

**IDENTIFICATION AND COMPARISON OF NON-CODING RNAS AND  
RIBONUCLEOPROTEIN COMPLEXES IN SEVERAL DIVERSE PROTIST  
SPECIES**

**DAVID CASEY MCWATTERS**  
Bachelor of Science, University of Lethbridge, 2013

A thesis submitted  
In partial fulfilment of the requirements for the degree of

**DOCTOR OF PHILOSOPHY**

**IN**

**BIOMOLECULAR SCIENCE**

Department of Biological Sciences  
University of Lethbridge  
LETHBRIDGE, ALBERTA, CANADA

© David Casey McWatters, 2022

**IDENTIFICATION AND COMPARISON OF NON-CODING RNAS AND  
RIBONUCLEOPROTEIN COMPLEXES IN SEVERAL DIVERSE PROTIST  
SPECIES**

DAVID CASEY MCWATTERS

Date of Defence: April 21, 2022

Dr. A. Russell Thesis Supervisor	Associate Professor	Ph.D.
Dr. I. Kovalchuk Thesis Examination Committee Member	Professor	Ph.D.
Dr. L. B. Selinger Thesis Examination Committee Member	Professor	Ph.D.
Dr. N. Thakor Internal External Examiner Department of Chemistry and Biochemistry Faculty of Arts and Science	Associate Professor	Ph.D.
Dr. J. Archibald External Examiner Dalhousie University Halifax, Nova Scotia	Professor	Ph.D.
Dr. S. Wiseman Chair, Thesis Examination Committee	Associate Professor	Ph.D.

## **DEDICATION**

For my Mom and Dad who from my earliest days fueled my passion for science. This thesis would not exist without your unconditional love and support.

## ABSTRACT

Many different classes of non-coding (nc)RNAs are found in species throughout the tree of life, each of which performs important cellular functions. The objective of my thesis was to identify and characterize ncRNAs and their associated protein complexes, with particular focus on small nucleolar RNAs (snoRNA), in unicellular eukaryotes whose genomes have undergone significant reduction or expansion. To do this I utilized the unicellular eukaryotes *Giardia lamblia*, *Giardia muris* and *Euglena gracilis*. A diRNP containing RNase P and snoRNA domains that targets tRNA<sup>Met</sup> for 2'-O-methylation and the U3 snoRNA were discovered and characterized in two *Giardia* species. Unique binding/assembly properties were determined for C/D snoRNP proteins in *G. lamblia*. A large collection of *E. gracilis* snoRNAs were discovered leading to a better understanding of Ψ-guide snoRNA structure and evolution. These findings highlight the diversity of ncRNA features that exist in less well studied eukaryotic species, and their unique functions.

## Supplementary materials

**Supplemental file 1.** Spreadsheet containing information regarding *G. muris* ncRNA library used in Chapter 2.

**Supplemental file 2.** Spreadsheet containing information regarding *G. lamblia* proteomics, RNA-seq and bioinformatic analysis used in Chapter 3.

**Supplemental file 3.** Additional image sets for immunofluorescent localization experiments discussed in Chapter 3.



## PREFACE – AUTHOR CONTRIBUTIONS

Work contained in Chapter 2 of this thesis “Analysis of *Giardia muris* ncRNAs reveals highly divergent spliceosomal, telomerase, and U3 RNAs and uncovers an RNase P-snoRNA diRNP” was performed in collaboration with Drs. Staffan Svärd and Jon Jerlström-Hultqvist at Uppsala University. Their contributions include the construction of the small ncRNA library, analysis of ncRNA synteny between *G. lamblia* and *G. muris* genomes, and assistance with identification of additional ncRNA candidates within the ncRNA library. David McWatters conceived of the study, performed all other computational analysis and experimental work, produced the figures, and wrote the manuscript/thesis chapter associated with the work.

