATTG
Eg-p8.3     ATGCACCATTCGAGTTGTTGTGCACATTATTGCACGAACCGTAGGGAGGTGCAAAGATTG 3'RACE
            ********* **** ***********  ****** **** ***** **************

Eg-p9       [LSU3503 (9)]
Eg-p9       CTGGGACCTCAATCGGGTGGCAGCGTCGGACCATGGTCGTTTGCTACCTACTCTTCCGTCCCAAGATTG
Eg-p9.1     CCGGGACCTCAATCGGGTGGCAGCGTCGTACCATGGTCGTTTGCTACCTACTCTTCCGTCCCAAGATTG
Eg-p9.2     CTGGGACCTCAATCGGGTGGCAGCGTCGTACCATGGTCGTTTGCTACCTACTCTTCCGTCCCAAGATTA 3'RACE
            * ************************** ************************************

Eg-p10      [LSU2842 (8)]
Eg-p10      TTGAATCCGCAATTTGGTTCTTGGGAGGCTCCAAAATGCCTCTCCAAGTCAGAATCTGCGGTAGAGTG
Eg-p10.1    TTGAATCCGCAATTTGGTTCTTGAGAGGCTCCAAAGTGCCTCTGCAAGTCAGAATCTGCGGTAGAGTG
Eg-p10.2    TTGAATCCGCAATTTGGTTCTTGGGAGGCTCCAAAGTGCCTCCCCAAGTCAGAATCTGCGGTAGAGTG
            ********************** ********** ******  ************************
```

```
Eg-p11     [SSU544]
Eg-p11     CCATCCACCGTTGTCTTTGCCAGTCCCGTCTGACCCGCATGGCTGGGCTGGTGGGTAGAGGTGGAGAGATGT
Eg-p11.1   CCATCCACCGTTGTCTTTGCCAGCCCCGTCTGACCCGCATGGCTGGGCTGGTGGGTAGAGGTGGAGAGATGT
Eg-p11.2   ---TCCACCGTTGTCTTTGCCAGTCCCGTCTGACCCGCATGGCTGGGCTGGTGGGTAGAGGCGGAAAGATGT 3'RACE
              ******************** ************************************ *** ******

Eg-p12     [LSU2904 (8)]
Eg-p12     ATAAGGCCTGAATAGGTTAGAGGGGACTGATGAAGTTGCGTCTCCTCAGCTCTCAGGCCAAGATGT 3'RACE
Eg-p12.1   ----GGCCTGAATAGGTTAGAGGGGACTGATGAAGTTGCGTCTCCTCAGTTCTCAGGCCAAGATGT 3'RACE
               ******************************************** ***************

Eg-p13     [LSU1568 (6)]
           ACACCCGTCCTTGTGGACTGGCCATCTTCTCAATGGATGCAGCTGGTTCTGAAAGGCGGGGAGACCG 3'RACE

Newly identified box AGA snoRNA species and their isoforms identified through bioinformatic analysis of a ncRNA library

Eg-p14     [LSU3235 (9)]
Eg-p14     AGGGGCACCCTCTGAGGCATCCTGGAGGCGCGCTGGCTCTCTCCTGGGTGCATGCTGTGGTGCCCAGATGG
Eg-p14.1   GGGGGCACCCTCTGAGGCATCCTGGAGGCGCGCTGGCTCTCTCCTGGGTGCATGCTGTTGTGCCCAGATGG
Eg-p14.2   -GGGGCACCCTCTGAGGCATCCTGGAGGC--GCTGGCTCTCTCCTGGGTGCATGCTGTGGTGCCCAGATGG
            ****************************  *************************** ************

Eg-p15     [LSU2752 (8)]
Eg-p15     TGCACTTTTAAGAGCCCCCGGCCCTGCCTTTCGGCATCGTGGGCTGTGGGGTGTCATAGGGTGCAGAGATGA
Eg-p15.1   TGCACTTTTAAGAGCTCCTGGCTCTGCCCTTCGGCATCGTGGGCTGTGGGGTGTCATAGGGTGCAGAGATGA
Eg-p15.2   TGCACTTTTACGAGCCCCCGGCCCTGCCTTTCGGCATCGTGGGCTGTGGGGTGTCATAGGGTGCAGAGATGA
           ********** **** ** *** ***** ****************************************

Eg-p16     [SSU1592]
Eg-p16     AAGCAGGACGTCAAATCCATGGGGGCAGATGCTGTGGTCCCTTTCGGAGCTAAATCCTGCCAGAAG-
Eg-p16.1   -AGCAGGACGTCAAATCCCTGGGGGCACATGCTGTGGTCCCTTTCGGAGCTAAATCCTGCCAGA---
Eg-p16.2   AAGCAGGACGTCAAATCCATGGATGCAGATGCTGTGGTCCCTTTCGGAGCTAAATCCTGCCAGAAGA
            ***************** ***  *** ***********************************
```

```
Eg-p17      [LSU3701 (10)]
Eg-p17      CAGCACAAACGCGTTACGTCGGTATCAGTCTCATGGTGCCGGGGTGTCAGATGTTGTGCCAGACCC
Eg-p17.1    CAGCACAAACGCGTCACGTCGGTATCACTCTCATGGTGCCGGGTTGTCAGATGTTGTGCCAGACC-
Eg-p17.2    CAGCACAAACGCGTCACGTCGCTATCAGCCTCATGGTGCCGGGGTGTCAGATGTTGTGCCAGACC-
Eg-p17.3    CAGCACAACCGCGTTACGTCGGTATCAGTCTCATGGTGCCGGGGTGTCAGATGTTGTGCCAGACC-
Eg-p17.4    CAGCACAAACGCGTTACGTCGGTATCAGTCTCTTGGTGCCGGGGTGTCAGATGTTGTGCCAGACCC
Eg-p17.5    CAGCACAAACGCGTTACGTCGGTATCAGTCTCATGATGCCGGGGTGTCAGATGTTGTGCCAGACCC
Eg-p17.6    CAGCACACACGCGTTACGTCGGTATCAGTCTCATGGTGCCGGGGTGTCAGATGTTGTGCCAGACC-
Eg-p17.7    CAGCACAGACGCGTTACGTCGGTATCAGTCTCATGGTGCCGGGGTGTCAGATGTTGTGCCAGACCC
Eg-p17.8    CAGCACAAACGCGTTACGTCGGTATCAGTCTCATGGTGCCGGGGCGTCAGATGTTGTGCCAGACC-
            *******  ***** ****** *****  *** ** *******  ********************

Eg-p18      [SSU1393]
Eg-p18      GTGCTGTGATCTGATGCCTGTGCCGGTGCCTTCTCTGGCACGGGTTCTTTCCCACAGCAAGAT--
Eg-p18.1    GCGCTGTGATCTGATGCCTGTGCTGGTTCCCACTCTGGCACGGGTTCTTTCCCACAGCAAGAT--
Eg-p18.2    GTGCTGTGATCTGATGCCTGTGCCGGTGCTTTATCTGGCACGGGTTCTTTCCCACAGCAAGATGA…
            * ********************** *** *    ******************************

Eg-p19      [LSU3953 (13)]
            CAGTGCCTGGCCTCTGAAACAGCGCGTGGGTGGCGTCCCCGCGCTGTCTTAAGCAGGCAGAGA

Eg-p20      [LSU2915 (8)]
Eg-p20      ACAGGGGTGCAGTATCGGGCAGATGGCAGCCTGGGTCAGTTGCATTTGCCTACTAAACCACCCCGAGATG-
Eg-p20.1    -CAGGGGTGCAGTATCGGGCAGATGGCAGCCTGGGTCAGTTGCATTTGCCTATTAAACCACCCCGAGATGT
Eg-p20.2    ACAGGGGTGCAGTATCGGGCAGATGGCAGCCTGGGTCAGTTGCATTTTCCTACTAAACCACCCCGAGATG-
             ********************************************** **** ****************

Eg-p21      [LSU2914 (8)]
Eg-p21      CTGCCAGGATTTTATCACCCCTCCTGACACATTTATTGTTGGAGGGGCTAGAGGTCCTGGTAGATGT
Eg-p21.1    ATGCCAGGCTTTTCTCACCCCTCCTGACAAATTTATTGTTGGAGGGGCTCGAGGTCCTGGTAGATGT
Eg-p21.2    CAGCCAGGATTTTATCACCCCTCCTGACACATTCATTGTTGCCGGCGCTAGAGGTCCTGGTAGATGT…
Eg-p21.3    -TGCCAGGACTTTATCACCCCTCCTGACACATTTATTGTTGGAGGGGCTAGAGGTCCTGGTAGATGT…
              ******  *** *************** *** *******  ** ********************
```

Eg-p22 **[LSU1023 (5)]**

```
Eg-p22      ATCGAGGTACTAGCCCAGGGCATGCGATCTTTCTGCATATCCCTTCTCAATGGCCTCGTAGACCT
Eg-p22.1    ATCGAGGTACTAGCCCAGGGCATGCGATCTTTCTGCATATCCCTTCTCAATGTCCTCGTAGAT--
            ************************************************** *********
```

Eg-p23 **[LSU1266 (5)]**

```
Eg-p23      CATGCACTTTCAGCACCTCTCCTTAAATCTTGAGGCTGGGGTTCCTGAGAGTGCACAGATCC
Eg-p23.1    CATGCACTTTAAGCACCTCTCCTTAAATCTTGAGGCTGGGGTTCCTGAGGGTGCACAGATCC
Eg-p23.2    CATGCACTTTAAGCACCTCTCCTTAAATCTTGAGGCTGGGGTTCCTGCGGGTGCCCAGATCC
Eg-p23.3    CATGCACTCCACGCACCTCTCCTTAAATCTTGAGGCTGGGGTTCCTGAGGGTGCACAGATCC
Eg-p23.4    CATGCACTCTAAGCACCTCTCCTTAAATCTTGAGGCTGGGGTTCCTGGGGGTGCACAGATCC
Eg-p23.5    CATGCACCTTAAGCACCTCTCCTTAAATCTTGAGGCTGGGGTTCCTGAGGCTGCACAGATCC
            *******     ********************************** *  ***********
```

Eg-p24 **[LSU280 (2)]**

```
Eg-p24      CTGGCAATATTTTCAGGCGGAGCTCAGTCCCGTCTCTGCTCCGGAGTGCAATTTGCCGAGATTG
Eg-p24.1    CTGGCAATATTTTCAGGCGGAGCTCAGTCCCGTCTCTGTTCCGGAGTGCAATTTGCCGAGATTG
            ************************************ ************************
```

Eg-p25 **[LSU1235 (5)]**

```
            CGTGGTGCTGGAACCAGCTCTGCACTGCACTGTGCAGTTGCAGACTAGAATGCACCACAGATGC
```

Eg-p26 **[LSU3697 (10)]**

```
Eg-p26      AAGGCAGCTCCAGAGCACCTTGGAAGCCAGTGCTTTTCCATGGTGCGTATCCTGCTGCCAAGATTG
Eg-p26.1    AAGGCAGCTCCAGCGCACCTTGGAAGCCAGTGCTTTTCCATGGTGCGTATCCTGCTGCCAAGATTG…
Eg-p26.2    AAGGCAGCTCCCGAGCACCTTGGAAGCCAGTGCTTTTCCATGGTGCGTATCCTGCTGCCAAGATTG…
            *********** * ******************************************************
```

Eg-p27 **[LSU3568 (9)]**

```
            ATGCACTGCTAAAGGCTTCGCAAGCATCGATGTTGTTGCTTGCGGAGAAACCAACGGTGCAAGATT
```

Eg-p28 **[LSU3042 (9)]**

```
            ACCGCTCAGACGACCACCTGCTGCAACATACAATGAGCTGCAGGTGGCATATGGTGCGGAGATCC
```

Eg-p29 **[SSU1660]**

```
Eg-p29      CCGCCTGTGATCAAGGGTGTTTCGGGCTTTTGCATGGCCGCTACATCAGAAAGGCAGGCAAGAATC
Eg-p29.1    CCGCCTGTGATCAAGGGTGTTTTGGGCTTTTGCATGGCCGCTACATCAGAAAGGCAGGCAAGAATC
            ********************** ****************************************
```

Eg-p30 **[SSU2065]**

```
Eg-p30      CCGGCGGCCCAAGGGGCGTGCGCTGGACACCGGGTTCCGCCACGCTGACAGATGGCCGGCCCAGACAT
Eg-p30.1    CCGGCGGCACAAGGGGCGTGTGCTGGACACCGGGTTCCGCCACGCTGACAGATGGCCGGCCCAGACAT
Eg-p30.2    CCGGCGGCACAAGGGGCGGGCGCTGGACACCGGGTTCCGCCACGCTGACAGATGGCCGGCCAAGACAT
Eg-p30.3    CCGGCGGCACAAGGGGCGTGCGCTGGACACCGGGTTCCGCCACGCTGACAGATGACCGGCCCAGACAT
Eg-p30.4    CCGGCGGCACAAGTCGCGTGCGCTCGACACCGGGTTCCGCCACGCTGACAGATGACCGGCCCAGACAT
Eg-p30.5    CCGGCGGCACACGGGGCGTGCGCTGGACACCGGGTTCCGCCACGCTGACAGATGACCGGCCCAGACAT
Eg-p30.6    CCGGCGGCCCAAGGGGCGTGCGCTGGACACCGGGTTCCGCCACGCTGACAGATGACCGGCCCAGACAT
            ******* ** *  *** * ********************************** **************
```

Eg-p31 **[LSU2119 (7)]**

```
Eg-p31      GCCCACCAGGAAACGACCCCTGCCCACACCCTTCACTCTGTGCAGGGGACTTGAAGGTGGAAGAAAC
Eg-p31.1    GCCCACCAGGAAACGACCCCTGCCCACACCCTTCACTCTGTGCAGGGGACTCGAAGGTGGAAGAAAC
Eg-p31.2    GCCCACCAGGAAACGACCCCTGCCCACAGCTTTCACTCTGTGCAGGGGACTCGAAGGTGGAAGAAAC
            **************************** * ******************** ***************
```

Eg-p32 **[LSU2361 (8)]**

```
Eg-p32      GAGCCTTGTTTAACCTCCAGCAGCCTGCTTGTTTGGGCTCTGGACTTCCCTCGAGGCAAGATTT…
Eg-p32.1    GAGCCTTGTTTAACCTCCAGCAGCCTGCTTGTTTGGGCTCTGGACTTCCCTCGAGGCAAGACTT…
Eg-p32.2    GAGCCTTGCTTAACCTCCAGCAGCCTGCTTGTTTGGGCTCTGGACTTCCCTCGAGGCAAGACTT…
Eg-p32.3    GAGCCTTGCTTAACCTCCAGCAGCCTGCTTGTTTGGGCTCTGGACTTCCCTCGAGGCAAGATTT…
Eg-p32.4    GAGCCTTGCTTCACCTCCAGCAGCCTGCTTGTTTGGGCTCTGGACTTCCCTCGAGGCAAGATTT…
Eg-p32.5    GAGCCTTGTTTCACCTCCAGCAGCCTGCTTGTTTGGGCTCTGGACTTCCCTCGAGGCAAGACTT…
Eg-p32.6    GAGCCTTGTTTCACCTCCAGCAGCCTGCTTGTTTGGGCTCTGGACTTCCCTCGAGGCAAGATTT…
Eg-p32.7    GAGCCTTGTTTAACCTCCAGCAGCCTGCTTGTTTGGGCTCTGGACTTCCCTCGAGTCAAGATTT…
Eg-p32.8    GAGCCTTGTTTAACCTCCAGCAGCCTGCTTGTTTGGGCTCTGGACTTCCATCGAGGCAAGATTT…
            ******** ** ************************************ ***** ***** **
```

Eg-p33 [LSU2943 (8)] related to Eg-p48

```
Eg-p33     CCCGTGCACATTATGTGTAGGCACGGCTTCGGAAGCCATACCGTGTCTACCGCCCATGTGCACCAGATGT
Eg-p33.1   CCCGTGCGCATTATGTGTAGGCACGGCTTCGGAAGCCATACCGTGTCTACCGCCCATGTGCACCAGATGT
Eg-p33.2   CCCGTGCACATTATGTGTAGGCACGGCTTCGGAAGCCATGCCGTGTCTACCGCCCATGTGCACCAGATGT
           ******* ****************************** ******************************
```

Eg-p34 [SSU2116]

```
           AACAAGGCGGGACAGGGAGCCGGCATTCAGATGCTGGTGTCCCATCGTGAAGCCTTGGAGATCC
```

Eg-p35 [LSU3542 (9)]

```
Eg-p35     AAGCACCGGGACGGTCGCACCGCTGGCTCGCC-TTTGTGCCGGCGGTGCATCCCAACGTGCCAGATGG
Eg-p35.1   AAGCACCGGGACGGTCGCACCGCTGGCTCGCC-TTTGTGCCGGCGGTGCATCCCAGCGTGCCAGATGG
Eg-p35.2   AAGCACCGGGACGGTCGCACTGCTGGTTCGAT-TTTGTGTTGGCAGTGCATCCCAACGTGCCAGATGG
Eg-p35.3   ACGCACCGGGACGGTCGCACCGCTCGCTCGTTGCTTGTGCGGGCGGTGCATCCCAACGTGCCAGATGG
Eg-p35.4   AAGCACCGGGACGGTCGCACCGCTCGCTCGTCATTTGTGCGAATGGTGCATCCCAACGTGCCAGATGG
Eg-p35.5   AAGCACCGGGACGGTCGCACCGCTGGCTCGCC-TTTGTGCCGGCGGTGCATCCCAATGTGCCAGATGG
           * ****************** *** * ***    *****      ********** ************
```

Eg-p36 [SSU1068]

```
Eg-p36     CACGCACCCATCCACATTGGGGCCACCACATCCTGCTTGCCGGACCCCGCCACACGGGTGCCAGATAT
Eg-p36.1   -ACGCACCCATCCACATTGGGGCCCCCACACCCTGCTTGCCGGACCCCGCCACACGGGTGCCAGATAT
Eg-p36.2   -ACGCACCCATCCACATTGGGGCCACCACATCCTGCTTGCCGGACCCCGCCACACGGGCGCCAGATAT
Eg-p36.3   -ACACACCCATCCACATTGGGGCCACCACATCCTGCTTGCCGGACCCCGCCACACGGGTGCCAGATAT
Eg-p36.4   CACGCACCCATACGCATTGGGGCCACCACATCCTGCTTGCCGGACCCCGCCACACGGGTGCCAGATAT
           ********** * ********** ***** *************************** *********
```

Eg-p37 [LSU3204 (9)]

```
Eg-p37     CTGCCCTGCCCGTAGTATACCTTGTGCTCGCTGCGCTTGGTAAGCAAATGCAGGGCAAGATAT…
Eg-p37.1   CTGCCCTGCCCGTAGTCTACCTTGTGCTCGCTGCGCTTGGTAAGCAAATGCAGGGCAAGATAT…
Eg-p36.2   CTGCCCTGCCCGTCGTATACCTTGTGCTCGCTGCGCTTGGTAAGCAAATGCAGGGCAAGATAT…
           ***************** ********************************************
```

```
Eg-p38      [SSU2305] * bp to intergenic region
Eg-p38      CAGGCACTCCAAAACTGGCACCCGACCCCC---GCTGGGTTGAGGTGTATGATCCGGTGCCCAGACCG
Eg-p38.1    CAGGCACTCCAAAACTGGCACCCGACCCCC---ACTGGGTTGAGGTGTATGATCCGGTGCCCAGACCG
Eg-p38.2    CCGGCACTCCAGAACTGGCATCTGACCCCTCTTGATGGGCCGCGGTGTATGATCCGGTGCCCAGACAC
Eg-p38.3    CCGGCACTCCAAAACTGGCATCTGACCCCTCTTGATGGGCCGCGGTGTATGATCCGGTGCCCAGACAC
            * ********* ******** * ******      ****  * ***********************

Eg-p39      [LSU1407 (5)]
Eg-p39      CCGGGGCTATGTGGCGCCTGCGACAGCTGTCAGTGCTGATCCAGGGAAGCATGCCCCCAGACCG
Eg-p39.1    CCGGGGCTATGTGGCGCCTGCGACAGCTGTCAGTGCTGATCCAGGGAAGCATGCCCCTAGACCG
Eg-p39.2    CCGGGGCTCTGTGGCGCCTGCGACAGCTGTCAGTGCTGATCCAGGGAAGCATGCCCCCAGACCG
            ******** ********************************************** ******

Eg-p40      [LSU567 (3)]
Eg-p40      CAGCAGCTCCTCCTTCGCTCAGTGGACCTTTTGATGCTGGGCTCACGTGGGCTGCCAGAAAT…
Eg-p40.1    CAGCAGCTCCTCCTTCGCTCCGTGGACCTTTTGATGCTGGGCTCACGTGGGCTGCCAGAAAT…
Eg-p40.2    CAGCAGCTCCTCCTTTGCTCAGTGGACCTTTTGATGCTGGGCTCACGTGGGCTGCCAGAAAT…
            *************** **** ****************************************

Eg-p41      [LSU1692 (6)]
Eg-p41      -CCGGTGTCTTGTACAAAGGCCTGCGCTCTCTGTGCGGAGGTCTTCTGGCAAAGGCACCCAGATGC
Eg-p41.1    -CCGGTGTCTTGTACAAAGGCCTGTGCTCTCTGTGCGGAGGTCTTCTGGCAAAGGCACCCAGATGC…
Eg-p41.2    GCCGGTGTCTTGTACAAAGGCCTGTGCTCTCTGTGCGGAGGTCTTCTGGCAAAGGCACCCAGA---
             ********************** ************************************

Eg-p42      [SSU89]
Eg-p42      TTGCACTCTGCCATTGGATCAGGCGATTATGCTTGTCGCCTGGTCCCAGAAAGGAGTGCCAGATTT…
Eg-p42.1    GTGCACTCTGCCATTGGATCAGGCAATTATGCTTGTCGCCTGGTCCCAGAAAGGAGTGCCAGATTT
Eg-p42.2    TTGCACTCTGCCATTGGATCAGGCGATTATGCTTGTCGCTTGGTCCCAGAAAGGAGTGCCAGATTT
Eg-p42.3    TTGCACTCTGCCATTGGATCAGGCGATTACGCTTGTCGCCTGGTCCCAGAAAGGAGTGCCAGATTT…
Eg-p42.4    TTGCACTCTGCCATTGGATCAGACGATTATGTATGTCGCCTGGTCCCAGAAAGGAGTGCCAGATTT…
            ********************* * **** *  ****** *************************
```

Eg-p43 **[LSU2746 (8)]**