Work contained in Chapter 4 of this thesis “RNA-Seq employing a novel rRNA depletion strategy reveals a rich repertoire of snoRNAs in *Euglena gracilis* including box C/D and  $\Psi$ -guide RNAs targeting the modification of rRNA extremities” was done collaboratively between members of Dr. Tony Russell’s lab. As outlined in the reprint agreement found in Chapter 4: David McWatters performed the bulk of bioinformatic analysis and wrote the final draft of the manuscript and produced figures. Ashley Moore performed experiments, initial bioinformatic analysis, wrote a first draft of the manuscript and produced figures. Andrew Hudson assisted with  $\Psi$ -guide snoRNA analysis.

## ACKNOWLEDGEMENTS

First, I need to acknowledge and express my immense gratitude to my supervisor and first mentor Dr. Tony Russell for giving me the opportunity to pursue my Ph.D. in his lab. Thank you for giving me the chance to follow my curiosity and pursue the research I found most exciting. I would not be the researcher I am today, and this thesis would not be what it is without his support, mentorship, encouragement, and openness.

A massive thank you to the members of my supervisory committee Drs. Igor Kovalchuk and Brent Selinger for their support and guidance in research and teaching during both my undergraduate and graduate degrees. I couldn't have been happier to have had the two of them along for the journey. Thank you to Dr. John Archibald for agreeing to serve as the external examiner for my thesis defence, I greatly appreciate him lending his time and expertise in the world of evolutionary biology to my work. I also want to express my appreciation for Dr. Steve Wiseman for chairing my thesis defence and Dr. Nehal Thakor for agreeing to act as my internal external examiner. Each of them has been wonderful to work along side at the U of L and supportive throughout my degree (though I somehow predate them both).

I also want to express my gratitude to Drs. Staffan Svärd and Scott Roy who have been amazing collaborators in research and gone out of their way to offer their support and guidance to me during my Ph.D. Much of what is contained in this thesis is built on a foundation of their fantastic work and I greatly appreciate their contributions to my success.

To all the members of the U of L Department of Biological Sciences, I hope I have been able to properly express how much I appreciate your support through the years, it has always felt like I could rely on each of you if I ever needed something. The same goes for

every undergraduate student I have had the pleasure of working with in the Russell lab. Working alongside young aspiring scientists that are so eager to learn has been one of the greatest pleasures of my time at the U of L, I hope each of you got as much out of it as I did.

Andy and Ashley; when I started in the Russell lab I had no idea what I was doing and without your patients, brilliance, and friendship things could have really gone off the rails in a hurry. I owe a huge part of the scientist I am now and any success I have had to the two of you. It did not hurt that your company made coming to the lab a whole bunch of fun! To the friends that I have made during my time at the U of L; Rhys, Nathan, Jackson, Connor, Dylan, Luc, Amanda, Tara, and Sarah your friendship is one of the greatest things I am leaving my degree with. Thank you for making my time in Lethbridge such a blast. While these years contained a lot of hard work, they have flown by and that is in no small part thanks to you.

Finally, I need to thank my family. Mom and Dad, you have supported my passion for science in every way that I could possibly imagine, but I am equally as grateful that you would have done the same no matter what I chose to pursue. I am a product of my parent's kindness, joy, compassion, and empathy and am so fortunate to be part of such a loving family. My sisters Jenn, Sarah, and Carla have been my number one supporters my whole life, a Ph.D. seems a whole lot more obtainable with a support system like them. Jalyce, I am sure I would have lost my mind during this process without you, thank you for your love and support all this time, I can't wait to see what is next for us in Boston and beyond.

## TABLE OF CONTENTS

<b>Thesis Exam Committee Members.....</b>	<b>i</b>
<b>Dedication.....</b>	<b>ii</b>
<b>Abstract.....</b>	<b>iii</b>
<b>Preface.....</b>	<b>iv</b>
<b>Acknowledgements.....</b>	<b>v</b>
<b>Table of Contents.....</b>	<b>vii</b>
<b>List of Tables.....</b>	<b>xi</b>
<b>List of Figures.....</b>	<b>xii</b>
<b>List of Abbreviations.....</b>	<b>xv</b>
 <b>Chapter 1: Introduction .....</b>	 <b>1</b>
1.1 Discovery of non-coding RNAs .....	1
1.1.1 Functions of non-coding RNAs .....	3
1.2 Overview of snoRNAs .....	7
1.2.1 C/D snoRNAs and snoRNPs .....	9
1.2.2 K-turn structure and binding .....	12
1.2.3 Biogenesis of box C/D snoRNPs .....	14
1.2.4 H/ACA snoRNAs and snoRNPs .....	17
1.2.5 snoRNPs in pre-rRNA processing .....	19
1.2.6 Additional snoRNA functions.....	23
1.2.7 ScaRNA .....	26
1.3 Understanding eukaryotic RNP diversity through studying protists.....	28
1.3.1 <i>Giardia lamblia</i> and the metamonads .....	29
1.3.2 <i>G. lamblia</i> ncRNAs .....	31
1.3.3 <i>G. lamblia</i> snoRNAs .....	32
1.3.4 <i>Euglena gracilis</i> and ncRNAs.....	33
1.3.5 <i>E. gracilis</i> snoRNAs .....	34
1.3.6 Genomic expansion and reduction .....	37
1.4 Objectives.....	37
 <b>Chapter 2: Analysis of <i>Giardia muris</i> ncRNAs reveals highly divergent spliceosomal, telomerase, and U3 RNAs and uncovers an RNase P-snoRNA diRNP .....</b>	 <b>39</b>
2.1 Introduction .....	39
2.2 Results .....	42
2.2.1 <i>G. muris</i> ncRNA candidate prediction .....	42
2.2.2 Identification of C/D and H/ACA box snoRNAs in <i>G. muris</i> .....	44
2.2.3 C/D snoRNA-guided modification sites are conserved between <i>G. muris</i> and the <i>Gu. theta</i> nucleomorph rRNAs .....	51
2.2.4 An RNase P-snoRNA diRNP targets tRNA <sup>Met</sup> 2'-O-methylation in <i>Giardia</i> species .....	53
2.2.5 Identification and analysis of Metamonad U3 snoRNAs.....	56
2.2.6 Small nuclear RNA candidates from <i>G. muris</i> .....	62
2.2.7 A highly-reduced Telomerase RNA (TER) component in <i>Giardia</i> spp. ....	66