```
Eg-p43      ACGCACGTGCCAGAGTCAGCAGGCCTGCCGGGAGTGCAGTGCCATGTGTTATGCATGTGCCAGAAGG
Eg-p43.1    ACGCACGTGCCAGAGTCAGCCGGCCTGCCGGGAGTGCAGTGCCATGTGTTATGCATGTGCCAGA---
Eg-p43.2    ACGCACGTGCCCGAGTCAGCAGGCCTGCCGGGAGTGCAGTGCCATGTGTTATGCATGTGCCAGAAGG
Eg-p43.3    ATGCACGTGCCAGAGTCAGCAGGCCTGCCGGGAGTGCAGTGCCATGTGTTATGCATGTGCCAGA---
            * ********* ******** ****************************************
```

Eg-p44 **[SSU1624]**

```
            AAGCCTTGTTATCCTCCAGCCGCCTTCTCTGGGCTCTGGCCTTCCCTCAAGGCAAGATTC
```

Eg-p45 **[LSU2837 (8)]**

```
Eg-p45      GCGCATGACTCACCAGCCGCAGGCTCTCGTGGATGTCGCGCCATGCGGCCAGCAGGAATGCAAGAGGT
Eg-p45.1    GCGCATGACTCACCAGCCGCAGGCCCTCGTGGATGTCGCGCCATGCGGCCAGCAGGAATGCAAGAGGT
            ************************ ****************************************
```

Eg-p46 **[LSU1926 (6)]**

```
            CAGGTGACGCACATTCCCTGGGTAGCATGAAACTTGCTTCCTTGGCGACTGTTCACCGAGATT
```

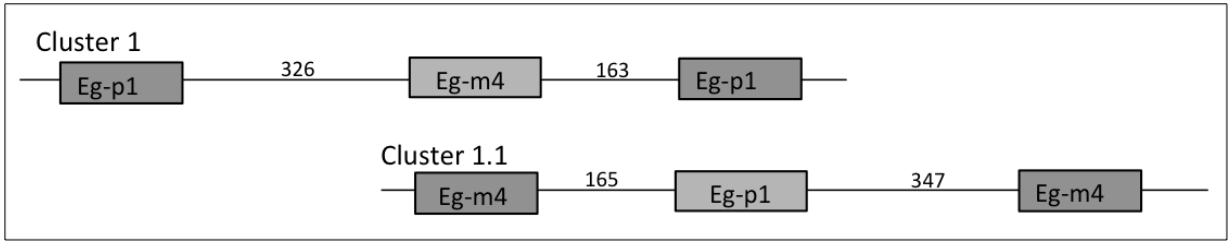
Eg-p47 **[LSU480 (3)]**

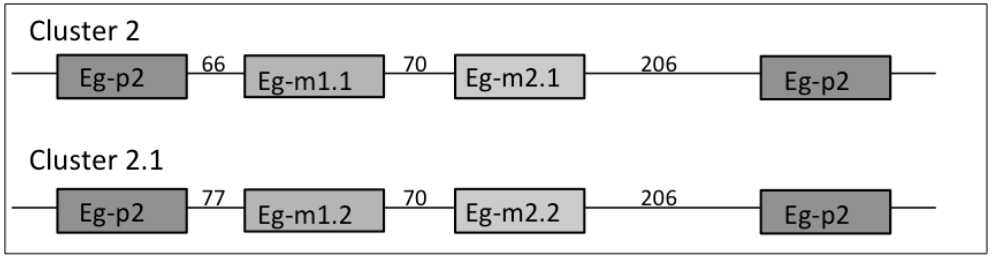
```
            ACTCCATCCCAAGCGGTTCTCAGTTTGCATAACTGATGAATCCTCCTCTGATGGAAAGAGTCTCG
```

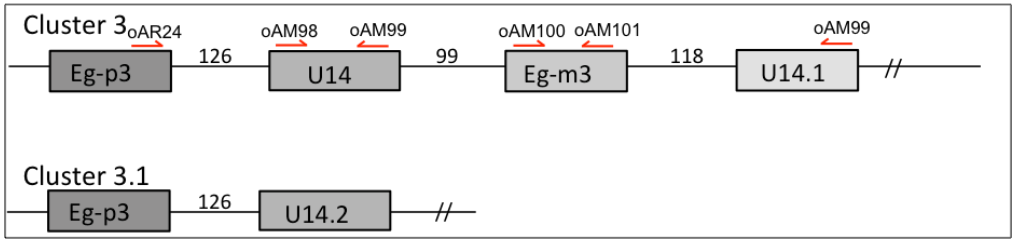
Eg-p48 **[LSU3591 (10)] (related to Eg-p33)**

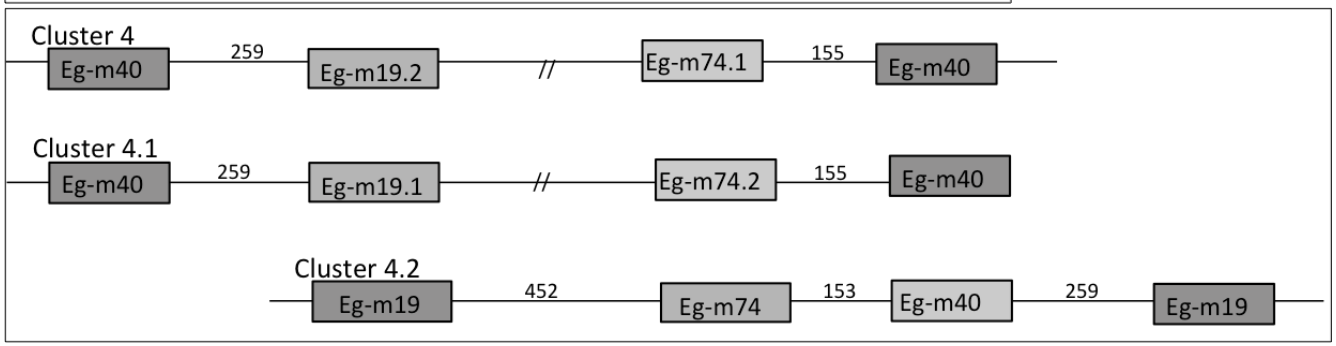
```
Eg-p48      CCGTGCACCTATTGCATGGCCAGGACAGCGCCTCCTTTGTCTGGCCATCGCCCCCCGTGCACCAGATCC
Eg-p48.1    CCGTGCACCTATTGCATGGCCAGGACAGCGCCTCCTTTGTCTGGCCATCGCCTCCCGTGCACCAGATCC
Eg-p48.2    CCGTGCACCTATTGCATGGCCAGGAAAGCGCCTCCTTTGTCTGGCCATCGCCCCCCGTGCACCAGATCC
            ************************* ************************** ***************
```

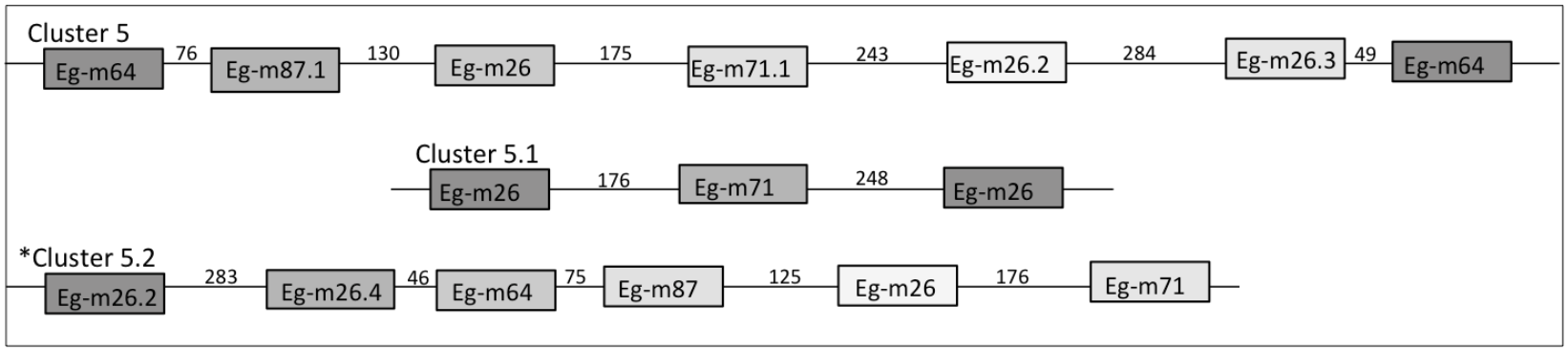
Figure S2. snoRNA gene clusters in *E. gracilis*. The snoRNA coding regions are shown as shaded boxes and intergenic sequence is represented by solid lines. The distance in base-pairs between each snoRNA coding region is indicated. A double slash indicates that the precise distance between the two genes is not known, primarily caused by difficulties in sequencing these intergenic regions. Clusters are grouped (for example, Cluster 1 and Cluster 1.1) when they share at least two snoRNA coding regions in common. Primer binding locations employed in RT-PCR experiments as described in Figure 16 and Figure S6 are shown as red arrows. Those clusters identified during our examination of raw preliminary *E. gracilis* genome sequence data provided by the MC Field laboratory database (http://web.me.com/mfield/Euglena_gracilis) are indicated with asterisks. Primer combinations used to amplify each cluster are described in Table S1. Cluster 2 GenBank accession no. JN051179; clusters 2.1-5.2 accession nos. JN051170-JN051178, respectively; clusters 6-34 accession nos. JN051180-JN051253, respectively.

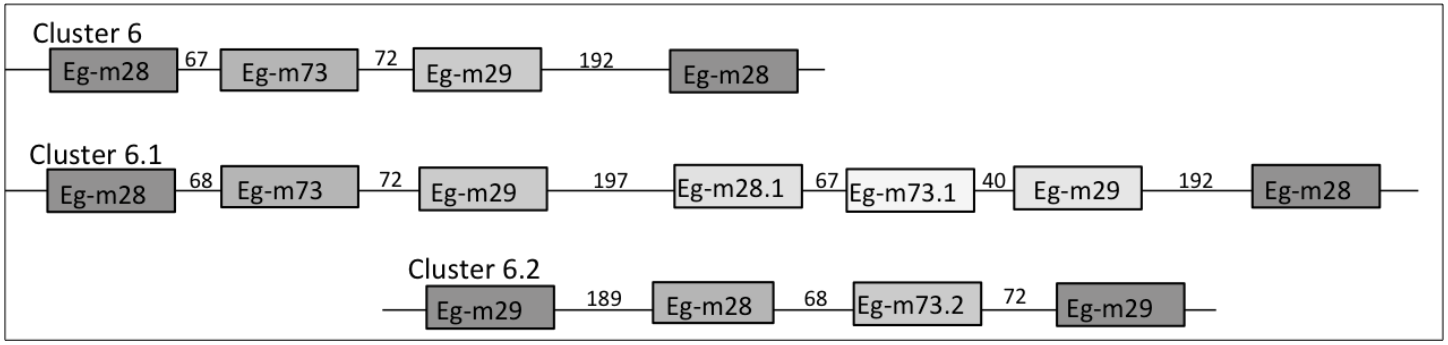


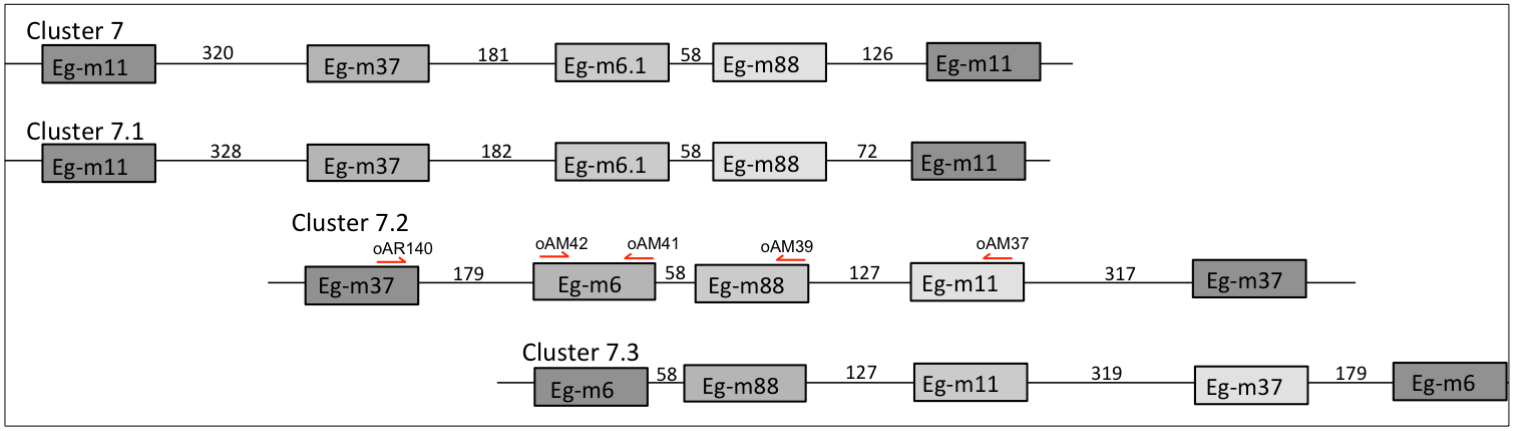


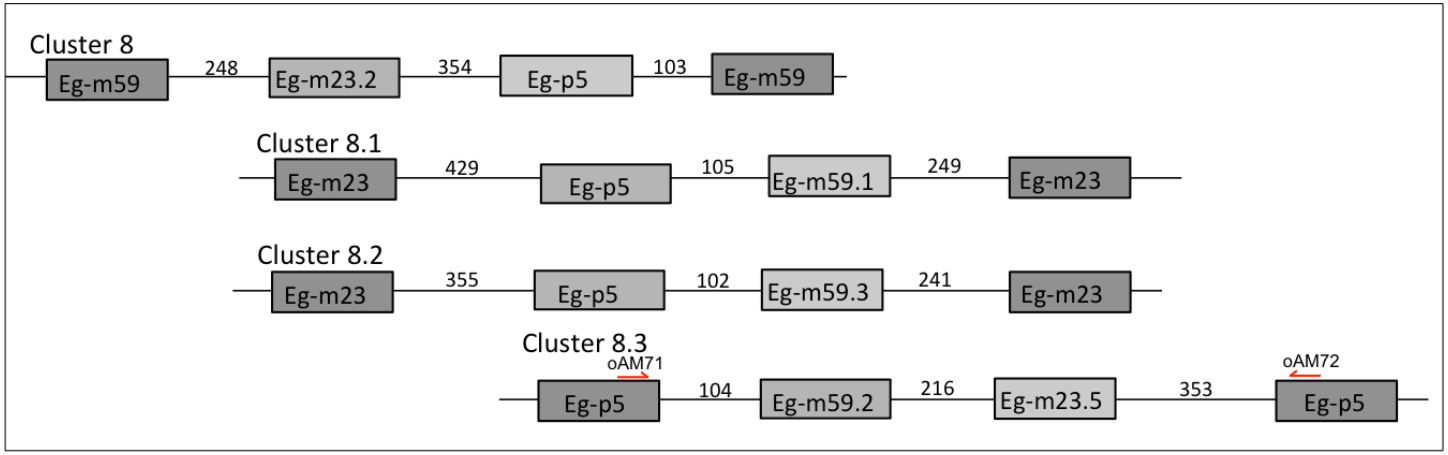

Cluster 3
oAR24
oAM98
oAM99
oAM100
oAM101
oAM99
Eg-p3
126
U14
99
Eg-m3
118
U14.1
//
Cluster 3.1
Eg-p3
126
U14.2
//

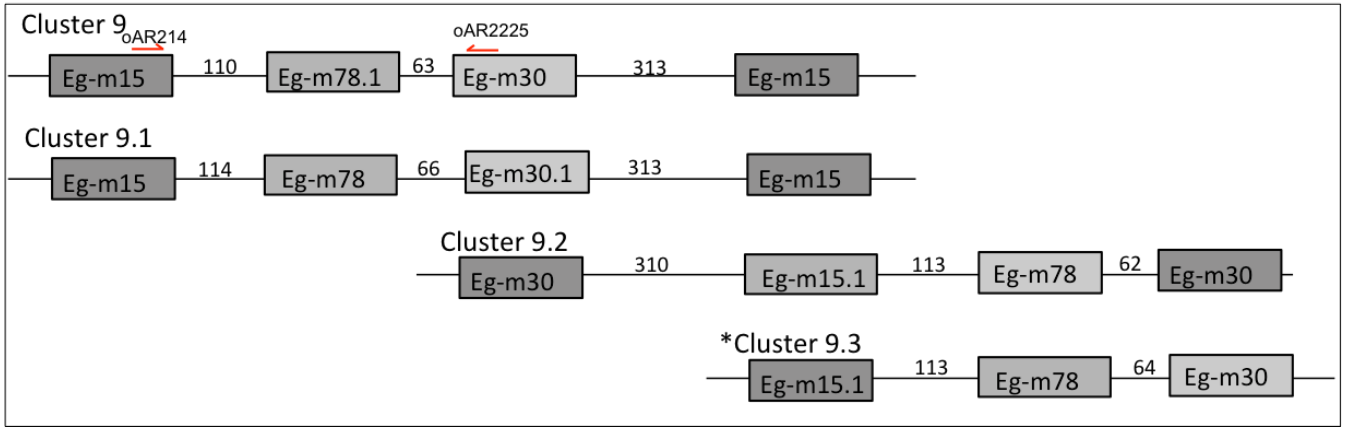

Cluster 4
Eg-m40
259
Eg-m19.2
//
Eg-m74.1
155
Eg-m40
Cluster 4.1
Eg-m40
259
Eg-m19.1
//
Eg-m74.2
155
Eg-m40
Cluster 4.2
Eg-m19
452
Eg-m74
153
Eg-m40
259
Eg-m19

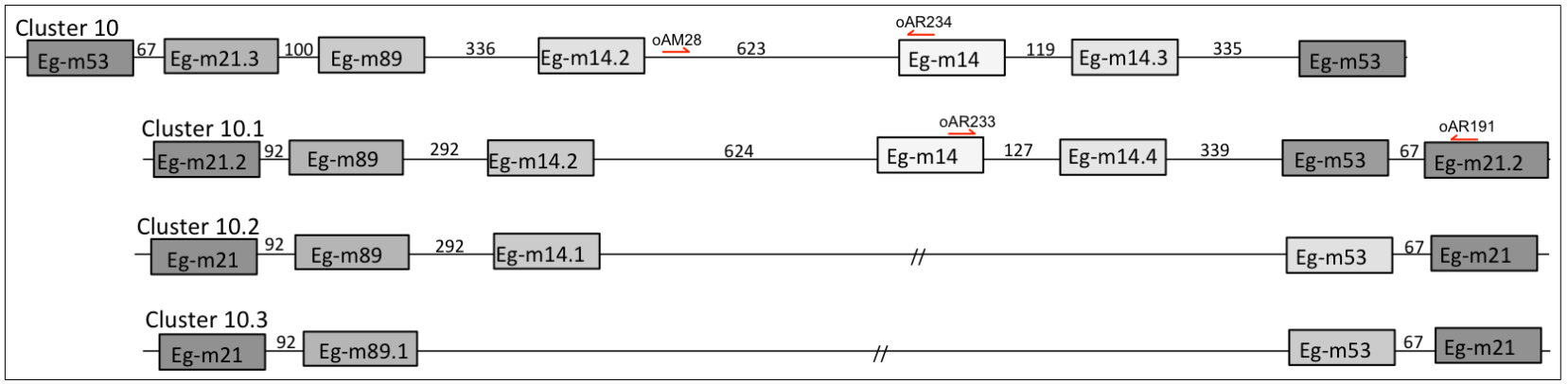

Cluster 5
Eg-m64
76
Eg-m87.1
130
Eg-m26
175
Eg-m71.1
243
Eg-m26.2
284
Eg-m26.3
49
Eg-m64
Cluster 5.1
Eg-m26
176
Eg-m71
248
Eg-m26
*Cluster 5.2
Eg-m26.2
283
Eg-m26.4
46
Eg-m64
75
Eg-m87
125
Eg-m26
176
Eg-m71

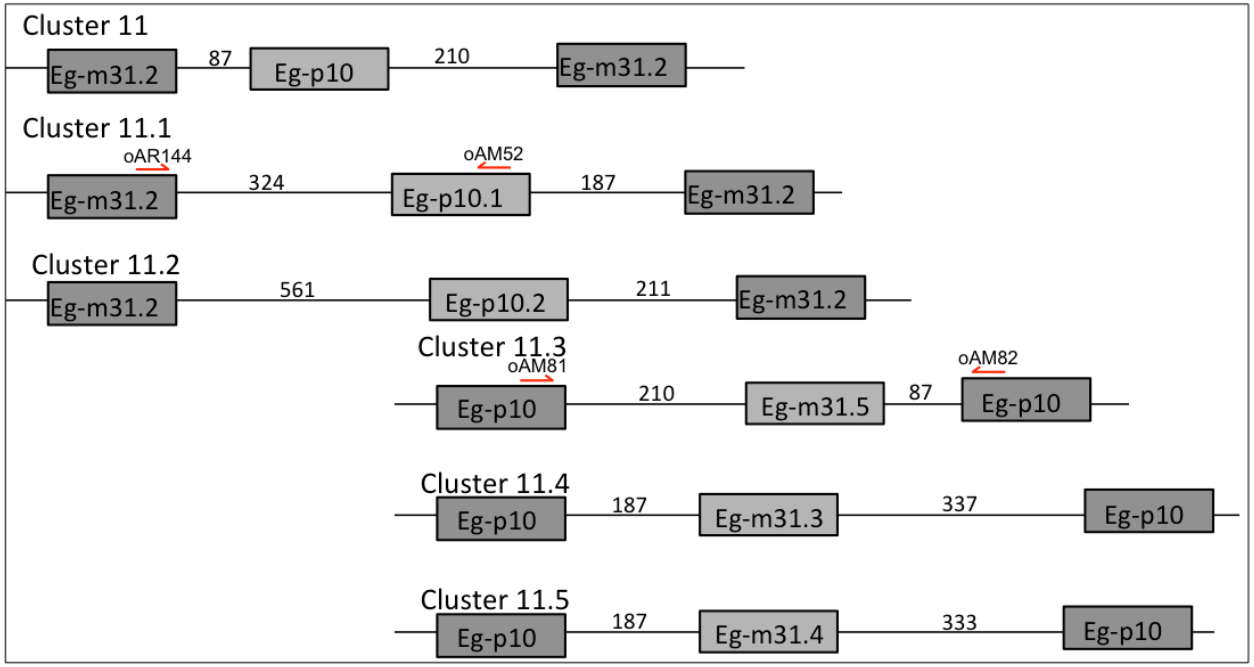

Cluster 6
Eg-m28
67
Eg-m73
72
Eg-m29
192
Eg-m28
Cluster 6.1
Eg-m28
68
Eg-m73
72
Eg-m29
197
Eg-m28.1
67
Eg-m73.1
40
Eg-m29
192
Eg-m28
Cluster 6.2
Eg-m29
189
Eg-m28
68
Eg-m73.2
72
Eg-m29

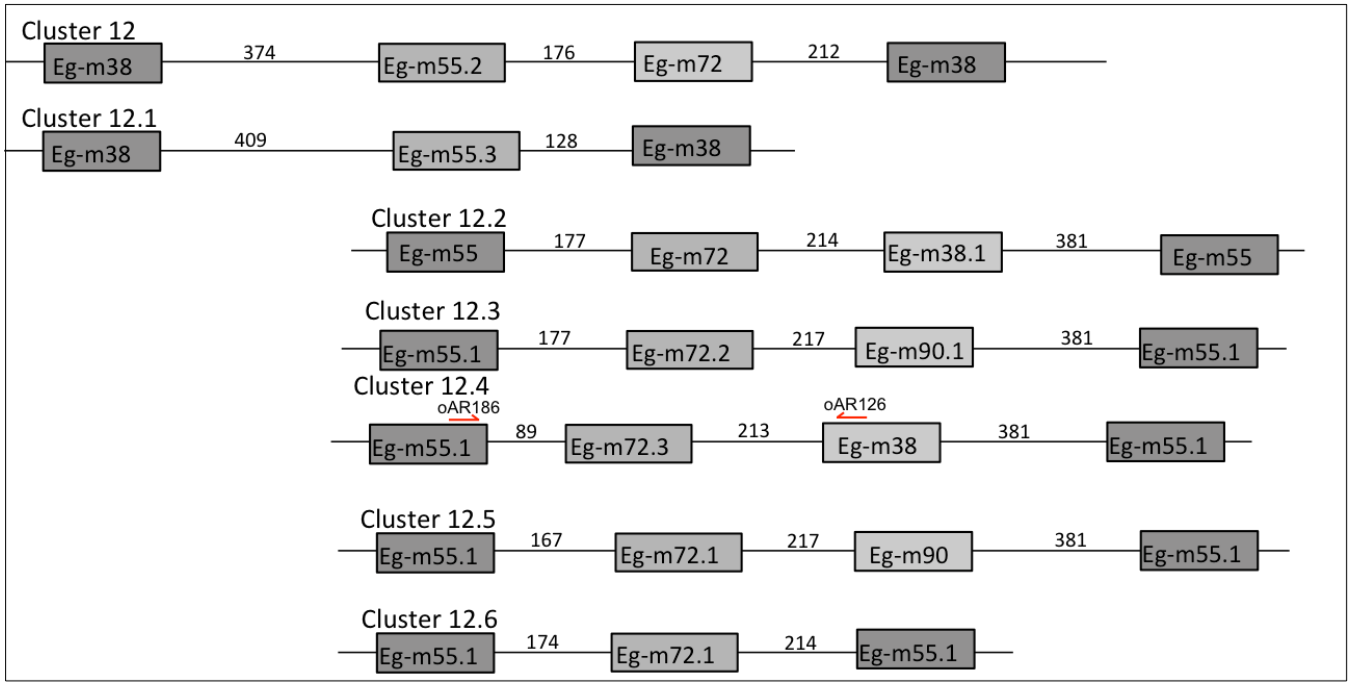

Cluster 7
Eg-m11
320
Eg-m37
181
Eg-m6.1
58
Eg-m88
126
Eg-m11
Cluster 7.1
Eg-m11
328
Eg-m37
182
Eg-m6.1
58
Eg-m88
72
Eg-m11
Cluster 7.2
oAR140
oAM42
oAM41
oAM39
oAM37
Eg-m37
179
Eg-m6
58
Eg-m88
127
Eg-m11
317
Eg-m37
Cluster 7.3
Eg-m6
58
Eg-m88
127
Eg-m11
319
Eg-m37
179
Eg-m6

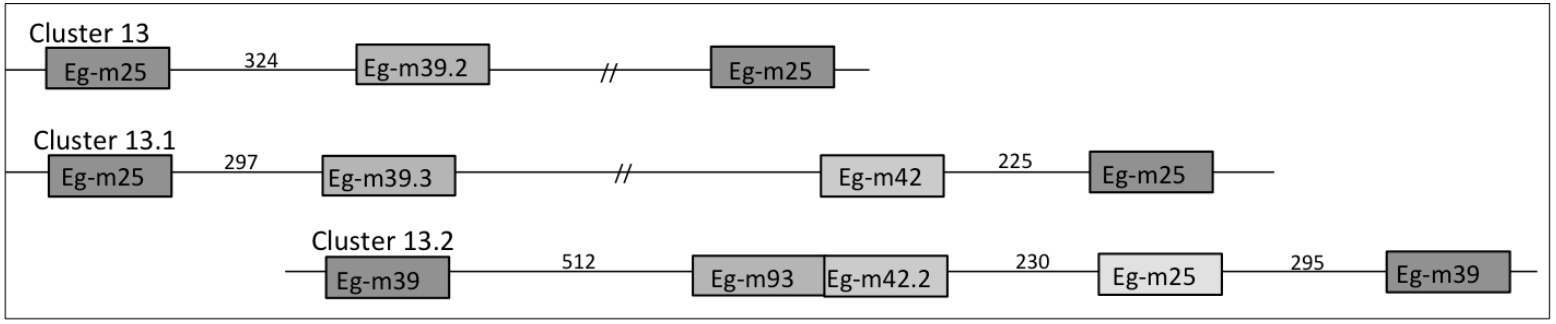

Cluster 8
Eg-m59
248
Eg-m23.2
354
Eg-p5
103
Eg-m59
Cluster 8.1
Eg-m23
429
Eg-p5
105
Eg-m59.1
249
Eg-m23
Cluster 8.2
Eg-m23
355
Eg-p5
102
Eg-m59.3
241
Eg-m23
Cluster 8.3
oAM71
oAM72
Eg-p5
104
Eg-m59.2
216
Eg-m23.5
353
Eg-p5

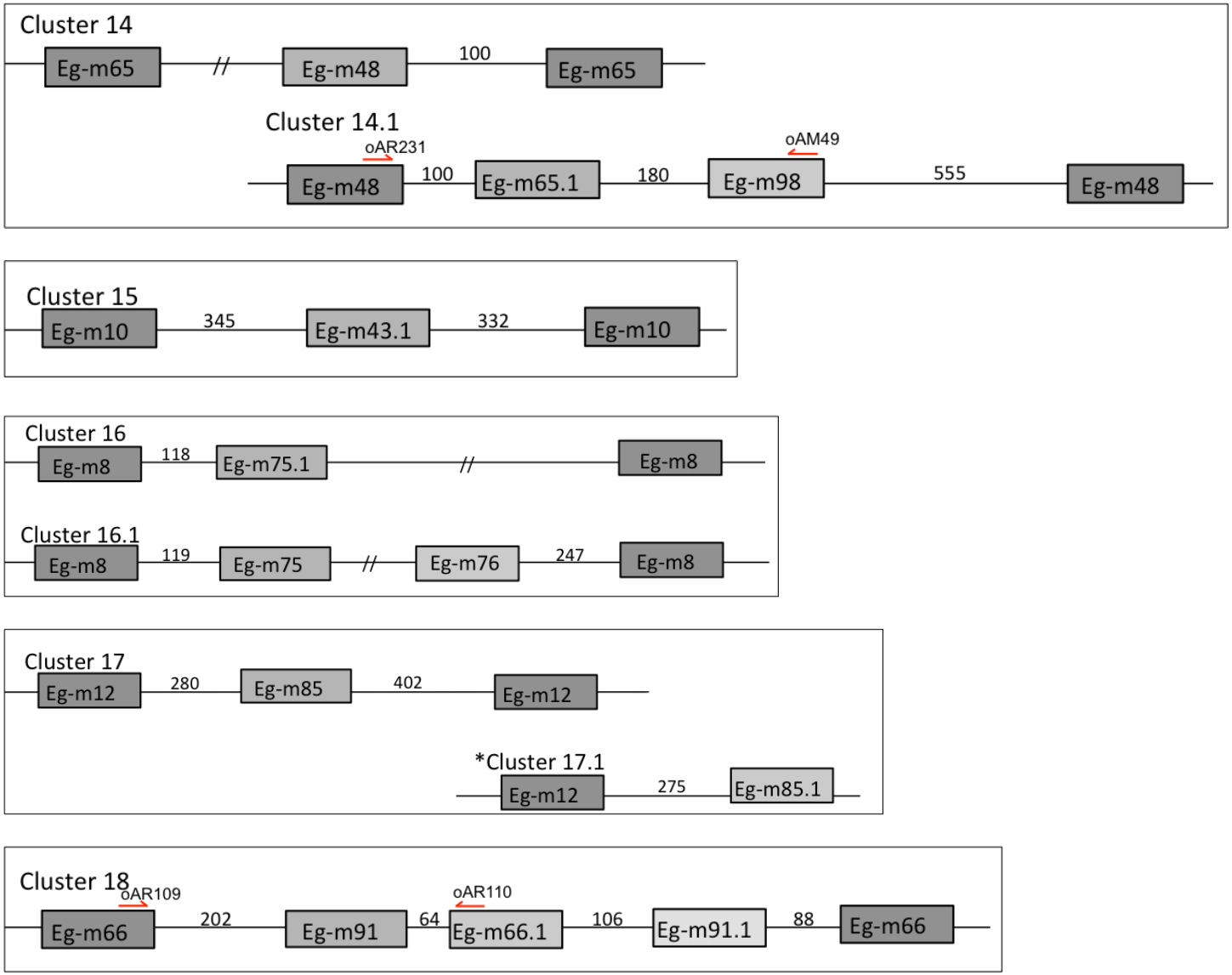

Cluster 9
oAR214
oAR2225
Eg-m15
110
Eg-m78.1
63
Eg-m30
313
Eg-m15
Cluster 9.1
Eg-m15
114
Eg-m78
66
Eg-m30.1
313
Eg-m15
Cluster 9.2
Eg-m30
310
Eg-m15.1
113
Eg-m78
62
Eg-m30
*Cluster 9.3
Eg-m15.1
113
Eg-m78
64
Eg-m30

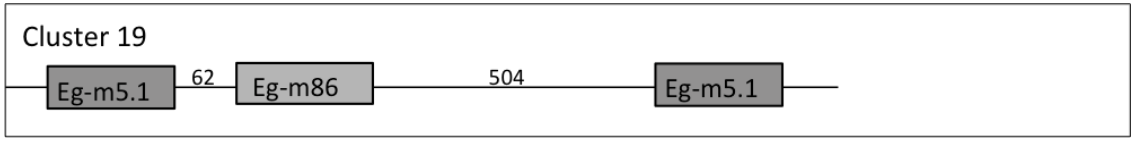

Cluster 10
Eg-m53
67
Eg-m21.3
100
Eg-m89
336
Eg-m14.2
oAM28
623
oAR234
Eg-m14
119
Eg-m14.3
335
Eg-m53
Cluster 10.1
Eg-m21.2
92
Eg-m89
292
Eg-m14.2
624
oAR233
Eg-m14
127
Eg-m14.4
339
Eg-m53
67
oAR191
Eg-m21.2
Cluster 10.2
Eg-m21
92
Eg-m89
292
Eg-m14.1
//
Eg-m53
67
Eg-m21
Cluster 10.3
Eg-m21
92
Eg-m89.1
//
Eg-m53
67
Eg-m21

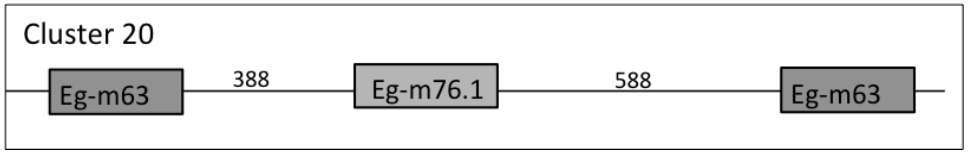

Cluster 11
Eg-m31.2
87
Eg-p10
210
Eg-m31.2
Cluster 11.1
oAR144
Eg-m31.2
324
oAM52
Eg-p10.1
187
Eg-m31.2
Cluster 11.2
Eg-m31.2
561
Eg-p10.2
211
Eg-m31.2
Cluster 11.3
oAM81
Eg-p10
210
Eg-m31.5
87
oAM82
Eg-p10
Cluster 11.4
Eg-p10
187
Eg-m31.3
337
Eg-p10
Cluster 11.5
Eg-p10
187
Eg-m31.4
333
Eg-p10

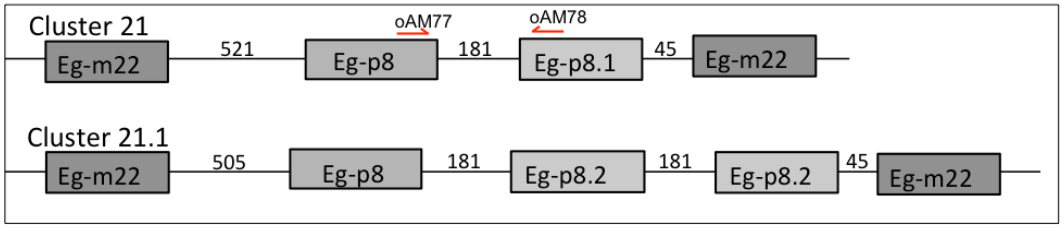

Cluster 12
Eg-m38
374
Eg-m55.2
176
Eg-m72
212
Eg-m38
Cluster 12.1
Eg-m38
409
Eg-m55.3
128
Eg-m38
Cluster 12.2
Eg-m55
177
Eg-m72
214
Eg-m38.1
381
Eg-m55
Cluster 12.3
Eg-m55.1
177
Eg-m72.2
217
Eg-m90.1
381
Eg-m55.1
Cluster 12.4
oAR186
oAR126
Eg-m55.1
89
Eg-m72.3
213
Eg-m38
381
Eg-m55.1
Cluster 12.5
Eg-m55.1
167
Eg-m72.1
217
Eg-m90
381
Eg-m55.1
Cluster 12.6
Eg-m55.1
174
Eg-m72.1
214
Eg-m55.1

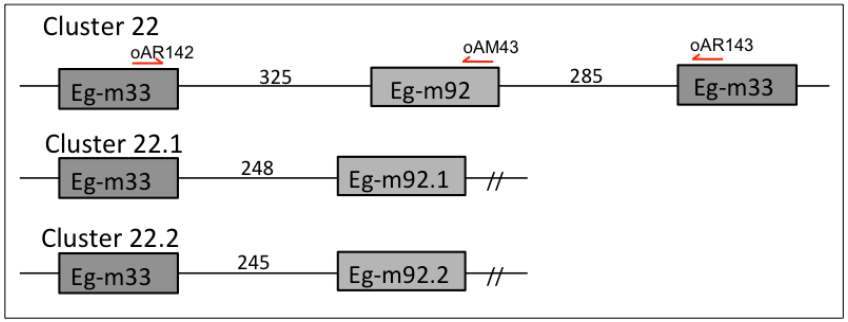

Cluster 13
Eg-m25
324
Eg-m39.2
//
Eg-m25
Cluster 13.1
Eg-m25
297
Eg-m39.3
//
Eg-m42
225
Eg-m25
Cluster 13.2
Eg-m39
512
Eg-m93
Eg-m42.2
230
Eg-m25
295
Eg-m39

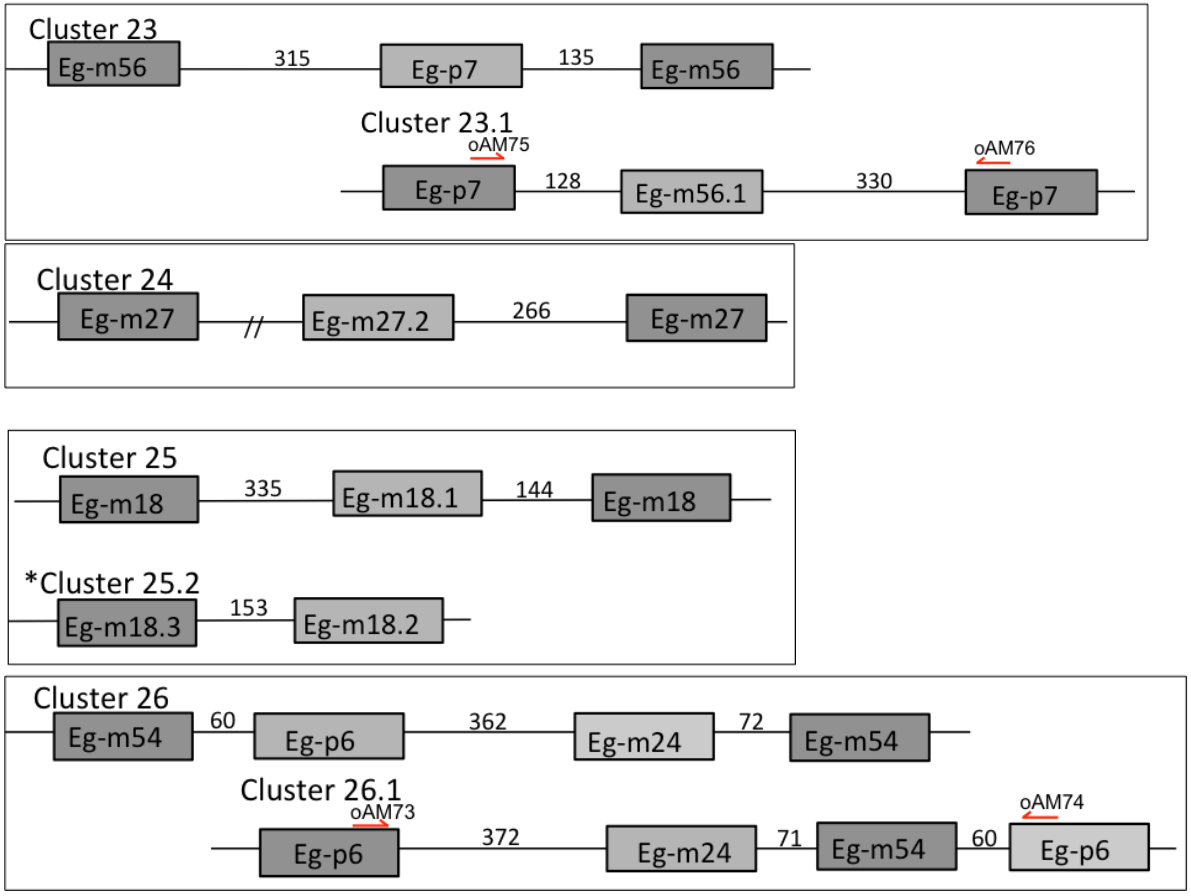

Cluster 14
Eg-m65
//
Eg-m48
100
Eg-m65
Cluster 14.1
oAR231
oAM49
Eg-m48
100
Eg-m65.1
180
Eg-m98
555
Eg-m48
Cluster 15
Eg-m10
345
Eg-m43.1
332
Eg-m10
Cluster 16
Eg-m8
118
Eg-m75.1
//
Eg-m8
Cluster 16.1
Eg-m8
119
Eg-m75
//
Eg-m76
247
Eg-m8
Cluster 17
Eg-m12
280
Eg-m85
402
Eg-m12
*Cluster 17.1
Eg-m12
275
Eg-m85.1
Cluster 18
oAR109
oAR110
Eg-m66
202
Eg-m91
64
Eg-m66.1
106
Eg-m91.1
88
Eg-m66

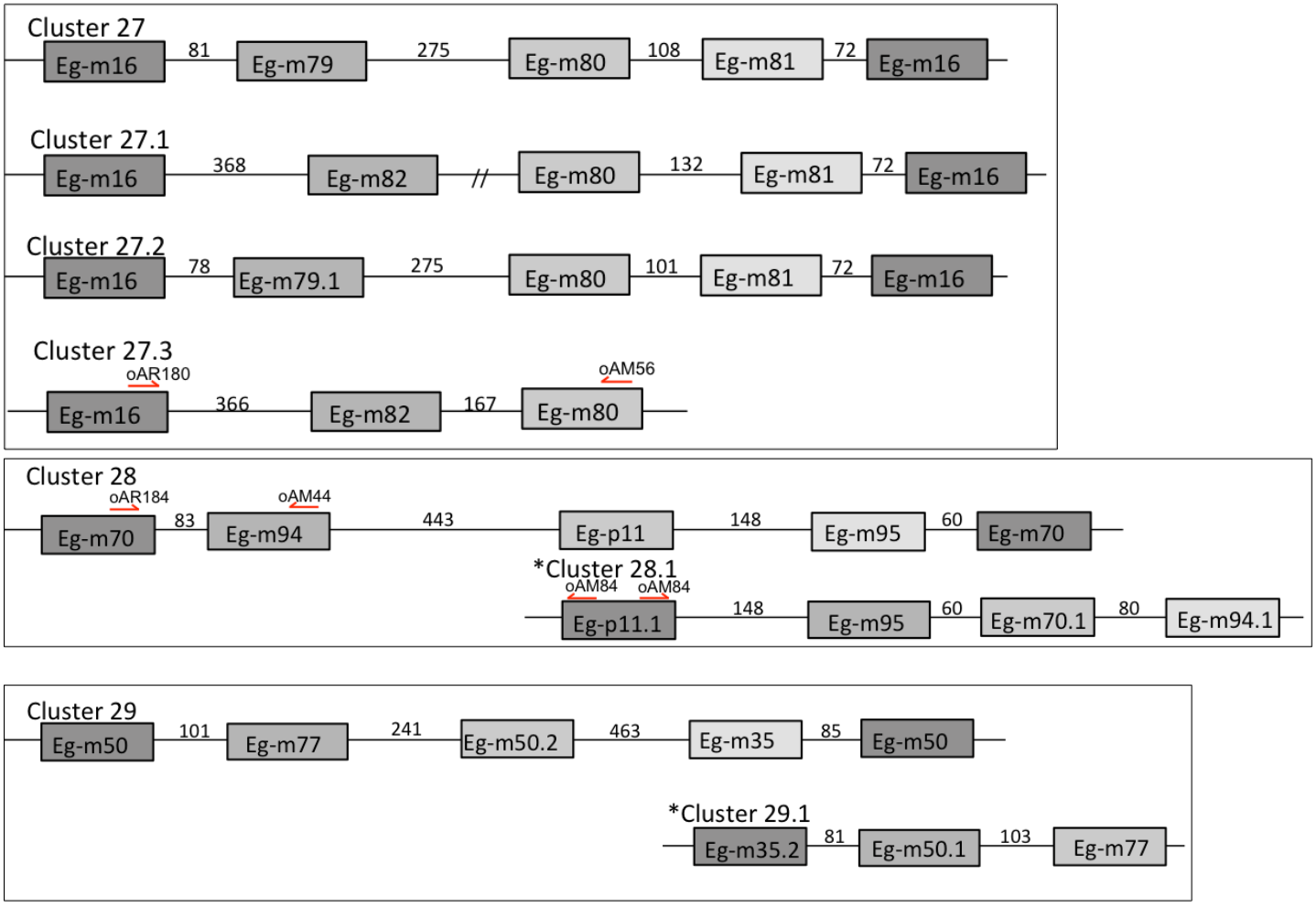

Cluster 19
Eg-m5.1
62
Eg-m86
504
Eg-m5.1

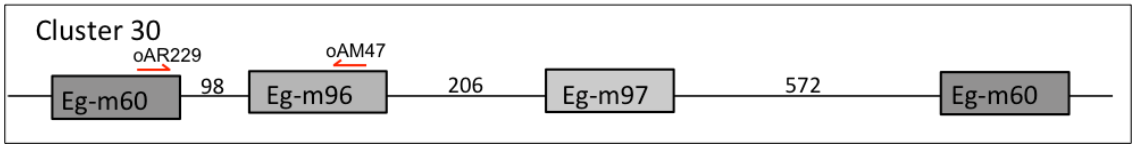

Cluster 20
Eg-m63
388
Eg-m76.1
588
Eg-m63

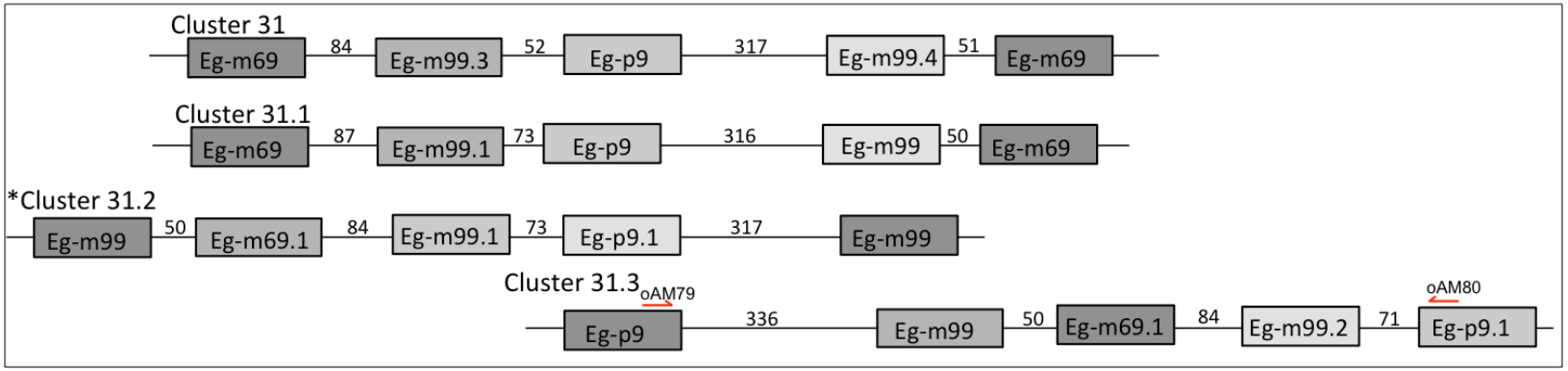

Cluster 21
Eg-m22
521
oAM77
Eg-p8
181
oAM78
Eg-p8.1
45
Eg-m22
Cluster 21.1
Eg-m22
505
Eg-p8
181
Eg-p8.2
181
Eg-p8.2
45
Eg-m22

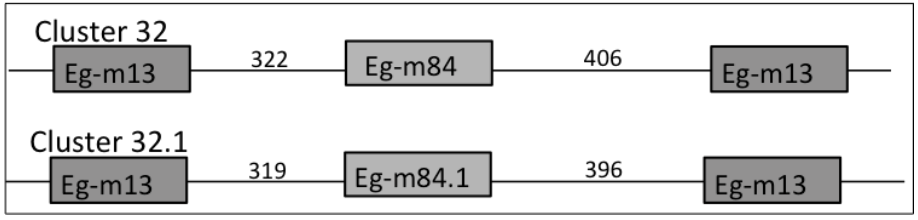

Cluster 22
oAR142
Eg-m33
325
oAM43
Eg-m92
285
oAR143
Eg-m33
Cluster 22.1
Eg-m33
248
Eg-m92.1
//
Cluster 22.2
Eg-m33
245
Eg-m92.2
//


Cluster 23
Eg-m56
315
Eg-p7
135
Eg-m56
Cluster 23.1
oAM75
oAM76
Eg-p7
128
Eg-m56.1
330
Eg-p7
Cluster 24
Eg-m27
//
Eg-m27.2
266
Eg-m27
Cluster 25
Eg-m18
335
Eg-m18.1
144
Eg-m18
*Cluster 25.2
Eg-m18.3
153
Eg-m18.2
Cluster 26
Eg-m54
60
Eg-p6
362
Eg-m24
72
Eg-m54
Cluster 26.1
oAM73
oAM74
Eg-p6
372
Eg-m24
71
Eg-m54
60
Eg-p6


Cluster 27
Eg-m16
81
Eg-m79
275
Eg-m80
108
Eg-m81
72
Eg-m16
Cluster 27.1
Eg-m16
368
Eg-m82
//
Eg-m80
132
Eg-m81
72
Eg-m16
Cluster 27.2
Eg-m16
78
Eg-m79.1
275
Eg-m80
101
Eg-m81
72
Eg-m16
Cluster 27.3
oAR180
Eg-m16
366
Eg-m82
167
oAM56
Eg-m80
Cluster 28
oAR184
oAM44
Eg-m70
83
Eg-m94
443
Eg-p11
148
Eg-m95
60
Eg-m70
*Cluster 28.1
oAM84
oAM84
Eg-p11.1
148
Eg-m95
60
Eg-m70.1
80
Eg-m94.1
Cluster 29
Eg-m50
101
Eg-m77
241
Eg-m50.2
463
Eg-m35
85
Eg-m50
*Cluster 29.1
Eg-m35.2
81
Eg-m50.1
103
Eg-m77


Cluster 30
oAR229
oAM47
Eg-m60
98
Eg-m96
206
Eg-m97
572
Eg-m60


Cluster 31
Eg-m69
84
Eg-m99.3
52
Eg-p9
317
Eg-m99.4
51
Eg-m69
Cluster 31.1
Eg-m69
87
Eg-m99.1
73
Eg-p9
316
Eg-m99
50
Eg-m69
*Cluster 31.2
Eg-m99
50
Eg-m69.1
84
Eg-m99.1
73
Eg-p9.1
317
Eg-m99
Cluster 31.3
oAM79
oAM80
Eg-p9
336
Eg-m99
50
Eg-m69.1
84
Eg-m99.2
71
Eg-p9.1


Cluster 32
Eg-m13
322
Eg-m84
406
Eg-m13
Cluster 32.1
Eg-m13
319
Eg-m84.1
396
Eg-m13


*Cluster 33
Eg-m47
61
Eg-p4
40
Eg-p12

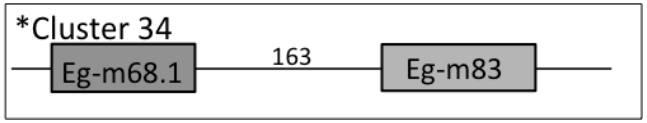
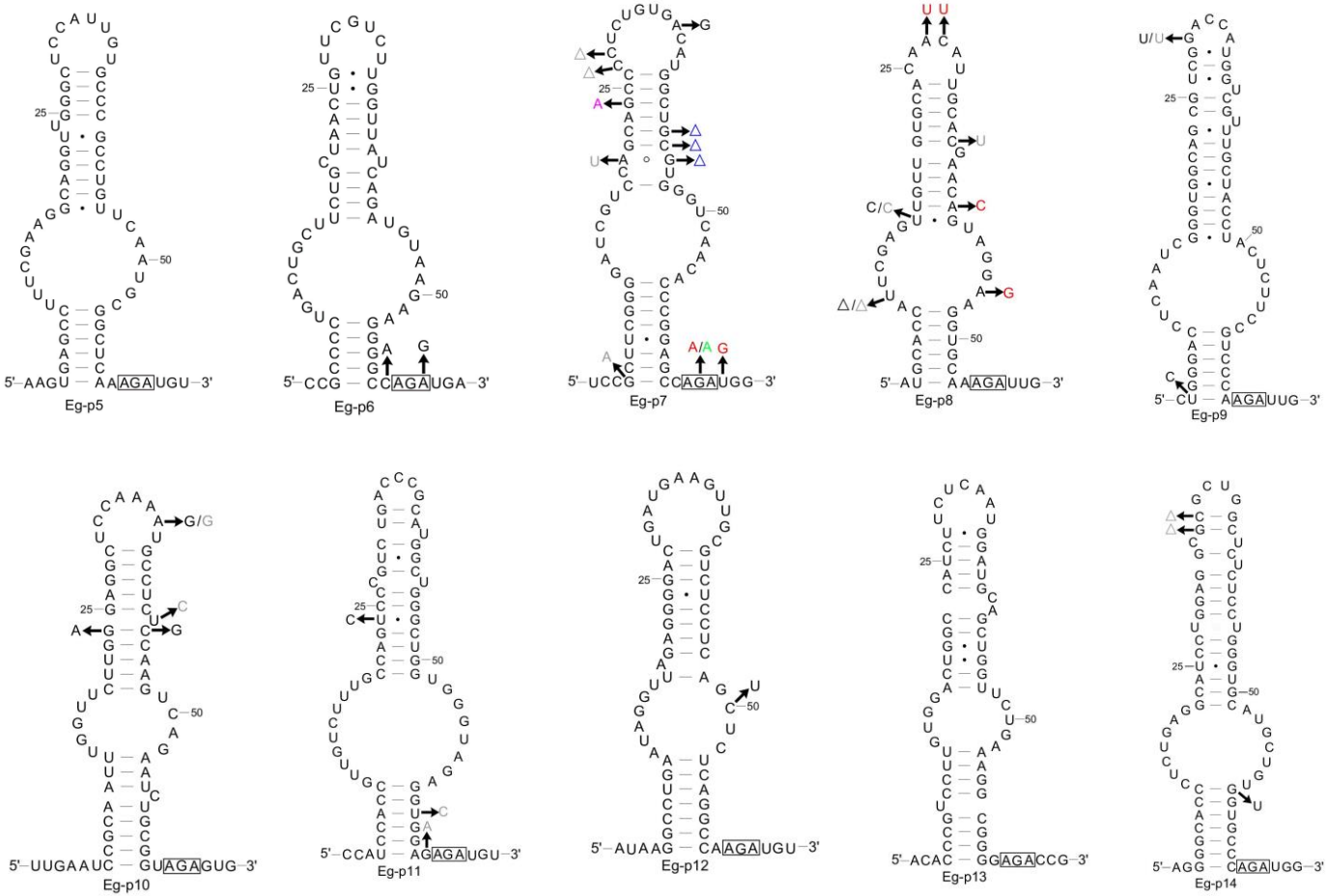
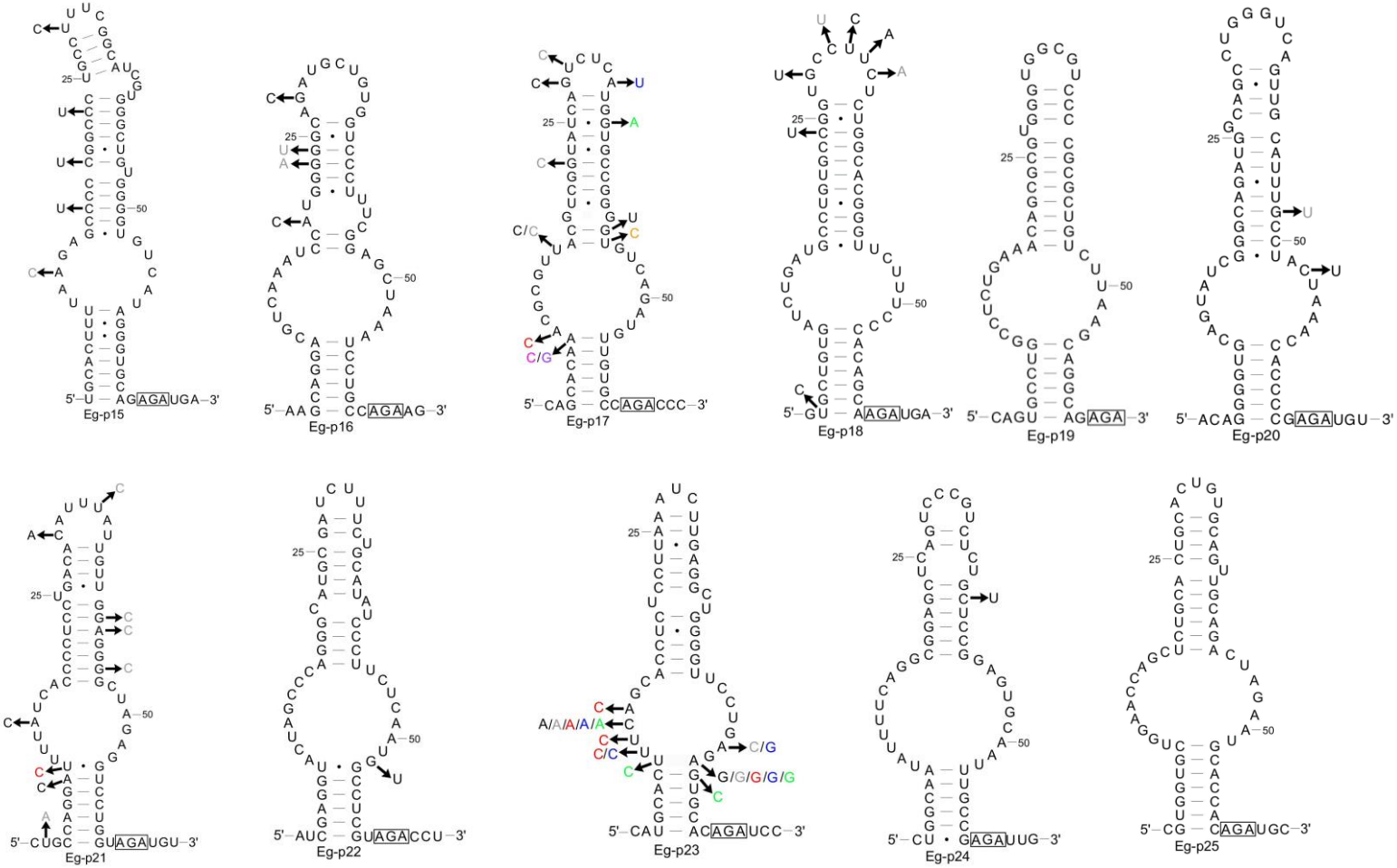
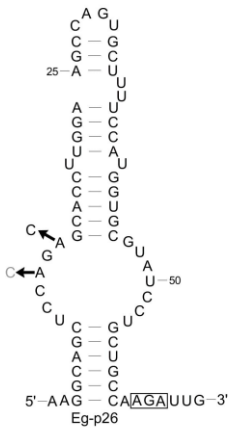
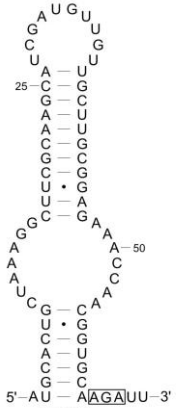

*Cluster 34
Eg-m68.1
163
Eg-m83

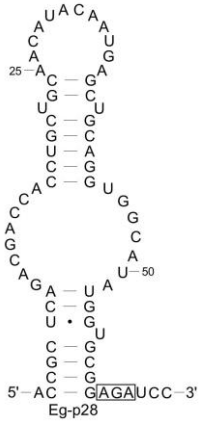
Figure S3. Predicted secondary structures of identified *E. gracilis* box AGA modification guide snoRNAs. In the structures, the boxed nucleotides highlight the AGA box elements and arrows indicate sequence variation between isoforms. Each color represents the nucleotide changes present in a single isoform. Black = Eg-p#.1; Grey = Eg-p#.2; Red = Eg-p#.3; Blue = Eg-p#.4; Green = Eg-p#.5; Pink = Eg-p#.6; Purple = Eg-p#.7; Orange = Eg-p#.8

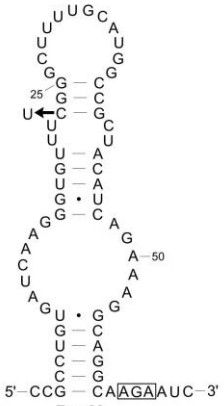


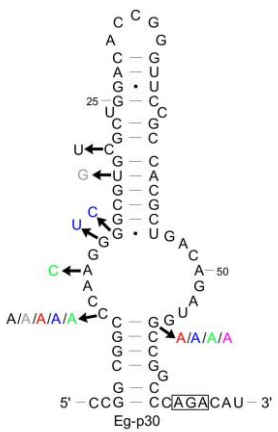

5'–U – AGAGAUGA–3'
Eg-p15
5'–AAG – CCAGAAG–3'
Eg-p16
5'–CAG – CCAGACCC–3'
Eg-p17
5'–GU – AAGAUGA–3'
Eg-p18
5'–CAGU – AGAGA –3'
Eg-p19
5'–ACAG – CGAGAUGU–3'
Eg-p20
5'–CUGC – GUAGAUGU–3'
Eg-p21
5'–AUC – GUAGACCU–3'
Eg-p22
5'–CAU – ACAGAUCC–3'
Eg-p23
5'–CU • GAGAUUG–3'
Eg-p24
5'–CG – CAGAUGC–3'
Eg-p25

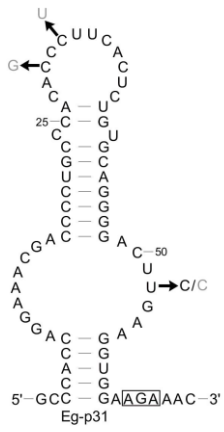

5'–AAG–CAAGAUUG–3'
Eg-p26

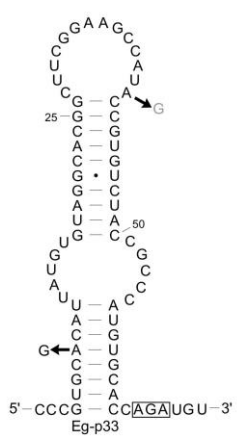

5'–AU–AAGAUU–3'
Eg-p27

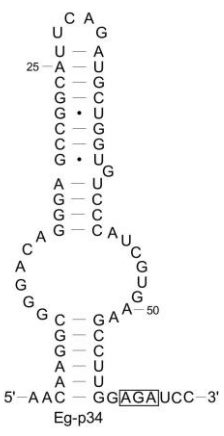

5'–AC–GAGAUCC–3'
Eg-p28

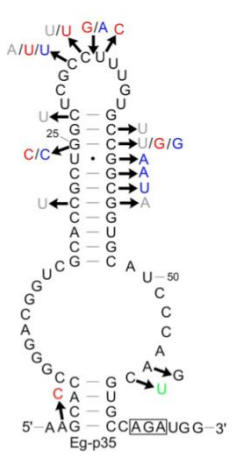

5'–CCG–CAAGAUC–3'
Eg-p29

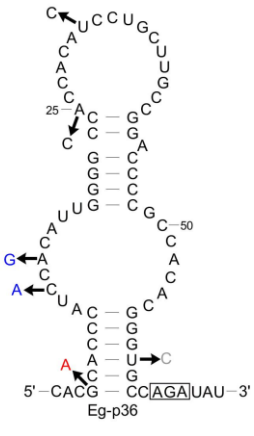

5'–CCG–CCAGACAU–3'
Eg-p30

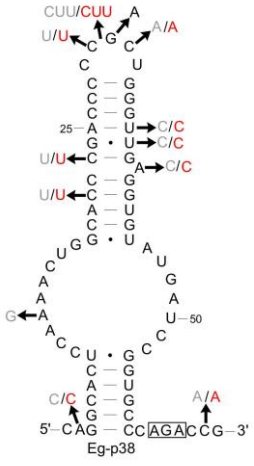

5'–GCC–GAAGAAAC–3'
Eg-p31


5'–CCCG–CCAGAUGU–3'
Eg-p33


5'–AAC–GGAGAUCC–3'
Eg-p34


5'–AAG–CCAGAUGG–3'
Eg-p35


5'–CACG–CCAGAUAU–3'
Eg-p36


5'–CAG–CCAGACCG–3'
Eg-p38

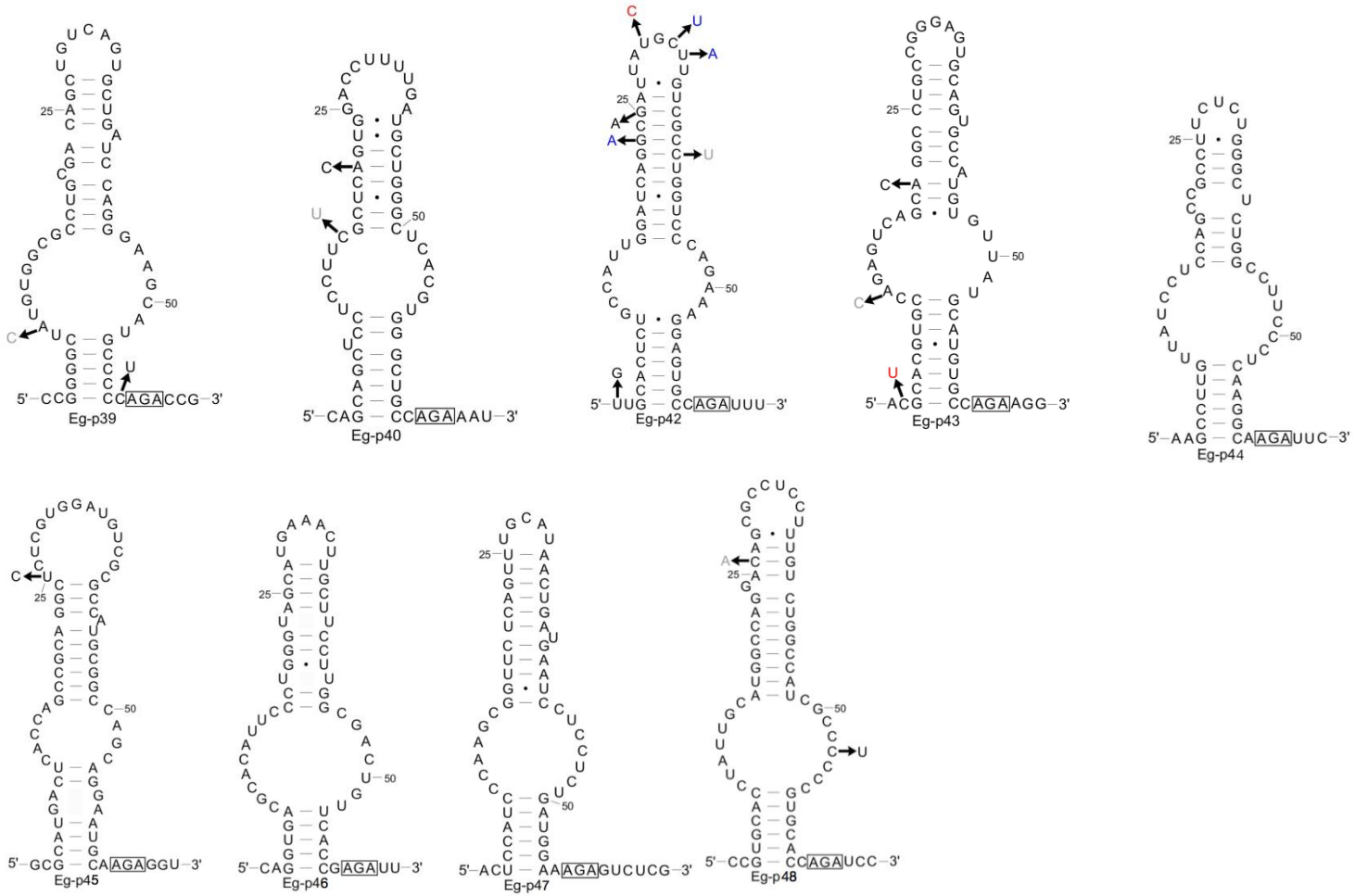

5'–CCG – CCAGACCG–3'
Eg-p39
5'–CAG – CCAGAAAU–3'
Eg-p40
5'–UUG – CCAGAUUU–3'
Eg-p42
5'–ACG – CCAGAAGG–3'
Eg-p43
5'–AAG – CAAGAUUC–3'
Eg-p44
5'–GCG – CAAGAGGU–3'
Eg-p45
5'–CAG – CGAGAUU–3'
Eg-p46
5'–ACU – AAAGAGUCUCG–3'
Eg-p47
5'–CCG – CCAGAUCC–3'
Eg-p48

Figure S4. ClustalW sequence alignments of snoRNA gene repeats characterized in the BAC genomic DNA library. After complete digestion with a restriction enzyme that cleaves within each gene repeat, the product was cloned and a number of individual colonies were sequenced. The aligned sequences are the unique variants of the gene repeats.

Clustal alignment of the 14 variants of snoRNA gene Cluster 23, produced after digestion with HindIII. Yellow highlighted nucleotides represent the coding region for Eg-m56 and the coding region for Eg-p7 is highlighted in green.