2.3 Discussion .....	70
2.3.1 Conservation of a ncRNA processing motif in <i>G. muris</i> .....	70
2.3.2 snoRNA repertoires are conserved between <i>Giardia</i> species .....	72
2.3.3 A conserved set of snoRNA-guided modifications present in functionally important positions of ribosomal RNA in species with reduced genomes ....	73
2.3.4 Cm34 tRNA modification targeted by the <i>Giardia</i> RNase P-snoRNA diRNP is conserved throughout eukaryotes and archaea.....	75
2.3.5 Reduction in U3 snoRNA size during metamonad evolution.....	77
2.3.6 Rapid divergence of snRNAs in the diplomonads .....	79
2.3.7 Reduced <i>Giardia</i> telomerase RNAs highlight convergent evolution of TBE features.....	81
2.4 Conclusions .....	83
2.5 Materials and Methods .....	84
2.5.1 RNA extraction and sequencing.....	84
2.5.2 Annotation of ncRNAs.....	84
2.5.3 Quantification of ncRNA candidates .....	87
2.5.4 Assessing homologous snoRNA target sites.....	87
2.5.5 Synteny analysis.....	87
2.5.6 Bioinformatics identification of diplomonad U3 snoRNA homologs .....	88
2.5.7 3' rapid amplification of cDNA ends (RACE) of the RNase P-GlsR15 fusion transcript .....	88
<b>Chapter 3: Analysis of two Snu13p homologs and their associated complexes from the diplomonad <i>Giardia lamblia</i> .....</b>	<b>90</b>
3.1 Introduction .....	90
3.2 Results .....	92
3.2.1 Diplomonads possess two distinct Snu13p homologs .....	92
3.2.2 <i>G. lamblia</i> Snu13p homologs bind K-turns more weakly than other eukaryotic homologs.....	96
3.2.3 Protein-protein interactions and RNA binding properties of <i>G. lamblia</i> Nop56, Nop58 and fibrillarin .....	99
3.2.4 Genes for several C/D snoRNP assembly factors appear to be absent from diplomonad genomes.....	103
3.2.5 C/D snoRNAs but not the U4 snRNA co-precipitate with both <i>G. lamblia</i> Snu13p homologs .....	105
3.2.6 Ribosome processing complexes co-precipitate with <i>G. lamblia</i> Snu13p homologs .....	109
3.2.7 Three abundant uncharacterized proteins co-precipitate with the <i>G.</i> <i>lamblia</i> Snu13p homologs .....	114
3.2.8 Cellular localization of <i>G. lamblia</i> Snu13p homologs and Ucp proteins ....	116
3.2.9 <i>G. lamblia</i> nucleoli can form at the basal region of nuclei .....	118
3.2.10 <i>G. lamblia</i> RNase P-GlsR15 RNA forms a hetero-diRNP .....	119
3.3 Discussion .....	122
3.3.1 Insights into C/D snoRNP assembly in <i>G. lamblia</i> .....	122
3.3.2 C/D snoRNP assembly machinery in <i>G. lamblia</i> .....	124
3.3.3 <i>G. lamblia</i> Snu13p homologs do not interact with the spliceosomal U4 snRNA in vivo .....	125

3.3.4	The <i>G. lamblia</i> SSU processome is similar in composition to other eukaryotes .....	127
3.3.5	<i>Giardia</i> specific proteins may be associated with ribosome function .....	129
3.3.6	<i>G. lamblia</i> nucleoli can be present at the apical or basal side of each nuclei .....	131
3.3.7	<i>G. lamblia</i> forms an RNase P-snoRNA diRNP in vivo .....	132
3.4	Conclusions .....	134
3.5	Materials and Methods .....	135
3.5.1	Culturing of <i>Giardia lamblia</i> .....	135
3.5.2	Phylogenetic analysis of L7Ae domain proteins.....	135
3.5.3	Cloning of <i>G. lamblia</i> proteins into pAC and pAN expression vectors .....	136
3.5.4	Transfection of <i>G. lamblia</i> cells with pAC and pAN vectors .....	137
3.5.5	Cloning of <i>G. lamblia</i> proteins for expression in <i>E. coli</i> .....	137
3.5.6	Overexpression and purification of recombinant <i>G. lamblia</i> proteins in <i>E. coli</i> .....	138
3.5.7	In vitro transcription of Fluorescently labeled RNA.....	139
3.5.8	Electrophoretic mobility shift assays (EMSA).....	141
3.5.9	Western blotting of tagged <i>G. lamblia</i> proteins.....	141
3.5.10	Protein co-precipitations with <i>G. lamblia</i> Snu13p homologs .....	142
3.5.11	Analysis and annotation of <i>G. lamblia</i> proteins.....	144
3.5.12	Immunofluorescence microscopy .....	145
3.5.13	RNA co-precipitations with <i>G. lamblia</i> Snu13p homologs .....	146
3.5.14	TGIRT™-III library preparation for RNA-seq .....	147
3.5.15	Analysis of RNA sequencing libraries.....	149
3.5.16	Glycerol gradient ultracentrifugation of tagged S4 cell lysate .....	151
<b>Chapter 4: RNA-Seq employing a novel rRNA depletion strategy reveals a rich repertoire of snoRNAs in <i>Euglena gracilis</i> including box C/D and Ψ-guide RNAs targeting the modification of rRNA extremities.....</b>		<b>152</b>
4.1	Introduction .....	153
4.2	Results and Discussion.....	155
4.2.1	Preventing amplification of unwanted RNAs during RNA-Seq library construction.....	155
4.2.2	Identification of new snoRNAs.....	161
4.2.3	Importance of rRNA spacer regions and timing of snoRNA-guided rRNA modification.....	162
4.2.4	Identification and characterization of novel Ψ-guide snoRNAs.....	165
4.2.5	Evolution of Ψ-guide snoRNA species in <i>Euglena</i> .....	169
4.2.6	Orphan snoRNAs in <i>Euglena gracilis</i> .....	172
4.2.7	tRNA identification.....	174
4.2.8	U1 snRNA sequences with variable 5' ends .....	175
4.3	Materials and Methods .....	177
4.3.1	Library construction.....	177
4.3.2	Preventing amplification of large subunit rRNA fragments .....	179
4.3.3	Bioinformatic analysis .....	179
<b>Chapter 5: Conclusions and Future Directions.....</b>		<b>182</b>

<b>References .....</b>	<b>188</b>
<b>Appendix 1 – Supplementary Material for Chapter 2: .....</b>	<b>216</b>
<b>Appendix 2 – Supplementary Material for Chapter 3: .....</b>	<b>241</b>
<b>Appendix 3 – Supplementary Material for Chapter 4: .....</b>	<b>253</b>
<b>Appendix 4 – Table of oligonucleotides used in this thesis: .....</b>	<b>286</b>

## LIST OF TALBES

<b>Table 2.1.</b> List of C/D and H/ACA box snoRNAs identified in <i>G. muris</i> and their homologs in other species. ....	45
<b>Table 3.1.</b> Predicted presence of C/D snoRNP assembly factors in metamonad species. ....	105
<b>Table 3.2.</b> Categorization of proteins detected in at least two replicates of protein co-precipitation experiments. ....	111
<b>Table 3.3.</b> Protein composition of the SSU processome sub-complexes in <i>G. lamblia</i> . .	112
<b>Table A.2.1.</b> Categorization of proteins from a single replicate $\beta$ -Snul3p protein co-precipitation experiment incubated with resin for 24 hours prior to elution.	251
<b>Table A.3.1.</b> Oligonucleotides used during synthesis of the <i>E. gracilis</i> small/capped RNA library .....	276
<b>Table A.3.2.</b> Blocker oligonucleotides used during PCR amplification of cDNA synthesized from <i>E. gracilis</i> small RNA or cap-enriched RNA. ....	276
<b>Table A.3.3.</b> Reads per million (RPM) for RNAs found in single end reads for size-selected and TMG-capped small RNA libraries. ....	277
<b>Table A.3.4.</b> Characteristics of box AGAUGN snoRNAs identified in <i>Euglena gracilis</i> .....	283
<b>Table A.3.5.</b> Identity of the first nucleotide upstream of the basal stem of AGAUGN snoRNAs in <i>Euglena gracilis</i> . ....	284
<b>Table A.3.6.</b> PCR conditions used for amplification of cDNA during library construction. ....	285
<b>Table A.4.1.</b> Oligonucleotide primers used in experimental Chapters 2-4. ....	286