```
1                AGCTTTGGCCAGCTGGCGACGGCTTGGAACCCTCGTTGCCGAACAGCCATGGGAGTGGGG 60
3                AGCTTTGGCCAGCTGGCTACGGCTTGGAACCCTCGTTGCCCAACAGCCATGGGAGTGGGG 60
7                AGCTTTGGCCAGCTGGCTACGGCTTGGAACCCTCGTTGCCGAACAGCCATGGGAGTGGGG 60
14               AGCTTTGGCCAGCTGGCTACGGCTTGGAACCCTCGTTGCCGAACAGCCATGGGAGTGGGG 60
2                AGCTTTGGCCAGCTGGCTTCGGCTTGGAACCCTCGTTGCCGAACAGCCATGGGAGTGGGG 60
4                AGCTTTGGCCAGCTGGCTACGGCTTGGAACCCTCGTTGCCGAACAGCCATGGGAGTGGGG 60
9                AGCTTTGGCCAGCTGGCTACGGCTTGGAACCCTCGTTGCCGAACAGCCATGGGAGTGGGG 60
13               AGCTTTGGCCAGCTGGCTACGGCTTGGAACCCTCGTTGCCGAACAGCCATGGGAGTGGGG 60
6                AGCTTTGGCCAGCTGGCTACGGCTTGGAACCCTCGTTGCCGAACAGCCATGGGAGTGGGG 60
12               AGCTTTGGCCAGCTGGCTACGGCTTGGAACCCTCGTTGCCGAACAGCCATGGGAGTGGGG 60
5                AGCTTTGGCCAGCTGGCTACGGCTTGGAACCCTCGTTGCCGAACAGCCATGGGAGTGGGG 60
8                AGCTTTGGCCAGCTGGCTACGGCTTGGAACCCTCGTTGCCGAACAGCCATGGGAGTGGGG 60
10               AGCTTTGGCCAGCTGGCTACGGCTTGGAACCCTCGTTGCCGAACAGCCATGGGAGTGGGG 60
11               AGCTTTGGCCAGCTGGCTACGGCTTGGAACCCTCGTTGCCGAACAGCCATGGGAGTGGGG 60
                 ****************  ********************* ******************

1                G-AGGAACGCGTAGAGCCGCACACTTTGGGGATGACGACATTGTTACTCTCATGCGTTGC 119
3                G-AGGAACGCGTAGAGCCGCACACTTTGGGGATGACGACATTGTTACTCTCATGCGTTGC 119
7                G-AGGAACGCGTAGAGCCGCACACTTTGGGGATGACGACATTGTTACTCTCATGCGTTGC 119
14               G-AGGAACGAGTAGAGCCGCACACTTTGGGGATGACGACATTGTTACTCTCATGCGTTGC 119
2                G-AGGAACGCGTAGAGCCGCACACTTTGGGGATGACGACATTGTTACTCTCATGCGTTGC 119
4                G-AGGAACGCATAGAGCCGCACACTTTGGGGATGACGACATTGTTACTCTCATGCGTTGC 119
9                G-AGGAACGCGTAGAGCCGCACACTTTGGGGATGACGACATTGTTACTCTCATGCGTTGC 119
13               G-AGGAACGCGTAGAGCCGCACACTTTGGGGATGACGACATTGTTACTCTCATGCGTTGC 119
6                G-AGGAACGCGTAGAGCCGCACACTTTGGGGATGACGACATTGTTACTCTCATGCGGTGC 119
12               G-AGGAACGCGTAGAGCCGCACACTTTGGGGATGACGACATTGTTACTCTCATGCGTTGC 119
5                G-AGGAACGCGTAGAGCCGCAAACTTTGGGGATGACGACATTGTTACTCTCATGCGTTGC 119
```

```
8               G-AGGAACGCGTAGAGCCGCACACTTTGGGGATGACGACATTGTTACTCTCATGCGTTGC 119
10              GGAGGAACGCGTAGAGCCGCACACTTTGGGGATGACGACATTGTTACTCTCATGCGTTGC 120
11              G-AGGAACGCGTAGAGCCGCACACTTTGGGGATGACGACATTGTTACTCTCATGCGTTGC 119
                * *******  ********** ********************************** ***

1               TGAGGCCCACTGATGTGGTATCGCCAGCTGACCCGTTATCCCCTCCCGACCGCTGGGTGG 179
3               TGAGGCCCACTGATGTGGTATCGCCAGCTGACCCGTTATCCCCTCCCGACCGCTGGGTGG 179
7               TGAGGCCCACTGATGTGGTATCGCCAGCTGACCCGTTATCCCCTCCCGACCGCTGGGTGG 179
14              TGAGGCCCACTGATGTGGTATCGCCAGCTGACCCGTTATCCCCTCCCGACCGCTGGGTGG 179
2               TGAGGCCCACTGATGTGGTATCGCCAGCTGACCCGTTATCCCCTCCCGACCGCTGGGTGG 179
4               TGAGGCCCACTGATGTGGTATCTCCAGCTGACCCGTTATCCCCTCCCGACCGCTGGGTGG 179
9               TGAGGCCCACTGATGTGGTATCGCCAGCTGACCCGTTATCCCCTCCCGACCGCTGGGTGG 179
13              TGAGGCCCACTGATGTGGTATCGCCAGCTGACCCGTTATCCCCTCCCGACCGCTGGGTGG 179
6               TGAGGCCCACTGATGTGGTATCGCCAGCTGACCCGTTGTCCCCTCCCGACCGCTGGGTGG 179
12              TGAGGCCCACTGATGTGGTATCGCCAGCTGACCCGTTATCCCCTCCCGACCGCTGGGTGG 179
5               TGAAGCCCACTGATGTGGTATCGCCAGCTGACCCGTTATCCCCTCCCGACCGCTGGGTGG 179
8               TGAGGCCCACTGATGTGGTATCGCCAGCTGACCCGTTATCCCCTCCCGACCGCTGGGTGG 179
10              TGAGGCCCACTGATGTGGTATCGCCAGCTGACCCGTTATCCCCTCCCGACCGCTGGGTGG 180
11              TGAGGCCCACTGATGTGGTATCGCCAGCTGACCCGTTATCCCCTCCCGACCGCTGGGTGG 179
                *** ****************** ************** ********************

1               CTGGTGGGCAGGGCATCTTCCATGGCATGTGGTGTTGTGGGCTGTGGCGAGGGATGGCTT 239
3               CTGGTGGGCAGGGCATCTTCCATGGCATGTGGTGTTGTGGGCTGTGGCGAGGGATGGCTT 239
7               CTGGTGGGCAGGGCATCTTCCATGGCATGTGGTGTTGTGGGCTGTGGCGAGGGATGGCTT 239
14              CTGGTGGGCAGGGCATGTTCCATGGCATGTGGTGTTGTGGGCTGTG-CGAGGGATGGCTT 238
2               CTGGTGGGCAGGGCATGTTCCATGGCATGTGGTGTTGTGGGCTGTGGCGAGGGATGGCTT 239
4               CTGGTGGGCAGGGCATGTTCCATGGCATGTGGTGTTGTGGGCTGTGGCGAGGGATGGCTT 239
9               CTGGTGGGCAGGGCATGTTCCATGGCATGTGGTGTTGTGGGCTGTGGCGAGGGATGGCTT 239
13              CTGGTGGGCAGGGCATGTTCCATGGCATGTGGTGTTGTGGGCTGTGGCGAGGGATGGCTT 239
6               CTGGTGGGCAGGGCATGTTCCATGGCATGTGGTGTTGTGGGCTGTGGCGAGGGATGGCTT 239
12              CTGGTGGGCAGGGCATGTTCCATGGCATGTGGTGTTGTGGGCTGTGGCGAGGGATGGCTT 239
5               CTGGTGGGCAGGGCATGTTCCATGGCATGTGGTGTTGTGGGCTGTGGCGAGGGATGGCTT 239
8               CTGGTGGGCAGGGCATGTTCCATGGCATGTGGTGTTGTGGGCTGTGGCGAGGGATGGCTT 239
10              CTGGTGGGCAGGGCATGTTCCATGGCATGTGGTGTTGTGGGCTGTGGCGAGGGATGGCTT 240
```

```
11              CTGGTGGGCAGGGCATGTTCCATGGCATGTGGTGTTGTGGGCTGTGGCGAGGGATGGCTT 239
                **************** ***************************** ************

1               GCGGGCACCAGAAAACCGGCTCGTCGCTTGCACCCCACCCTGGAGGAGGCCACAGGGCTG 299
3               GCGGGCACCAGAAAACCGGCTCGTCGCTTGCACCCCACCCTGGAGGAGGCCACAGGGCTG 299
7               GCGGGCACCAGAAAACCGGCTCGTCGCTTGCACCCCACCCTGGAGGAGGCCACAGGGCTG 299
14              GCGGGCACCAGAAAACCGGCTCGTCGCTTGCACCCCACCCTGGAGGAGGCCACAGGGCTG 298
2               GCGGGCACCAGAAAACCGGCTCACCGCTTGCACCCCACCCTGGAGGAGGCCACAGGGCTG 299
4               GCAGGCACCAGAAAACCGGCTCACCGCTTGCACCCCACCCTGGAGGAGGCCACAGGGCTG 299
9               GCGGGCACCAGAAAACCGGCTCACCGCTTGCACCCCACCCTGGAGGAGGCCACAGGGCTG 299
13              GCGGGCACCAGAAAACCGGCTCACCGCTTGCACCCCACCCTGGAGGAGGCCACAGGGCTG 299
6               GCGGGCACCAGAAAACCGGCTCACCGCTTGCACCCCACCCTGGAGGAGGCCACAGGGCTG 299
12              GCGGGCACCAGAAAACCGGCTCACCGCTTGCACCCCACCCTGGAGGAGGCCACAGGGCTG 299
5               GCGGGCACCAGAAAACCGGCTCACCGCTTGCACCCCACCCTGGAGGAGGCCACAGGGCTG 299
8               GCGGGCACCAGAAAACCGGCTCACCGCTTGCACCCCACCCTGGAGGAGGCCACAGGGCTG 299
10              GCGGGCACCAGAAAACCGGCTCACCGCTTGCACCCCACCCTGGAGGAGGCCACAGGGCTG 300
11              GCGGGCACCAGAAAACCGGCTCACCGCTTGCACCCCACCCTGGAGGAGGCCACAGGGCTG 299
                ** ******************  ************************************

1               GAGTGGCTGCGATGGGTGTTTTCTCAGCAGTGCCGTGCACTGGGGTCCGTTGGCCATCTG 359
3               GTGTGGCTGCGATGGGTGTTTTCTCAGCAGTGCCGTGCACTGGGGTCCGTTGGCCATCTG 359
7               GAGTGGCTGCGATGGGTGTATTCTCAGCAGTGCCGTGCACTGGGGTCCGTTGGCCATCTG 359
14              GAGTGGCTGCGATGGGTGTTTTCTCAGCAGTGCCGTGCACTGGGGTCCGTTGGCCATCTG 358
2               GAGTGGCTGCGATGGGTGTTTTCTCAGCAGTGCCGTGCACTGGGGTCCGTTGGCCATCTG 359
4               GAGTGGCTGCGATGGGTGTTTTCTCAGCAGTGCCGTGCACTGGGGTCCGTTGGCCATCTG 359
9               GAGTGGCTGCGATGGGTGTTTTCTCAGCAGTGCCGTGCACTGGGGTCCGTTGGCCATCTG 359
13              GAGTGGCTGCGATGGGTGTTTTCTCAGCAGTGCCGTGCACTGGGGTCCGTTGGCCATCTG 359
6               GAGTGGCTGCGATGGGTGTTTTCTCAGCAGTGCCGTGCACTGGGGTCCGTTGGCCATCTG 359
12              GAGTGGCTGCGATGGGTGTTTTCTCAGCAGTGCCGTGCACTGGGGTCCGTTGGCCATCTG 359
5               GAGTGGCTGCGATGGGTGTTTTCTCAGCAGTGCCGTGCACTGGGGTCCGTTGGCCATCTG 359
8               GTGTGGCTGCGATGGGTGTTTTCTCAGCAGTGCCGTGCACTGGGGTCCGTTGGCCATCTG 359
10              GTGTGGCTGCGATGGGTGTTTTCTCAGCAGTGCCGTGCACTGGGGTCCGTTGGCCATCTG 360
11              GTGTGGCTGCGATGGGTGTTTTCTCAGCAGTGCCGTGCACTGGGGTCCGTTGGCCATCTG 359
                * ***************** ***************************************
```

```
1               GCCCAGCATTGGCCGGACTGGTTTTTGGGGAGGGCGGTGGCACGGGGGCTTATGGGCCTG 419
3               GCCCAGCATTGGCCGGACTGGTTTTTGGGGAGGGCGGTGGCACGGGGGCTTATGGGCCTG 419
7               GCCCAGCATTGGCCGGACTGGTTTTTGGGGAGGGCGGTGGCACGGGGGCTTATGGGCCTG 419
14              GCCCAGCATTGGCCGGACTGGTTTTTGGGGAGGGCGGTGGCACGGGGGCTTATGGGCCTG 418
2               GCCCAGCATTGGCCGGACTGGTTTTTGGGGAGGGCGGTGGCACGGGGGCTTATGGGCCTG 419
4               GCCCAGCATTGGCCGGACTGGTTTTTGGGGAGGGCGGTGGCACGGGGGCTTATGGGCCTG 419
9               GCCCAGCATTGGCCGGACTGGTTTTTGGGGAGGGCGGTGGCACGGGGGCTTATGGGCCTG 419
13              GCCCAGCATTGGCCGGACTGGTTTTTGGGGAGGGCGGTGGCACGGGGGCTTATGGGCCTG 419
6               GCCCAGCATTGGCCGGACTGGTTTTTGGGGAGGGCGGTGGCACGGGGGCTTATGGGCCTG 419
12              GCCCAGCATTGGCCGGACTGGTTTTTGGGGAGGGCGGTGGCACGGGGGCTTATGGGCCTG 419
5               GCCCAGCATTGGCCGGACTGGTTTTTGGGGAGGGCGGTGGCACGGGGGCTTATGGGCCTG 419
8               GCCCAGCATTGGCCGGACTGGTTTTTGGGGAGGGCGGTGGCACGGGGGCTTATGGGCCTG 419
10              GCCCAGCATTGGCCGGACTGGTTTTTGGGGAGGGCGGTGGCACGGGGGCTTATGGGCCTG 420
11              GCCCAGCATTGGCCGGACTGGTTTTTGGGGAGGGCGGTGGCACGGGGGCTTATGGGCCTG 419
                ************************************************************

1               GCGCCAGTGATGTCCTTCCCTCTTTGGTGGCATTTCTGTGGTGCAACGTCCGCTTCGGGG 479
3               GCGCCAGTGATGTCCTTCCCTCTTTGGTGGCATTTCTGTGGTGCAATGTCCGCTTCGGGG 479
7               GCGCCAGTGATGTCCTTCCCTCTTTGGTGGCATTTCTGTGGTGCAACGTCCGCTTCGGGG 479
14              GCGCCAGTGATGTCCTTCCCTCTTTGGTGGCATTTCTGTGGTGCAACGTCCGCTTCGGGG 478
2               GCGCCAGTGATGTCCTTCCCTCTTTGGTGGCATTTCTGTGGTGCAATGTCCGCTTCGGGG 479
4               GCGCCAGTGATGTCCTTCCCTCTTTGGTGGCATTTCTGTGGTGCAATGTCCGCTTCGGGG 479
9               GCGCCAGTGATGTCCTTCCCTCTTTGGTGGCAGTTCTGTGGTGCAACGTCCGCTTCGGGG 479
13              GCGCCAGTGATGTCCTTCCCTCTTTGGTGGCAGTTCTGTGGTGCAACGTCCGCTTCGGGG 479
6               GCGCCAGTGATGTCCTTCCCTCTTTGGTGGCATTTCTGTGGTGCAACGTCCACTTCGGGG 479
12              GCGCCAGTGATGTCCTTCCCTCTTTGGTGGCATTTCTGTGGTGCAACGTCCGCTTCGGGG 479
5               GCGCCAGTGATGTCCTTCCCTCTTTGGTGGCATTTCTGTGGTGCAACGTCCGCTTCGGGG 479
8               GCGCCAGTGATGTCCTTCCCTCTTTGGTGGCATTTCTGTGGTGCAACGTCCGCTTCGGGG 479
10              GCGCCAGTGATGTCCTTCCCTCTTTGGTGGCATTTCTGTGGTGCAACGTCCGCTTCGGGG 480
11              GCGCCAGTGATGTCCTTCCCTCTTTGGTGGCATTTCTGTGGTGCAACGTCCGCTTCGGGG 479
                ******************************** ************ **** ********
```

```
1               ATCGTCCAGCAGCCCCTCTGTGACATGGCT---TGGGTCAACACCCGGAGCCAGATGGGG 536
3               ATCGTCCAGCAACCCCTCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAGATGGGG 539
7               ATCGTCCAGCAGCCCCTCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAGATGGGG 539
14              ATCGTCCAGCAGCCCCTCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAGATGGGG 538
2               ATCGTCCAGCAGCCCCTCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAGATGGGG 539
4               ATCGTCCAGCAGCCCCTCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAGATGGGG 539
9               ATCGTCCAGCAGCCCCTCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAGATGGGG 539
13              ATCGTCCAGCAGCCCCTCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAGATGGGG 539
6               ATCGTCCTGCAGCC--TCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAGATGGGG 537
12              ATCGTCCAGCAGCCCCTCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAGATGGGG 539
5               ATCGTCCAGCAGCCCCTCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAGATGGGG 539
8               ATCGTCCAGCAGCCCCTCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAGATGGGG 539
10              ATCGTCCAGCAGCCCCTCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAAATGGGG 540
11              ATCGTCCAGCAGCCCCTCTGTGACATGGCTGCGTGGGTCAACACCCGGAGCCAAAGGGGG 539
                ******* *** **  ************** ******************** * ****

1               GTGTGACGAGACGGGGTGGTCACGTGCCCTGTTCC-CAGTGCTCGGAAGCT 586
3               GTGTGACGAGACGGGGTGGTCACGTGCCCTGTTCC-CAGTGCTCGGAAGCT 589
7               GTGTGACGAGATGGGGTGGTCACGTGCCCTGTTCC-CAGTGCTCGGAAGCT 589
14              GTGTGACGAGATGGGGTGGTCACGTGCCCTGTTCC-CAGTGCTCGGAAGCT 588
2               GTGTGACGAGATGGGGTGGTCACGTGCCCTGTTCC-CAGTGCTCGGAAGCT 589
4               GTGTGACGAGATGGGGTGGTCACGTGCCCTGTTCC-CAGTGCTTGGAAGCT 589
9               GTGTGACGAGATGGGGTGGTCACGTGCCCTGTTCC-CAGTGCTCGGAAGCT 589
13              GTGTGACGAGATGGGGTGGTCACGTGCCCTGTTCC-CAGTGCTCGGAAGCT 589
6               GTGTGACGAGACGGGGTGGTCACGTGCCCTGTTCC-CAGTGCTCGGAAGCT 587
12              GTGTGACGAGATGGGGTGGTCACGTGCCCTGTTCC-CAGTGCTCGGAAGCT 589
5               GTGTGACGAGATGGGGTGGTCACGTGCCCTGTTCC-CAGTGCTCGGAAGCT 589
8               GTGTGACGAGATGGGGTGGTCACGTGCCCTGTTCC-CAGTGCTCGGAAGCT 589
10              GTGTGACGAGATGGGGTGGTCACGTGCCCTGTTCC-CAGTGCTCGGAAGCT 590
11              GTGTGACGAGATGGCGTGGTCGAATACC-TGTTGAGCAGTGCTCGAAAGCT 589
                *********** ** ******   * ** ****   ******* * *****
```

Clustal alignment of the 33 variants of snoRNA gene Cluster 2, produced after digestion with SphI. Yellow highlighted nucleotides represent the coding region for Eg-p2, the coding region for Eg-m1 is highlighted in green and Eg-m2 is highlighted in blue.