## LIST OF FIGURES

<b>Figure 1.1.</b> Conservation of ncRNA classes in the three domains of life. ....	4
<b>Figure 1.2.</b> General structure of C/D and H/ACA snoRNA and target RNA nucleotide modification structures. ....	11
<b>Figure 1.3.</b> Consensus structure of a K-turn motif. ....	14
<b>Figure 1.4.</b> Secondary structure of U3 snoRNA from <i>Saccharomyces cerevisiae</i> alone and base-pairing to the 35S pre-rRNA. ....	21
<b>Figure 1.5.</b> Phylogenetic tree representing the current proposed distribution of eukaryotic supergroups. ....	31
<b>Figure 1.6.</b> Unique pre-rRNA processing features of <i>Euglena gracilis</i> . ....	36
<b>Figure 2.1.</b> <i>G. muris</i> C/D snoRNAs predominantly target rRNA positions also targeted in <i>G. lamblia</i> . ....	46
<b>Figure 2.2.</b> <i>G. muris</i> snoRNAs form distinct abundance clusters among ncRNA candidates. ....	48
<b>Figure 2.3.</b> Newly identified H/ACA snoRNAs from <i>G. muris</i> . ....	50
<b>Figure 2.4.</b> An RNase P-snoRNA fusion RNA and GmsR5 have predicted tRNA targets for 2'-O-methylation. ....	54
<b>Figure 2.5.</b> Metamonad U3 snoRNA candidates maintain critical U3 features. ....	59
<b>Figure 2.6.</b> Size reduction of the U3 snoRNA in diplomonads. ....	61
<b>Figure 2.7.</b> Divergent <i>G. muris</i> spliceosomal snRNAs. ....	65
<b>Figure 2.8.</b> Comparison of <i>Giardia</i> telomerase RNA (TER) structures with other eukaryotic TERs. ....	70
<b>Figure 3.1.</b> Diplomonads possess two distinct Snul3p homologs. ....	95
<b>Figure 3.2.</b> Diplomonad Snup13p homologs form a clade within in a phylogenetic tree. ....	95
<b>Figure 3.3.</b> EMSA analysis of <i>G. lamblia</i> $\alpha$ -Snul3p and $\beta$ -Snul3p binding to the <i>G. lamblia</i> U4 5' SL or GlsR9 C/D snoRNA. ....	97
<b>Figure 3.4.</b> <i>G. lamblia</i> Snul3p homologs bind with relatively low affinity to K-turn containing RNAs. ....	97
<b>Figure 3.5.</b> EMSA analysis of <i>G. lamblia</i> $\alpha$ -Snul3p and $\beta$ -Snul3p binding to variant Gl U4 K-turn motifs. ....	99
<b>Figure 3.6.</b> <i>G. lamblia</i> Nop56p and Nop58p co-purify with His-tagged fibrillarin. ....	101
<b>Figure 3.7.</b> Analysis of Nop56-fibrillarin and Nop58-fibrillarin binding GlsR9 snoRNA <i>in vitro</i> . ....	101
<b>Figure 3.8.</b> C/D snoRNAs and RNase P are significantly enriched in co-precipitations with $\alpha$ -Snul3p and $\beta$ -Snul3p. ....	109
<b>Figure 3.9.</b> Ucp1-3 proteins detected in protein co-precipitations with <i>G. lamblia</i> Snul3p homologs and S4. ....	113
<b>Figure 3.10.</b> Cellular localization of tagged <i>G. lamblia</i> proteins. ....	117
<b>Figure 3.11.</b> <i>G. lamblia</i> nucleoli can localize to the basal side of one or both nuclei in individual cells. ....	119
<b>Figure 3.12.</b> EMSA analysis of Snul3p proteins binding the <i>G. lamblia</i> RNase P P12 K-turn. ....	120
<b>Figure 4.1.</b> Primer blocking strategy to prevent rRNA amplification during small RNA library preparation. ....	158
<b>Figure 4.2.</b> Sequencing results of a <i>Euglena gracilis</i> small RNA library before and after use of primer blocking to prevent amplification of rRNA fragments. ....	159