```
14              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
15              ----------CGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 50
19              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
18              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
16              ------------------------------------------------------------
17              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGNNGCTCGTCGCCGGGGTGTGGAC 59
2               CTGGCTGCC-GG-TTGGCGAGCCGGGGGGTTGGGGCATGGCTCGTCGCCGGGGTGTGGAC 58
5               CTGCCTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGCATGGCTCGTCGCCGGGGTGTGGAC 59
7               CTGGCTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGCATGGCTCGTCGCCGGGGTGTGGAC 59
6               CTGGCTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGCATGGCTCGTCGCCGGGGTGTGGAC 59
3               CTGGGTGCC-GG-TTGGCGAGCCGGGGGGTTGGGGCATGGCTCGTCGCCGGGGTGTGGAC 58
4               CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGCATGGCTCGTCGCCGGGGTGTGGAC 59
8               CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGCATGGCTCGTCGCCGGGGTGTGGAC 59
9               CTTGGTGCCCGGGTTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 60
10              CTTGGTGCCCGGGTTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 60
11              CTTGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGCGGAC 59
32              CTGGGTGTCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
33              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
12              CTGGGTGCCCGG-TTGGCGAGCTGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
13              CTGGGTGCCCGG-TTGGCGAGCTGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
25              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
26              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
24              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
27              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
28              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
29              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
30              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
31              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
21              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
22              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
```

```
23              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
20              CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGTGTGGCTCGTCGCCGGGGTGTGGAC 59
1               CTGGGTGCCCGG-TTGGCGAGCCGGGGGGTTGGGGCATGGCTCGTCGCCGGGGTGTGGAC 59
14              GGACCTTGGCGGACCTTTCTGGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
15              GGACCTTGGCGGACCTTTCTGGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 110
19              GGACCTTGGCGGACCTTTCTGGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
18              GGACCTTGGCGGACCTTTCTGGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
16              ------------------CTGGGTGTTCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 42
17              GGACCTTGGCGGACCTTTCTCGGTGTTCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
2               GGACCTTGGCGGACCTTTCTCGGTGTTCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 118
5               GGACCTTGGCGGACCTTTCTCGGTGTTCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
7               GGACCTTGGCGGACCTTTCTCGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
6               GGACCTTGGCGGACCTTTCTCGGTGTCCCCGTCACGCTTGACGTGGGGAGCAACTGCAGG 119
3               GGACCTTGGCGGACCTTTCTGGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 118
4               GGACCTTGGCGGACCTTTCTGGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
8               GGACCTTGGCGGACCTTTCTGGGTGTCCCCGTCACGGTTGGCGTGGGGAGCAATTGCAGG 119
9               GGACCTTGGCGGACCTTTCTGGGTGTCCCCGTCACGCTTGACGTGGGGAGCAACTGCAGG 120
10              GGACCTTGGCGGACCTTTCTGGGTGTCCCCGTCACGCTTGACGTGGGGAGCAACTGCAGG 120
11              GGACCTTGGCGGACCTTTCTGGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
32              GGACCTTGGCGGACCTTTCTCGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
33              GGACCTTGGCGGACCTTTCTGGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
12              GGACCTTGGCGGACCTTTCTGGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
13              GGACCTTGGCGGACCTTTCTCGGTGTCCCCGTCACGCTTGACGTGGGGAGCAACTGCAGG 119
25              GGACCTTGGCGGACCTTTCTGGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
26              GGACCTTGGCGGACCTTTCTGGGTGTTCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
24              GGACCTTGGCGGACCTTTCTGGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
27              GGACCTTGGCGGACCTTTCTCGGTGTTCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
28              GGACCTTGGCGGACCTTTCTCGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
29              GGACCTTGGCGGACCTTTCTCGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
30              GGACCTTGGCGGACCTTTCTCGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
31              GGACCTTGGCGGACCTTTCTCGGTGTCCCCGTCACGCTTGACGTGGGGAGCAACTGCAGG 119
21              GGACCTTGGCGGACCTTTCTCGGTGTCCCCGTCACGCTTGACGTGGGGAGCAACTGCAGG 119
22              GGACCTTGGCGGACCTTTCTGGGTGTCCCCGTCACGGTTGGCGTGGGGAGCAACTGCAGG 119
23              GGACCTTGGCGGACCTTTCTGGGTGTTCCCGTCACGGTTGGCGTGGGGAGCAACTGCAGG 119
```

```
20              GGACCTTGGCGGACCTTTCTGGGTGTCCCCATCACGCTTGGTGTGGGGAGCAACTGCAGG 119
1               GGACCTTGGCGGACCTTTCTCGGTGTCCCCGTCACGCTTGGCGTGGGGAGCAACTGCAGG 119
                         ** ***** *** ***** ***  *********** ******

14              ACCGGGGGTGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAG-CAATGCGGCACGCA 178
15              ACCGGGGGTGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAG-CAATGCGGCACGCA 169
19              ACCGGGG-TGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 178
18              ACCGGGG-TGTGGCGGGGTGTCTCGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 178
16              ACCGGGG-TGTGGCGGGGTGTCTTGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 101
17              ACCGGGG-TGTGGCGGGGTGTCTTGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 178
2               ACCGGGGGTGTGGCGGGGTGTCTTGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 178
5               ACCGGGGGTGTGGCGGGGTGTCTTGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 179
7               ACCGGGGGTGTGGCGGGGTGTCTCGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 179
6               AC-GGGGGTGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 178
3               ACCGGGG-TGTGGCGGGGTGTCTCGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 177
4               ACCGGGG-TGTGGCGGGGTGTCTCGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 178
8               ACCGGGG-TGTGGCGGGGTGTCTCGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 178
9               ACCGGGG-TGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 179
10              ACCGGGG-TGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 179
11              ACCGGGG-TGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 178
32              ACGGGGG-TGTGGCGGGGTGTCTCGGGGGTGCGCAGTTTTTGCAGGCAATGCGGCACGCA 178
33              ACCGGGG-TGTGGCGGGGTGTCTCGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 178
12              ACCGGGGGTGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAG-CAATGCGGCACGCA 178
13              ACCGGGGGTGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAG-CAATGCGGCACGCA 178
25              ACCGGGGGTGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAG-CAATGCGGCACGCA 178
26              ACCGGGGGTGTGGCGGGGTGTTCCGGGGGTGCGCGGTTTTTGCAG-CAATGCGGCACGCA 178
24              ACCGGGGGGGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAG-CAATGCGGCACGCA 178
27              AC-GGGGGTGTG----------------------------------------CACGCA 136
28              AC-GGGGGTGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAG-CAATGCGGCACGCA 177
29              ACCGGGGGTGTGGCGGGGTGTCTCGGGGGTGCGCGGTTTTTGCAG-CAATGCGGCACGCA 178
30              ACCGGGGGTGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAG-CAATGCGGCACGCA 178
31              ACCGGGGGTGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAG-CAATGCGGCACGCA 178
21              AC-CGGGGTGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 178
22              AC-GGGGGTGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 178
```

```
23              AC-GGGGGTGTGGCGGGGTGTTCCGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 178
20              AC-GGGGGTGTGGCGGGGTGTCCCGGGG-TGCGCGGTTTTTGCAGGCAATGCGGCACGCA 177
1               AC-GGGGGTGTGGCGGGGTGTCCCGGGGGTGCGCGGTTTTTGCAGGCAATGCGGCACGCA 178
                **  ***  ***                                              ******

14              GTTT-GCAGCCC-------ATGCCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 226
15              GTTT-GCAGCCC-------ATGCCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 217
19              GTTT-GCAGCCC-------ATGCCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 226
18              GTTTTGCAGCCC-------ATGCCACTGCGTATACCCCGCT----GCCCAGCCTTGGCAG 227
16              GTTT-GCAGCCC-------ATGCCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 149
17              GTTT-GCAGCCC-------ATGCCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 226
2               GTTT-GCAGCCC-------ATGTCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 226
5               GTTT-GCAGCCC-------ATGTCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 227
7               GTTT-GCAGCCC-------ATGTCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 227
6               GTTT-GCAGCCC-------ATGTCACTGCGTATACCCCGCT----GCCCAGCCTTGGCAG 226
3               GTTT-GCAGCCC-------ATGCCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 225
4               GTTT-GCAGCCC-------ATGCCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 226
8               GTTT-GCAGCCC-------ATGCCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 226
9               GTTT-GCAGCCC-------ATGTCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 227
10              GTTT-GCAGCCC-------ATGTCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 227
11              GTTT-GCAGCCC-------ATGTCACTGCGTATACCCCGCT----GCCCAGCCTTGGCAG 226
32              GTCT-GCAGCCCATGAGCCCATTCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 233
33              GTTT-GCAGCCC-------ATGCCACTGCGTATATGCCGCT----GCCCAGCCTTGGCAG 226
12              GTTT-GCAGCCC-----------ACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 221
13              GTTT-GCAGCCC-----------ACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 221
25              GTTT-GCAGCCC-------ATGCCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 226
26              GTTT-GCAGCCC-------ATGCCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 226
24              GTTT-GCAGCCC-------ATGCCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 226
27              GTTT-GCAGCCC-------ATGCCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 184
28              GTTT-GCAGCCC-------ATGCCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 225
29              GTTT-GCAGCCC-------ATGCCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 226
30              GTTT-GCAGCCC-------ATGCCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 226
31              GTTT-GCAGCCC-------ATGCCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 226
21              GTTT-GTAGCCC-------ATGCCACTGCGTATACGCCGGTATATGCCCAGCCTTGGCAG 230
```

```
22          GTTT-GTAGCCC-------ATGTCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 226
23          GTTT-GTAGCCC-------ATGTCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 226
20          GTTTTGTAGCCC-------ATGTCACTGCGTATACGCCGCT----GCCCAGCCTTGGCAG 226
1           GTTT-GCAGCCC-------ATGCCACTGCGTATACCCCGCT----GCCCAGCCTTGGCAG 226
            ** * * *****            **********  *** *    ***************

14          CATGTCGTCCGAAAGGCATGGCAGACGCACGTG-----------TTGTTTTTTT---CA 271
15          CATGTCGTCCGAAAGGCATGGCAGACGCACGTG-----------TTGTTTTTTT---CA 262
19          CATGTCGTCCGAAAGGCATGGCAGACGCACGTG-----------TTGTTTTTTT---CA 271
18          CATGTCGTCCGAAAGGCATGGCAGACGCACGTG-----------TTGTTTTTTT---CA 272
16          CATGTCGTCCGAAAGGCATGGCAGACGCACGTG-----------TTGTTTTTTT---CA 194
17          CATGTCGTCCGAAAGGCATGGCAGACGCACGTG-----------TTGTTTTTTT---CA 271
2           CATGTCGTCCGAAAAGCATGGCAGACGCACGTG-----------TTGTTTTTTT---CA 271
5           CATGTCGTCCGAAAAGCATGGCAGACGCACGTG-----------TTGTTTTTTT---CA 272
7           CATGTCGTCCGAAAAGCATGGCAGACGCACGTG-----------TTGTTTTTTT---CA 272
6           CATGTCGTCCGAAAGGCATGGCAGACGCACGTG-----------TTGTTTTTTT---CA 271
3           CATGTCGTCCGAAAGGCATGGCAGACGCACGTG-----------TTGTTTTTTT---CA 270
4           CATGTCGTCCGAAAGGCATGGCAGACGCACGTG-----------TTGTTTTTTT---CA 271
8           CATGTCGTCCGAAAGGCATGGCAGACGCATGTG-----------TTGTTTTTTT---CA 271
9           CATGTCGTCCGAAAGGCATGGCAGACGCACGTG-----------TTGTTTTTTC---CA 272
10          CATGTCGTCCGAAAGGCATGGCAGACGCACGTG-----------TTGTTTTTTC---CA 272
11          CATGTCGTCCGAAAGGCATGGCAGACGCACGTG-----------TTGTTTTTTT---CA 271
32          CATGTCGTCCGAAAGGCATGGCAGACGCACGTG-----------TTGTTTTTTT---CA 278
33          CATGTCGTCCGAAAGGCATGGCAGACGCACGTG-----------TTGTTTTTTT---CA 271
12          CATGTCGTCCGAAAAGCATGGCAGACGCACGGGAGGTTCTGCCGATTGTTTTTTTT--CA 279
13          CATGTCGTCCGAAAAGCATGGCAGACGCACGGGAGGTTCTGCCGATTGTTTTTTTT--CA 279
25          CATGTCGTCCGAAAAGCATGGCAGACGCACGGGAGGTTCTGCCGGTTGTTTTTCT---CA 283
26          CATGTCGTCCGAAAAGCATGGCAGACGCACGGGAGGTTCTGCCGGTTGTTTTTCT---CA 283
24          CATGTCGTCCGAAAAGCATGGCAGACGCACGGGAGGTTCTGCCGGTTGTTTTTCT---CA 283
27          CATGTCGTCCGAAAAGCATGGCAGACGCACGGGAGGTTCTGTCGATTGTTTTTTT---CA 241
28          CATGTCGTCCGAAAAGCATGGCAGACGCACGGGAGGTTCTGTCGATTGTTTTTTT---CA 282
29          CATGTCGTCCGAAAAGCATGGCAGACGCACGGGAGGTTCTGTCGATTGTTTTTTT---CA 283
30          CATGTCGTCCGAAAAGCATGGCAGACGCACGGGAGGTTCTGCCGGTTGTTTTTCT---CA 283
31          CATGTCGTCCGAAAAGCATGGCAGACGCACGGGAGGTTCTGCCGGTTGTTTTTTT---CA 283
```

```
21              CATGTCGTCCGAAAGGCATGGCAGACGCACGGGAGGTTCTGCCGGTTGTTTTTTT---CA 287
22              CATGTCGTCCGAAAGGCATGGCAGACGCACGGGAGGTTCTGCCGGTTGTTTTTTTTTTTCA 286
23              CATGTCGTCCGAAAGGCATGGCAGACGCACGGGAGGTTCTGCCGGTTGTTTTTTTTTTTCA 286
20              CATGTCGTCCAAAAGGCATGACAGACGCACGGGACGTTCTGCCGGTTGATTTTTT---CA 283
1               CATGTCGTCCGAAAGGCATGGCAGACGCACGGGAGGTTCTGCCGGTTGTTTTTCT---CA 283
                ********** *** ***** ******** * *           *** ****     **

14              TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTACTGGCATTGAAACATGATGACCC 331
15              TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTACTGGCATTGAAACATGATGACCC 322
19              TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTACTGGCATTGAAACATGATGACCC 331
18              TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTACTGGCATTGAAACATGATGACCC 332
16              TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTACTGGCATTGAAACATGATGACCC 254
17              TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTACTGGCATTGAAACATGATGACCC 331
2               TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTACTGGCATTGAAACATGATGACCC 331
5               TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTACTGGCATTGAAACATGATGACCC 332
7               TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTACTGGCATTGAAACATGATGACCC 332
6               TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTACTGGCATTGAAACATGATGACCC 331
3               TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTACTGGCATTGAAACATGATGACCC 330
4               TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTACTGGCATTGAAACAGGATGACCC 331
8               TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTACTGGCATTGAAACATGATGACCC 331
9               TCCTCCCTCCACCCCTGTCTCCGATTGAATGCCTGTACTGGCATTGAAACAGGATGACCC 332
10              TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTACTGGCATTGAAACATGATGACCC 332
11              TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTGCTGGCATTGAAACATGATGACCC 331
32              TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTACTGGCATTGAAACATGATGACCC 338
33              TCCTCCCTCCACCCCTGTCCCCGACTGAATGCCTGTGCTGGCACTGAAACATGATGACCC 331
12              TCCTCCCTCCACCCCTGTCCCCGACTGAATGCCTGTGCTGGCACTGAAACATGATGACCC 339
13              TCCTCCCTCCACCCCTGTCCCCGACTGAATGCCTGTGCTGGCACTGAAACATGATGACCC 339
25              TCCTCCCTCCACCCCTGTCCCCGACTGAATGCCTGTGCTGGCACTGAAACATGATGACCC 343
26              TCCTCCCTCCACCCCTGTCCCCGACTGAATGCCTGTGCTGGCACTGAAACATGATGACCC 343
24              TCCTCCCTCCACCCCTGTCCCCGACTGAATGCCTGTGCTGGCACTGAAACATGATGACCC 343
27              TCCTCCCTCCACCCCTGTCCCCGACTGAATGCCTGTGCTGGCACTGAAACATGATGACCC 301
28              TCCTCCCTCCACCCCTGTCCCCGACTGAATGCCTGTGCTGGCACTGAAACATGATGACCC 342
29              TCCTCCCTCCACCCCTGTCCCCGACTGAATGCCTGTGCTGGCACTGAAACATGATGACCC 343
30              TCCTCCCTCCACCCCTGTCCCCGACTGAATGCCTGTGCTGGCACTGAAACATGATGACCC 343
```

```
31              TCCTCCCTCCACCCCTGTCCCCGACTGAATGCCTGTGCTGGCACTGAAACATGATGACCC 343
21              TCCTCCCTCCACCCCTGTCCCCGATTGAATGCCTGTGCTGGCACTGAAACATGATGACCC 347
22              TCCTCCCTCCACCCCTGTCCCCGACTGAATGCCTGTGCTGGCATTGAAACATGATGACCC 346
23              TCCTCCCTCCACCCCTGTCCCCGACTGAATGCCTGTGCTGGCATTGAAACATGATGACCC 346
20              TCCTCCCTCCACCCCTGTCCCCGACTGAATGCCTGTGCTGGCACTGAAACATGATGACCC 343
1               TCCTCCCCCCACCCCTGTCCCCGACTGAATGCCTGTGCTGGCACTGAAACATGATGACCC 343
                ******* *********** **** *********** ****** ******* ********

14              CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 391
15              CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 382
19              CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 391
18              CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 392
16              CATTCTA-CCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 313
17              CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTGTTTCCCTGTTGGGCTGA 391
2               CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 391
5               CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 392
7               CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 392
6               CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 391
3               CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 390
4               CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGG 391
8               CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 391
9               CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 392
10              CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 392
11              CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 391
32              CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 398
33              CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 391
12              CATTCTAACCCAGCTCGATGACGCCGGTGTGGTTATTCCTTTTGTTCCCTGTTGGGCTGA 399
13              CATTCTAACCCAGCTCGATGACGCCGGTGTGGTTATTCCTTTTGTTCCCTGTTGGGCTGA 399
25              CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTCTTTTTGTTCCCTGTTGGGCTGA 403
26              CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTCTTTTTGTTCCCTGTTGGGCTGA 403
24              CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTCTTTTTGTTCCCTGTTGGGCTGA 403
27              CATTCTAACCCAGCTCGATGACGCTGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 361
28              CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 402
29              CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 403
```

```
30          CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 403
31          CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTGCTTTTGTTCCCTGTTGGGCTGA 403
21          CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTCCTTTTGTTCCCTGTTGGGCTGA 407
22          CATTCTAACCCAGCTCGATGACGCCGGTGTAATTATTCCTTTTGTTCCCTGTTGGGCTGA 406
23          CATTCTAACCCAGCTCGATGACGCCGGTGTAATTATTCCTTTTGTTCCCTGTTGGGCTGA 406
20          CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTCTTTTTGTTCCCTGTTGGGCTGA 403
1           CATTCTAACCCAGCTCGATGACGCCGGTGTGATTATTCCTTTTGTTCCCTGTTGGGCTGA 403
            ******* *************** *****  *****  ***  ***************

14          CGGATGTGCATGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 451
15          CGGATGTGCATGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 442
19          CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 451
18          CGGATGTGCATGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 452
16          CGGATGTGCATGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 373
17          CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 451
2           CGGATGTGCATGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 451
5           CGGATGTGCATGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 452
7           CGGATGTGCATGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 452
6           CGGATGTGCATGGGGCGGTGGTTGTGCGCTGCGCTTGGTTGTCTCCGCCAGGCTTTCGTC 451
3           CGGATGTGCATGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 450
4           CGGATGTGCATGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 451
8           CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 451
9           CGGATGTGCGTAGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 452
10          CGGATGTGCGTAGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 452
11          CGGACGTGCGTAGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 451
32          CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 458
33          CGGATGTGCATGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 451
12          CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 459
13          CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 459
25          CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 463
26          CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 463
24          CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 463
27          CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 421
28          CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 462
```

```
29              CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 463
30              CGGATGTGCATGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 463
31              CGGATGTGCATGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 463
21              CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 467
22              CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 466
23              CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 466
20              CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 463
1               CGGATGTGCGTGGGGCGGTGGTTGTGCGCTGCGCTCGGTTGTCTCCGCCAGGCTTTCGTC 463
                **** **** * ********************** ************************

14              ACCCTGCAGCCAATGATGAGCCCCT-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 510
15              ACCCTGCAGCCAATGATGAGCCCCT-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 501
19              ACCCTGCAGCCAATGATGAGCCCCT-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 510
18              ACCCTGCAGCCAATGATGAGCCCCT-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 511
16              ACCCTGCAGCCAATGATGAGCCCCT-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 432
17              ACCCTGCAGCCAATGATAAGCCCCT-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 510
2               ACCCTGCAGCCAATGATGAGCCCCT-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 510
5               ACCCTGCAGCCAATGATGAGCCCCT-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 511
7               ACCCTGCAGCCAATGATGAGCCCCT-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 511
6               ACCCTGCAGCCAATGATGAGCCCCT-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 510
3               ACCCTGCAGCCAATGATGAGCCCCT-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 509
4               ACCCTGCAGCCAATGATGAGCCCCT-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 510
8               GCCCTGCAGCCAATGATGAGCCCCCCGCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 511
9               GCCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCACAGATGAGAT 511
10              GCCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCACAGATGAGAT 511
11              GCCCTGCAGCCAATGATGAGCCCCT-GCTTGCCACTTCTCTTTGATGCCGCAAATGAGAT 510
32              GCCCTGCAGCCAATGATGAGCCCCT-GCTTGCCACTCCTCTTTGATGCCACAGATGAGAT 517
33              GCCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 510
12              GCCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 518
13              GCCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 518
25              GCCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCGCAGATGAGAT 522
26              GCCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCGCAGATGAGAT 522
24              GCCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 522
27              GCCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCGCAGATGAGAT 480
```

```
28              GCCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCACAGATGAGAT 521
29              GCCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCGCAGATGAGAT 522
30              GCCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 522
31              ACCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 522
21              GCCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 526
22              GCCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 525
23              GCCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 525
20              GCCCTGCAGCCAATGATGAGCCCCC-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 522
1               ACCCTGCAGCCAATGATGAGCCCCT-GCTTGCCACTCCTCTTTGATGCCGCAAATGAGAT 522
                 ********************* ******  ********** ************ ** *******

14              TCGTAGCGCGCCTGATAGCGTCTGTATGGG 540
15              TCGTAGCGCGCCTGATAGCGTCTGTATGGG 531
19              TCGTAGCGCGCCTGATAGCGTCTGTATGGG 540
18              TCGTAGCGCGCCTGATAGCGTCTGTATGGG 541
16              TCGTAGCGCGCCTGATAGCGTCTGTATGGG 462
17              TCGTAGCGCGCCTGATAGCGTCTGTTTGGG 540
2               TCGTAGCGCGCCTGATAGCGTCTGTATGGG 540
5               TCGTAGCGCGCCTGATAGCGTCTGTATGGG 541
7               TCGTAGCGCGCCTGATAGCGTCTGTATGGG 541
6               TCGTAGCGCGCCTGATAGCGTCTGTATGGG 540
3               TCGTAGCGCGCCTGATAGCGTCTGTATGGG 539
4               TCGTAGCGCGCCTGATAGCGTCTGTATGGG 540
8               CCGTAGCGCGCCTGATAGCGTCTGTATGGG 541
9               TCGTAGCGCGCCTGATAGCGTCTGTTTGGG 541
10              TCGTAGCGCGCCTGATAGCGTCTGTTTGGG 541
11              TCGTAGCGCGCCTGATAGCGTCTGTTTGGG 540
32              TCGTAGCGCGCCTGATAGCGTCTGTTTGGG 547
33              TCGTAGCGCGCCTGATAGCGTCTGTATGGG 540
12              TCGTAGCGCGCCTGATAGCGTCTGTTTGGG 548
13              TCGTAGCGCGCCTGATAGCGTCTGTTTGGG 548
25              TCGTAGCGCGCCTGATAGCGTCTGTATGGG 552
26              TCGTAGCGCGCCTGATAGCGTCTGTATGGG 552
24              TCGTAGCGCGCCTGATAGCGTCTGTTTGGG 552
```

```
27              TCGTAGCGCGCCTGATAGCGTCTGTTTGGG 510
28              TCGTAGCGCGCCTGATAGCGTCTGTATGGG 551
29              TCGTAGCGCGCCTGATAGCGTCTGTATGGG 552
30              TCGTAGCGCGCCTGATAGCGTCTGTATGGG 552
31              TCGTAGCGCGCCTGATAGCGTCTGTATGGG 552
21              TCGTAGCGCGCCTGATAGCGTCTGTTTGGG 556
22              TCGTAGCGCGCCTGATAGCGTCTGTATGGG 555
23              TCGTAGCGCGCCTGATAGCGTCTGTATGGG 555
20              TCGTAGCGCGCCTGATAGCGTCTGTTTGGG 552
1               TCGTAGCGCGCCTGATAGCGTCTGTTTGGG 552
                 ************************ ****
```

Clustal alignment of the 29 variants of snoRNA gene Cluster 25, produced by digestion with EcoRI. Yellow highlighted nucleotides represent the coding region for Eg-m18.

```
1       AATTCCATCTTTTGCTCCCCCTCCCTCCTGCTGCTTTTCGGTCTCGCCGGGCCGCGGTGT
2       AATTCCATCTTTTGCTCCCCCTCCCTTCTGCTGCTTTTCGGTCTCGCCGGGCCGTGGTGT
3       AATTCCATCTTTTGCTCCCCCTCCCTCCTGCTGCTTTTCGGTCTCGCCGGGCCGCGGTGT
4       AATTCCATCTTTTGCTCCCCCTCCCTCCTGCTGCTTTTCGGTCTCGCCGGGCCGCGGTGT
        ************************** *************************** *****

1       GCCCGCTGCCTGAGGGAAAATGCATGGGCCCATCGCCGCGCGTGTGTTCGTGCGG-GGCC
2       GCCCGCTGCCTGAGGGAAAATGCATGGGCCCATCGCCGCGCGCGTGTTCGTGCCGGGGCC
3       GCCCGCTGCCTGCGGGAAAATGCATGGGCCCATCGCCGCGCGTGTGTTCGTGCCGGGGCC
4       GCCCGCTGCCTGAGGGAAAATGCATGGGCCCATCGCCGCGCGTGTGTTCGTGCTGGGGCC
        ************ ***************************** ********** * ****

1       GCCGCCTGCCTGGGTGATGTGTCCTCCTCCACCCTTTTTTTTTCATTATTGTTTTATTGG
2       GCCGCCTGCCTGGGTGATGTGTCCTCCTCCACCCTTTTTTTTTCATTATTGTTTTATTGG
3       GCCGCCTGCCTGGGTGATGTGTCCTCCTCCACCCTTTTTTTTTCATTATTGTTTTATTAG
4       GCCGCCTGCCTGGGTGATGTGTCCTCCTCCACCCTTTTTTTTTCATTATTGTTTTATTGG
        ********************************************************** *
```

```
1       ACACCCAGCTCGGGTGGGGAGTTTGATCGCAACAGCCGCTGCTCTGCTTCAACGGGTTGG
2       ACACCCAGCTCAGGTGGGGAGTTTGATCGCAACAGCCGCTGCTCTGCTTCAACGGGTTGG
3       ACACCCAGCTCGGGTGGGGAGTTTGATCGCAACAGCCGCTGCTCTGCTTCAACGGGTTGG
4       ACACCCAGCTCGGGTGGGGAGTTTGATCGCAACAGCCGCTGCTCTGCTTCAACGGGTTGG
        *********** **********************************************

1       GGTTTTGGCCTTCAGTGGCACACTGCTTGCCTTCCTGCACTGTATGGCGCTGTTGCGAAG
2       GGTTTTGGCCTTCAGTGGCACACTGCTTGCCTTCCTGCACTGTATGGCGCTGTTGCGAAG
3       GCTTTTGGCCTTCAGTGGCACACTGCTTGCCTTCCTGCACTGTATGGCGCTGTTGCGAAG
4       GGTTTTGGCCTTCAGTGGCACACTGCTTGCCTTCCTGCACTGTATGGCGCTGTTGCGAAG
        * ************************************************************

1       GATGTCGGTTTCAATTAAACGCTGACCGCGTGATTTGCCTTCTGCCTGACTGGGCGTGGG
2       GATGTCGGTTTCAATTAAACGCTGACCGCGTGATTTGCCTTCTGCCTGACTGGGCGTGGG
3       GATGTCGGTTTCAATTAAACGCTGACCGCGTGATTTGCCTTCTGCCTGACTGGGCGTGGG
4       GATGTCGGTTTCAATTAAACGCTGACCGCGTGATTTGCCTTCTGCCTGACTGGGCGTGGG
        ************************************************************

1       GAGCTGTGGTGACAGGCGCAGGGCTGGCCTGGGGCAACACCC-------ACC--------
2       GAGCTGTGGTGACAGGCGCAGGGCTGGCCTGGGGTAACACCC-------ACC--------
3       GAGCTGTGGTGACAGGCGCAGGGCTGGCCTGGGGTAACACCCTGTGGGCACCCACCCACC
4       GAGCTGTGGTGACAGGCGCAGGGCTGGCCTGGGGTAACACCCTGTGGGCACCCACCCACC
        ********************************** *******       ***

1       CACCCTGTGGGC---CGGGAGCTGAGAGGGGGGCGTTCAGGGTGTGGGGGTTGGGTGCAC
2       CACCCTGTGGGCGACCGGGAGCTGAGAGGGGGGCGTTCAGGGTGTGGGGGTTGGGTGCAC
3       CACCCTGTGGGCGACCGGGAGCTGAGAGGGGGGCGTTCAGGGTGTGGGGGTTGGGTGCAC
4       CACCCTGTGGGCGACCGGGAGCTGAGAGGGGGGCGTTCAGGGTGTGGGGGTTGGGTGCAC
        ************   *********************************************

1       TCATCTCGGAGGGCGGACTGCGCCGGACGTTGCGAATTGTGACCCAGGATGTCGGTTTCA
2       TCATCTCGGAGGGCGGACTGCGCCGGACGTTGCGAATTGTGACCCAGGATGTCGGTTTCA
3       TCATCTCGGAGGGCGGACTGCGCCGGACGTTGCGAATTGTGACCCAGGATGTCGATTTCA
4       TCATCTCGGAGGGCGGACTGCGCCGGACGTTGCGAATTGTGACCCAGGATGTCGATTTCA
```

```
        ************************************************************ *****

1       ATTAAACGCTGCCACCTGATGTGTGCACTAACTGAGGCAAGTCCGCCAGGGATTGCACAG
2       ATTAAACGCTGCCACCTGATGTGTGCACTAACTGAGGCAAGTCCGCCGGGGATTGCACAG
3       ATTAAACGCTGCCACCTGATGTGTGCACTCACTGAGGCAAGTCCGCCAGGGATTGCACAG
4       ATTAAACGCTGCCACCTGATGTGTGCACTCACTGAGGCAAGTCCGCCAGGGATTGCACAG
        ***************************** ***************** ************

1       GTTGAACGTGGGGAATT
2       GTTGAACGTGGGGAATT
3       GTTGAACGTGGGGAATT
4       GTTGAACGTGGGGAATT
        ****************
```

Clustal alignment of the variants of Eg-h1 gene cluster, produced by digestion with NcoI. Yellow highlighted nucleotides represent the coding region for Eg-h1.

```
10              ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTC 45
11              ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTC 45
16              ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTC 45
21              ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTC 45
12              ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTC 45
3               ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTC 45
20              ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTC 45
22              ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTC 45
5               ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTC 45
1               ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTC 45
7               ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTG 45
17              ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTG 45
6               ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTG 45
9               ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTG 45
8               ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTG 45
18              ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTG 45
19              ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTG 45
```

```
14              ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTC 45
15              --------------------------------GGCAAAACTCGTGCTATATTATTTTTC 27
2               ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTC 45
13              ---------------CATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTC 45
4               CATGGCAAAACTCGCTATGGCAAAACTCGCTATGGCAAAACTCGTGCTATATTATTTTTC 60
                                           *************************

10              CCGGAATGGCTGCAGGGCCTGCCTGCGGGGGGGCCCTGTGGTCACCGAATGTGCTCACTG 105
11              CCGGAATGGCTGCAGGGCCTGCCTGCGGGGGGGCCCTGTGGTCACCGAATGTGCTCACTG 105
16              CCGGAATGGCTGCAGGGCCTGCCTCCGGGGGG-CCCTGTGGTCACCGAATGTGCTCACTG 104
21              CCGGAATGGCTGCAGGGCCTGCCTCCGGGGGG-CCCTGTGGTCACCGAATGTGCTCACTG 104
12              CCGGAATGGCTGCAGGGCCTGCCTCCGGGGGG-CCCTGTGGTCACCGAATGTGCTCACTG 104
3               CCGGAATGGCTGCAGGGCCTGCCTGCGGGGGGGCCCTGTGGTCACCGAATGTGCTCACTG 105
20              CCGGAATGGCTGCAGGGCCTGCCTCCGGGGGGGCCCTGTGGTCACCGAATGTGCTCACTG 105
22              CCGGAATGGCTGCAGGGCCTGCCTCCGGGGGG-CCCTGTGGTCACCGAATGTGCTCACTG 104
5               CCGGAATGGCTGCAGGGCCTGCCTCCGGGGGG-CCCTGTGGTCACCGAATGTGCTCACTG 104
1               CCGGAATGGCTGCAGGGCCTGCCTCCGGGGGG-CCCTGTGGTCACCGAATGTGCTCACTG 104
7               CCGGAATGGCTGCAGGGCCTGCCTCCG--------------TCACCGAATGTGCTCACTG 91
17              CCGGAATGGCTGCAGGGCCTGCCTCCG--------------TCACCGAATGTGCTCACTG 91
6               CCGGAATGGCTGCAGGGCCTGCCTCCG--------------TCACCGAATGTGCTCACTG 91
9               CCGGAATGGCTGCAGGGCCTGCCTCCG--------------TCACCGAATGTGCTCACTG 91
8               CCGGAATGGCTGCAGGGCCTGCCTCCG--------------TCACCGAATGTGCTCACTG 91
18              CCGGAATGGCTGCAGGGCCTGCCTCCG--------------TCACCGAATGTGCTCACTG 91
19              CCGGAATGGCTGCAGGGCCTGCCTCCG--------------TCACCGAATGTGCTCACTG 91
14              CCGGAATGGCTGCAGGGCCTGCCTCCG--------------TCACCGAATGTGCTCACTG 91
15              CCGGAATGGCTGCAGGGCCTGCCTCCG--------------TCACCGAATGTGCTCACTG 73
2               CCGGAATGGCTGCAGGGCCTGCCTCCG--------------TCACCGAATGTGCTCACTG 91
13              CCGGAATGGCTGCAGGGCCTGCCTCCG--------------TCACCGAATGTGCTCACTG 91
4               CCGGAATGGCTGCAGGGCCTGCCTCCG--------------TCACCGAATGTGCTCACTG 106
                ************************ **              ******************

10              AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 165
11              AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 165
16              AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 164
```

```
21          AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 164
12          AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 164
3           AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 165
20          AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 165
22          AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 164
5           AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 164
1           AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 164
7           AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 151
17          AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 151
6           AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 151
9           AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 151
8           AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 151
18          AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 151
19          AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 151
14          AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 151
15          AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 133
2           AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 151
13          AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCGTTCGGAGAGCAAGAAGG 151
4           AGCAGGGTGCGCCCCTCCAATGGCGCCCCCCTGGCCCACGCCATTCGGAGAGCAAGAAGG 166
            ****************************************** ****************

10          GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCGGCGCTCGCCACCTTGGTACGCATCATAC 224
11          GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCGGCGCTCGCCACCTTGGTACGCATCATAC 224
16          GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCGGCGCTCGCCACCTTGGTACGCACCATAC 223
21          GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCGGCGCTCGCCACCTTGGTACGCATCATAC 223
12          GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCGGCGCTCGCCACCTTGGTACGCATCATAC 223
3           GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCTGTGCTCGCCACCTTGGTACGCATCATAC 224
20          GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCGGCGCTCGCCACCTTGGTACGCATCATAC 224
22          GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCGGCGCTCGCCACCTTGGTACGCATCATAC 223
5           GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCGGCGCTCGCCACCTTGGTACGCATCATAC 223
1           GGCTCTGGAGCGCCGGGGGG-CCCCCCTCCCGGCGCTCGCCACCTTGGTACGCATCATAC 223
7           GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCGGCGCTCGCCACCTTGGTACGCATCATAC 210
17          GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCGGCGCTCGCCACCTTGGTACGCATCATAC 210
6           GGCTCTGGAGCGCCGGGGGG-CCCCCCTCCCGGCGCTCGCCACCTTGGTACGCATCATAC 210
```

```
9                GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCGGCGCTCGCCACCTTGGTACGCATCATAC 210
8                GGCTCTGGAGCGCCGGGGGGGCCCGCCTCCCGGCGCTCGCCACCTTGGTACGCATCATAC 211
18               GGCTCTGGAGCGCCGGGGGGGCCCGCCTCCCGGCGCTCGCCACCTTGGTACGCATCATAC 211
19               GGCTCTGGAGCGCCGGGGGGGCCCGCCTCCCGGCGCTCGCCACCTTGGTACGCATCATAC 211
14               GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCTGTGCTCGCCACCTTGGTACGCATCATAC 210
15               GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCTGTGCTCGCCACCTTGGTACGCATCATAC 192
2                GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCTGTGCTCGCCACCTTGGTACGCATCATAC 210
13               GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCTGTGCTCGCCACCTTGGTACGCATCATAC 210
4                GGCTCTGGAGCGCCGGGGGG-CCCGCCTCCCGGCGCTCGCCACCTTGGTACGCATCATAC 225
                 ******************** *** ****** * ******************** *****

10               CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 284
11               CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 284
16               CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 283
21               CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 283
12               CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 283
3                CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 284
20               CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGG-ATCTTCACGTGCGTAACA 283
22               CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 283
5                CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 283
1                CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 283
7                CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 270
17               CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 270
6                CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 270
9                CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 270
8                CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 271
18               CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 271
19               CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 271
14               CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 270
15               CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 252
2                CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 270
13               CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 270
4                CTTGTCCTCTGTGTGGCCCCGCGAGGGGCGATAACACTTGGGATCTTCACGTGCGTAACA 285
                 ************************************** ******************
```

```
10              CATCGGCGGAGCGTTCGTTTTT-GTTTTCAGAAACCACAACCGAGGTGCAGGCAATGTTG 343
11              CATCGGCGGAGCGTTCGTTTTT-GTTTTCAGAAACCACAACCGAGGTGCAGGCAATGTTG 343
16              CATCGGCGGAGCGTTCGTTTTT-GTTTTCAGAAACCACAACAGAGGTGCAGGCAATGTTG 342
21              CATCGGCGGAGCGTTCGTTTT--GTTTTCAGAAACCACAACCGAGGTGCAGGCAATGTTG 341
12              CATCGGCGGAGCGTTCGTTTTT-GTTTTCAGAAACCACAACCGAGGTGCAGGCAATGTTG 342
3               CATCGGCGGAGCGTTCGTTTTT-GTTTTCAGAAACCACAACAGAGGTGCAGGCAATGTTG 343
20              CATCGGCGAAGCGTTCGTTTTT-GTTTTCAGAAACCACAACCGAGGTGCAGGCA------ 336
22              CATCGGCGAAGCGTTCGTTTT--GTTTTCAGAA-CCACA-CCGAGGTGCAGGCA------ 333
5               CATCGGCGAAGCGTTCGTTTTT-GTTTTCAGAAACCACAACCGAGGTGCAGG-------- 334
1               CATCGGCGAAGCGTTCGTTTTT-GTTTTCAGAAACCACAACCGAGGTGCAGG-------- 334
7               CATCGGCGAAGCGTTCGTTTTT-GTTTTCAGAAACCACAACCGAGGTGCAGG-------- 321
17              CATCGGCGAAGCGTTCGTTTTT-GTTTTCAGAAACCACAACCGAGGTGCAGG-------- 321
6               CATCGGCGAAGCGTTCGTTTTT-GTTTTCAGAAACCACAACCGAGGTGCAGG-------- 321
9               CATCGGCGAAGCGTTCGTTTTTTGTTTTCAGAAACC---------------------- 306
8               CATAGGCGAAGCGTTCGTTTTT-GTTTTCAGAAACCACAACCGAGGTGCAGG-------- 322
18              CATAGGCGAAGCGTTCGGTTTT-GTTTTCAGAAACCACAACCGAGGTGCAGG-------- 322
19              CATCGGCGAAGCGTTCGTTTTT-GTTTTCA------------------------------ 300
14              CATCGGCGGAGCGTTCGTTTTT-GTTTTCAGAAACCACAACAGAGGTGCAGGCAATGTTG 329
15              CATCGGCGGAGCGTTCGTTTTT-GTTTTCAGAAACCACAACAGAGGTGCAGGCAATGTTG 311
2               CATCGGCGGAGCGTTCGTTTTT-GTTTTCAGAAACCACAACAGAGGTGCAGGCAATGTTG 329
13              CATCGGCGGAGCGTTCGTTTTT-GTTTTCAGAAACCACAACCGAGGTGCAGGCAATGTTG 329
4               CATCGGCGAAGCGTTCGTTTTT-GTTTTCAGAAACCACAACCGAGGTGCAGGCAATGTTG 344
                *** **** ******** ***  *******

10              CAGCAGGCATGTGGGGGGTGGGGGAGGTGGAAGGAGGCGGCAGTCGTTTTTGAGTTCTCA 403
11              CAGCAGGCATGTGGGGGGTGGGGAGGGTGGAAGGAGGCGGCAGTCGGTTTTGAGTTCTCA 403
16              CAGCAGGCATGTGGGGG-TGGG--AGGTG-AAGGAGGCG-CAGTC--------------- 382
21              CAGCA------------------------------------------------------ 346
12              CAGCAGGCATGTGGGGG-TGGGG-AGGTGGAAGGAGGCGGCAGTCGGTTTTGAGTTCTCA 400
3               CAGCAGGGATGTGGGGG-TGGGG-AGGTGGAAGGAGGCGGCAGTCGGTTTTGAGTTCTCA 401
20              ---------TGTGGGG--TGGG---AGTTGAAGGAGGC---------------------- 360
22              ---------TGTGGGG--TGGGG--AGTG-AAGGAGCCG-CAGTCG-TTTTGAGTTCTCA 377
5               ------GCATGTGGGGG-TGGGG-AGGTGGAAGGAGGCGGCAGTCGGTTTTGAGTTCTCA 386
1               ------GCATGTGGGGG-TGGGG-AGGTGGAAGGAGGCGGCAGTCGGTTTTGAGTTCTCA 386
```

```
7               ------GCATGTGGGGG-TGGGG-AGGTGGAAGGAGGCGGCAGTCGGTTTTGAGTTCTCA 373
17              ------GCATGTGGGGG-TGGGG-AGGTGGAAGGAGGCGGCAGTCGGTTTTGAGTTCTCA 373
6               ------GCATGTGGGGG-TGGGG-AGGTGGAAGGAGGCGGCAGTCGGTTTTGAGTTCTCA 373
9               ------------------------------------------------------------
8               ------GTATGTGGGGG-TGGGG-AGGTGGAAGGAGGCGGCAATCGGTTTTGAGTTCTCA 374
18              ------GTATGTGGGGG-TGGGG-AGGTGGAAGGAGGCGGCAATCGGTTTTGAGTTCTCA 374
19              ------------------------------------------------------------
14              CAGCAGGGATGTGGGGG-TGGGG-AGGTGGAAGGAGGCGGCAGTCGGTTTTGAGTTCTCA 387
15              CAGCAGGGATGTGGGGG-TGGGG-AGGTGGAAGGAGGCGGCAGTCGGTTTTGAGTTCTCA 369
2               CAGCAGGGATGTGGGGG-TGGGG-AGGTGGAAGGAGGCGGCAGTCGGTTTTGAGTTCTCA 387
13              CAGCAGGGATGTGGGGG-TGGGG-AGGTGGAAGGAGGCGGCAGTCGGTTTTGAGTTCTCA 387
4               CAGCAGGGATGTGGGGG-TGGGG-AGGTGGAAGGAGGCGGCAGTCGGTTTTGAGTTCTCA 402


10              GGTTCGGGACCCGTCAGCGTGGCAA---------------------------------- 428
11              GGTTCGGGACCCGTCAGCGTGGCAA---------------------------------- 428
16              ------------------------------------------------------------
21              ------------------------------------------------------------
12              GGTTCGGGACCCGTCAGCGTGGCAATGGGGATTCGGAAACCGGC---------------- 444
3               GGTTCGGGACCCGTCAGCGTGGCAATGGGGATTCGGAAACCGGC---------------- 445
20              ------------------------------------------------------------
22              G-TTCGGA--CCGTCAGC------------------------------------------ 392
5               GGTTCGGGACCCGTCAGCGTGGCAATGGGGATTCGGAAACCGGCGCCAACCACACACCGG 446
1               GGTTCGGGACCCGTCAGCGTGGCAATGGGGATTCGGAAACCGGCGCCAACCACACACCGG 446
7               GGTTCGGGACCCGTCAGCGTGGCAATGGGGATTCGGAAACCGGCGCCAACCACACACCGG 433
17              GGTTCGGGACCCGTCAGCGTGGCAATGGGGATTCGGAAACCGGCGCCAACCACACACCGG 433
6               GGTTCGGGACCCGTCAGCGTGGCAATGGGGATTCGGAAACCGGCGCCAACCACACACCGG 433
9               ------------------------------------------------------------
8               GGTTCGGGACCCGTCAGCGTGGCAATGGGGATTCGGAAACCGCC---AACCACACACCGG 431
18              GGTTCGGGACCCGTCAGCGTGGCAATGGGGATTCGGAAACCGCC---AACCACACACCGG 431
19              ------------------------------------------------------------
14              GGTTCGGGGCCCGTCAGCGTGGCAATGGGGATTCGGAAACCG---CCAACCACACACCGG 444
15              GGTTCGGGGCCCGTCAGCGTGGCAATGGGGATTCGGAAACCG---CCAACCACACACCGG 426
2               GGTTCGGGACCCGTCAGCGTGGCAATGGGGATTCGGAAACCG---CCAACCACACACCGG 444
```

```
13              GGTTCGGGGCCCGTCAGCGTGGCAATGGGGATTCGGAAACCG---CCAACCACACACCGG 444
4               GGTTCGGGACCCGTCAGCGTGGCAATGGGGATTCGGAAACCGGCGCCAACCACACACCG- 461


10              ---------------------------------------------
11              ---------------------------------------------
16              ---------------------------------------------
21              ---------------------------------------------
12              -TCACCACAGCCACCCCTCACC-------TGCCACCATG------- 475
3               -TCACCACAGCCACCCCTCACC-------TGCCACCATG------- 476
20              ---------------------------------------------
22              ---------------------------------------------
5               CTCACCACAGCCACCCCTCACC-------TGCCACCTGCCCATG-- 483
1               CTCACCACAGCCACCCCTCACC-------TGCCACCTGCCCATG-- 483
7               CTCACCACAGCCACCCCTCACC-------TGCCACCTGCCCATG-- 470
17              CTCACCACAGCCACCCCTCACC-------TGCCACCTGCCCATG-- 470
6               CTCACCACAGCCACCCCTCACC-------TGCCACCTGCCCATG-- 470
9               ---------------------------------------------
8               CTCACCACAGCCACCCCTCACC-------CCCCACCTGCCACCATG 470
18              CTCACCACAGCCACCCCTCACC-------CCCCACCTGCCACCATG 470
19              ---------------------------------------------
14              CTCACCACAGCCACCCCTCACCCCCCACCTGCCACCATG------- 483
15              CTCACAACAGCCACCCCTCACCCCC-ACCTGCAAC-ATG------- 463
2               CTCACCACAGCCACCCCTCACCCCCCACCTGCCACCATG------- 483
13              CTCACCACAGCCACCCCTCACCCCCCACCTGCCACCATG------- 483
4               CTCACCACAGCCACCCCTCACC-------TGCTACCATG------- 493
```

Figure S5. Additional examples of gene duplication as a mechanism of box C/D snoRNA evolution and methylation clustering in *E. gracilis*.

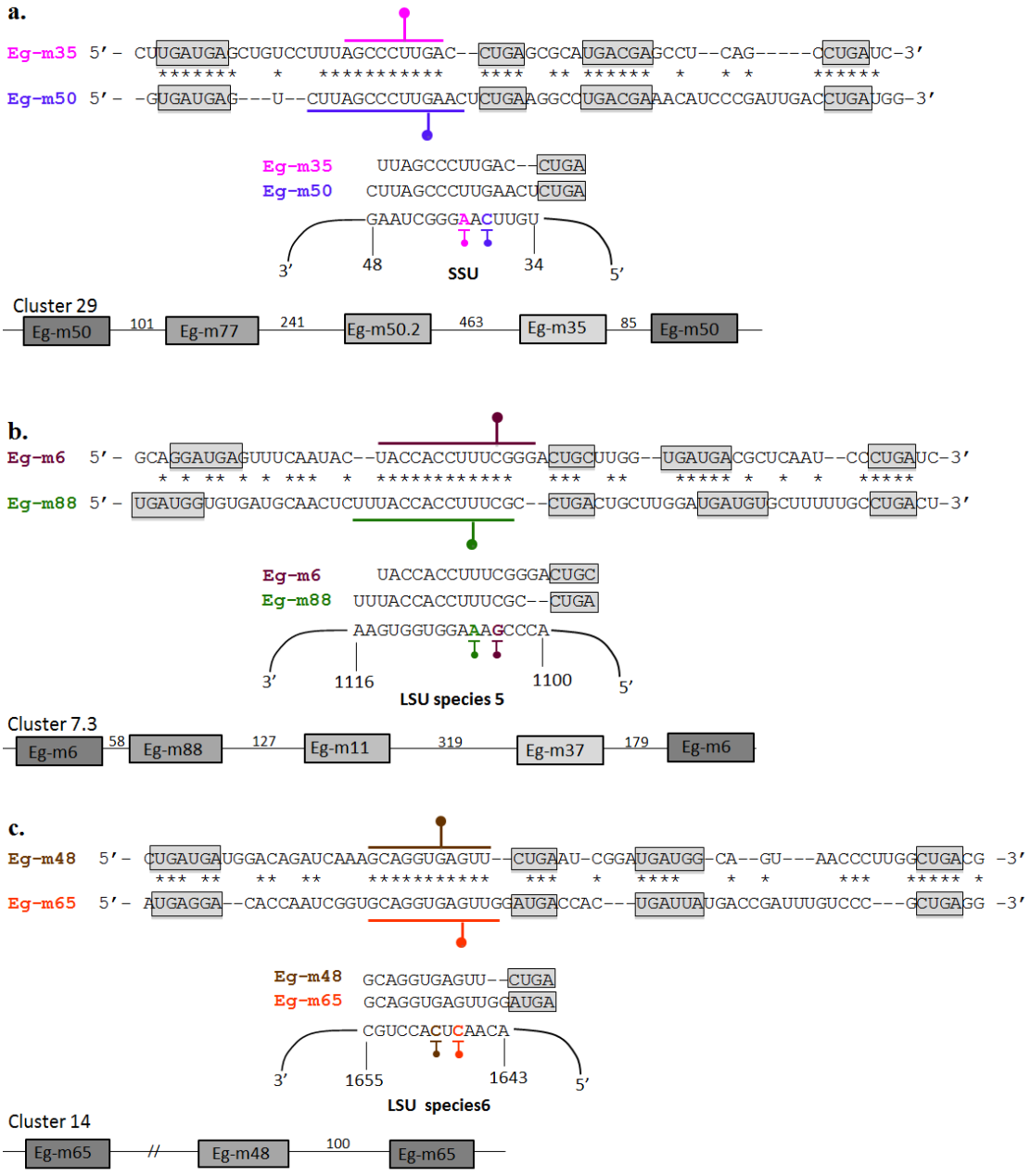


Figure S6. Additional examples of polycistronic expression of snoRNA gene clusters in *E. gracilis*.

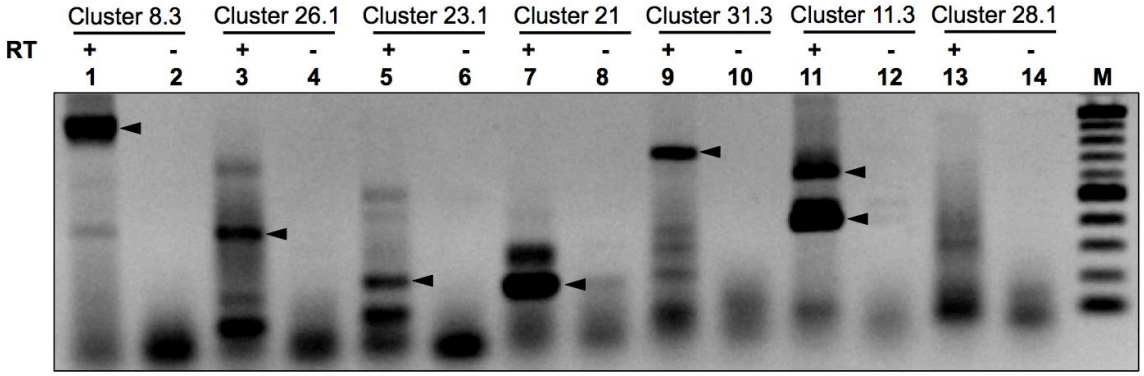


Figure S7. Additional examples of primer blocking to reduce rRNA amplification during small RNA library preparation.

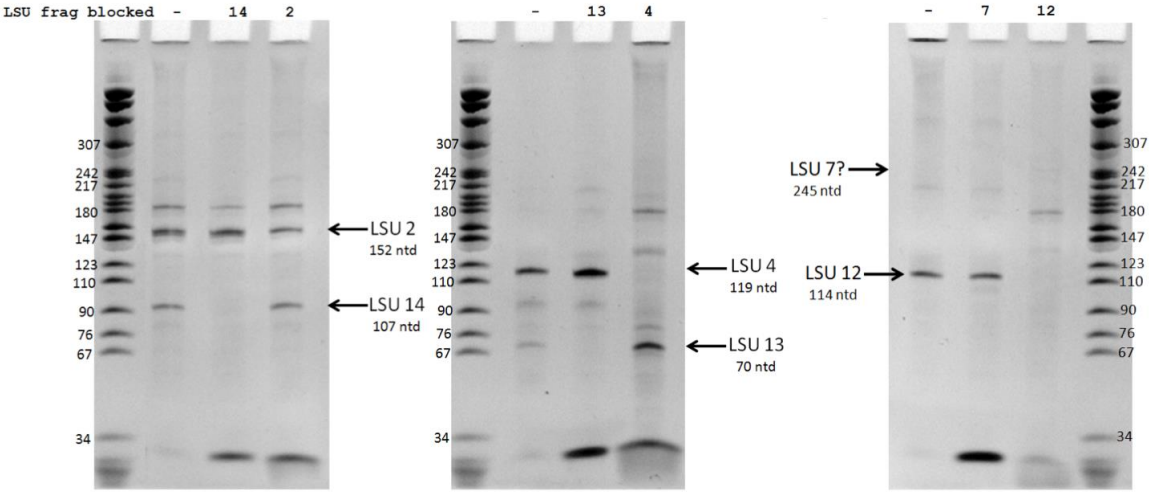


Figure S8. Sequences of predicted orphan snoRNAs identified from a *Euglena* small RNA library. These RNAs do not target any known rRNA or snRNA modification sites. The snoRNAs are indicated by library sequence cluster number. The box elements are highlighted as in Figure S1.

Orphan C/D box snoRNAs

>Cluster741:[5,64]
TGATGACNATTGAAGCATTGGACGCCAGATGACATTCTGAAACAATCCAACTCTTTCTGA

>Cluster3959:[4,48]
TCTTGATGATGATGGCGACGATGATAATGATGATGATGTTGATGCTGATGAT

>Cluster5070:[5,59] (clusters 27808, 18969)
GCGATGATGATTGCTTTTGATGAGTCCCATGATTCTGCCCAACTTTTGTTCGTCGCTGACTCTGAG

>Cluster11692:[15,63]
TGATGACCTAACGATGCATGCCGACCATGTCTGATGTTCGAGACCCTGA

>Cluster11704:[5,64]
CAGAGGATGGATGTGAATCTCGTTGAACCGAACGCATTGTGAAGGAGATGGCTTGGATTGCTGACCN

>Cluster14763:[7,65]
TGATGTTTCTGTTGATCTGAGTATTCTGAGCATTCGTGAAGGCAGACCATACTCTCTGA

>Cluster16253:[5,60]
GGATGTCCACTGGTTTGCACCGCTGATGAGGTTTGATGTGATTCTCTTTTGACTGA

>Cluster18492:[5,58]
GGATGAATTCATTTGTAGTGTTCTCTGACTCCGATGACGTCCACCAGGATCTGA

>Cluster20468:[5,67]
TGATGTTTTTAATTCAAGTTCTGATTCGAGTTCATCTGATTCCGATTCCAATTCCGATTCTGA

>Cluster21446:[5,60]
TGATGTTTTTCCAGAATTCGCGCTCTGACGGAAGCTGATGCTCTGCTATTTACTGA

>Cluster23369:[5,65]
CAAATGATGTGTCACATTTCCCTGGCACTGAGGTGCACGCCACCACTGATGCTTTTCATCCCTGACCC

>Cluster25741:[7,56]
TGATGGCCATTGCTGACATCCTTACTGAATCTGTTGATCACCCAGACTGA

>Cluster29708:[3,61]
TGATGAGAGCGGTGGGTGGACTGTCTCTGAATTTCCACTTCTTTTGCTTGCTTTGCTGA

>Cluster34036:[7,66] similar to 34454
GGATGAGCTGTGGTGATGGCGATGAATGAACACGGCGTTTGCTGGCTTTGATGAAACTGA

>Cluster34454:[12,73] similar to 34036
AGATGCCCGAAGCAAAGTGGCGATGAATGAACACTGCGTTTGCTGGCTTTGATGAAATCTGA

>Cluster35831:[5,54]
TGATGATCCACATCCTGACACCTCTTGTCTGCCGCGCACTTGGGCCCTGA

>Cluster50269:[5,61]
TGATGGCACGCCCAAACGTGAGCAGTGAAATTTCTTGAATGAAGGCTCACTGGCTGA

>Cluster60084:[4,64]
TGATGATCACCCCCACCGCCCCCGAGTTCCGAATCCAATGCTGGACGTCACCCTTGGCTGA

>Cluster64265:[15,65]
TGATGCCGCACTTCGTCGCTGAATGCACACTCCTCTTGCTGGCTTTGCTGA

Orphan box AGA snoRNAs

>Cluster62083:[2,64]
TGGCACTGCCACAGGCTCCACCTGCGCCCACGCTCTCGCTTGTGGAGAAACCAACGGTGCCAGATGCTCGCT

>Cluster14789:[2,63]
GTGCAGCTTCAAAGCGACTGCTCCATGTGTGTGTGGATGCCAGTTTCATCCTGAGCTGCAAGAGGT

>Cluster21320:[3,64]
CCCGAGGCAACTTGGTTGGGACCAGCTTGCTCTCATGCTTGTCACCGGCTTCATACCTCGAAGATCT

>Cluster26394:[3,58]
AAGCCTTGTTTGCGTTCCATGGCTTCTCATGAGCTTTGGACCTTCCCTCAAGGCAAGATTT

>Cluster48916:[2,59]
ATGCAGCTTTTACGCGACTGCCGCCTGCATGGTTGCCAGTCACATCCTGAGCTGCAAGATTT

>Cluster54246:[1,66]
TCTGATGGGGCGGGCCATCGGGGGTGAATGAACACTGCGCTTGCTGGCTTTGATGAAATCAGAAGA

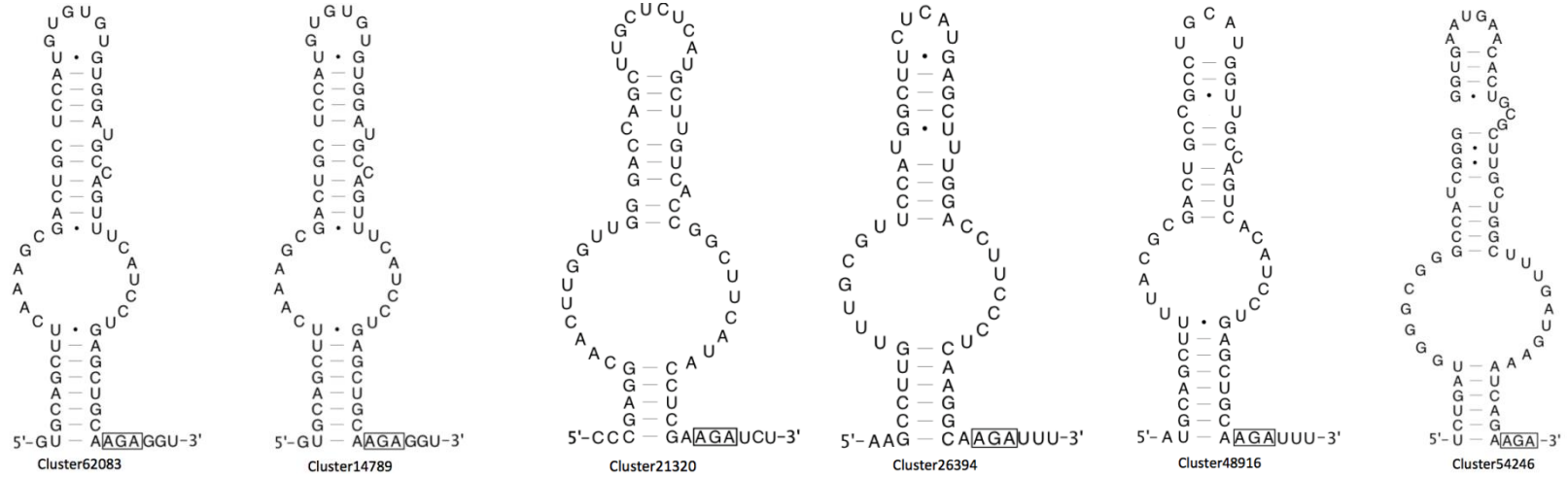


Figure S9. Examples of *E. gracilis* cultures treated with fibrillarin dsRNA (and control treatments with no dsRNA, indicated with a 'C'). For treatment conditions, refer to Table 5. For OD_{600} readings of the cultures, refer to Table S16.

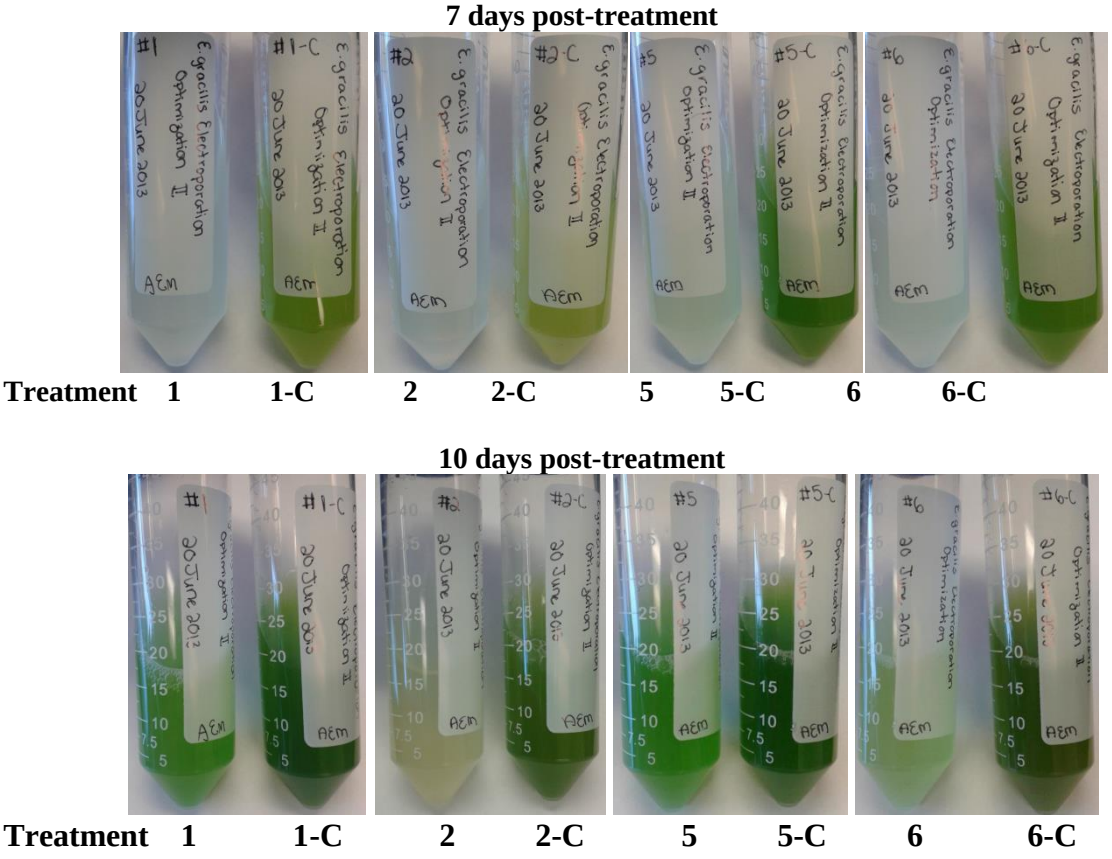


Figure S10. The inhibitory effect observed after electroporating *Euglena* cells with dsRNA is sequence specific. Cells treated with non-specific dsRNA did not exhibit the same growth defect as cells treated with anti-fibrillarin dsRNA.

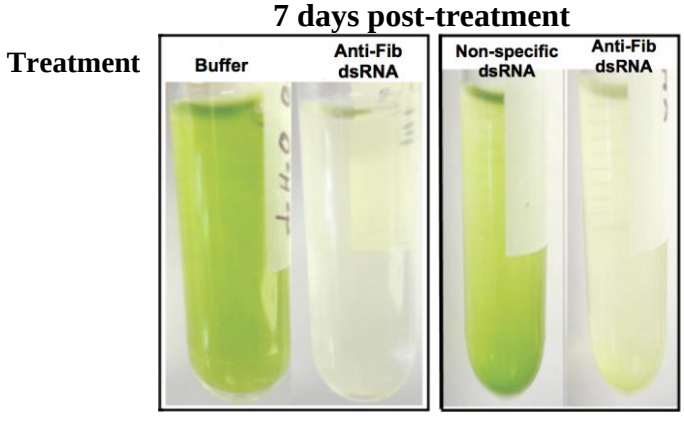


Figure S11. Loss of targeted 2′-O-methylation after the introduction of snoRNA-specific dsRNA was observed 9 and 25 days post-treatment. The products indicated with black arrow heads are the 2′-O-methylation sites guided by the targeted snoRNAs. The white arrow heads indicate products that are the result of predicted reverse transcriptase stop sites. See Figure 23 for further explanation.