<b>Figure 4.3.</b> Identified <i>E. gracilis</i> snoRNAs whose guide regions base-pair with pre-rRNA intergenic sequence. ....	165
<b>Figure 4.4.</b> Examples of predicted secondary structures of <i>E. gracilis</i> AGAUGN box snoRNA species and isoforms. ....	167
<b>Figure 4.5.</b> Evolution of <i>E. gracilis</i> $\Psi$ -guide snoRNA species.....	171
<b>Figure 4.6.</b> Sequence logo of variant <i>E. gracilis</i> U1 sequences.....	177
<b>Figure A.1.1.</b> Most snoRNA homologs from <i>Giardia muris</i> and <i>Giardia lamblia</i> diverge significantly outside of guide regions. ....	220
<b>Figure A.1.2.</b> Predicted base-pairing of <i>G. muris</i> C/D snoRNAs to RNA targets.....	221
<b>Figure A.1.3.</b> <i>G. muris</i> H/ACA snoRNA candidates form conserved dual hairpin structures. ....	223
<b>Figure A.1.4.</b> Conserved snoRNA guided modification sites in <i>G. muris</i> and the <i>Gu. theta</i> nucleomorph.....	229
<b>Figure A.1.5.</b> <i>G. muris</i> and <i>Gu. theta</i> nucleomorph C/D snoRNAs share little conserved sequence outside of their guide elements.....	230
<b>Figure A.1.6.</b> <i>G. muris</i> and the <i>Gu. theta</i> nucleomorph snoRNAs target homologous positions to human snoRNAs in functionally important regions of the rRNA. ....	234
<b>Figure A.1.7.</b> The mature form of RNase P–GlsR15 is a single transcript and is conserved in <i>G. muris</i> .....	235
<b>Figure A.1.8.</b> GlsR1 lacks many key features of U3 snoRNAs.....	235
<b>Figure A.1.9.</b> Secondary structure predictions for metamonad U3 snoRNAs. ....	236
<b>Figure A.1.10.</b> Comparison of snRNA sequences from <i>G. muris</i> and <i>G. lamblia</i> .....	237
<b>Figure A.1.11.</b> Secondary structure predictions for <i>Giardia</i> U5 snRNAs. ....	238
<b>Figure A.1.12.</b> Alignment of telomerase RNAs (TER) from <i>Giardia</i> species. ....	239
<b>Figure A.1.13.</b> <i>G. muris</i> ncRNA 3' end processing motif.....	239
<b>Figure A.1.14.</b> Helix I base-pairing potential between U3 and the 18S rRNA is not conserved in all metamonads. ....	240
<b>Figure A.2.1.</b> Two Snu13p homologs are present in <i>G. lamblia</i> , <i>G. muris</i> and <i>S. salmonicida</i> . ....	242
<b>Figure A.2.2.</b> Diplomonad Snu13p homologs deviate from the general eukaryotic and archaeal consensus sequences at key residues. ....	244
<b>Figure A.2.3.</b> EMSA analysis of <i>G. lamblia</i> $\beta$ -Snu13p binding to the <i>G. lamblia</i> GlsR5 C/D snoRNA. ....	244
<b>Figure A.2.4.</b> Nop56-fibrillarin and Nop58-fibrillarin duplexes from <i>G. lamblia</i> specifically bind C/D snoRNAs. ....	245
<b>Figure A.2.5.</b> Nop56-fibrillarin and Nop58-fibrillarin bind to C/D snoRNAs more tightly than either Snu13p homolog. ....	246
<b>Figure A.2.6.</b> Truncated NUFIP $\Delta$ C cannot form a complex with either <i>G. lamblia</i> Snu13p homolog. ....	247
<b>Figure A.2.7.</b> RNA specifically precipitates with TAP tagged proteins and input libraries are consistent between transfected <i>G. lamblia</i> strains. ....	248
<b>Figure A.2.8.</b> <i>G. lamblia</i> ncRNA Candidate 12 is a C/D snoRNA.....	248
<b>Figure A.2.9.</b> Ucp protein sequences differ significantly between <i>G. lamblia</i> and <i>G. muris</i> . ....	250
<b>Figure A.2.10.</b> Localization of ribosomal protein S4 and putative Bcd1 candidate in <i>G. lamblia</i> . ....	250

<b>Figure A.2.11.</b> Sequence divergence between Snu13p homologs in different diplomonads and other eukaryotes. ....	251
<b>Figure A.2.12.</b> Tagged <i>G. lamblia</i> S4 rprotein sediments in glycerol gradients.....	252
<b>Figure A.2.13.</b> Ucp3 co-precipitation enriches distinct RNAs. ....	252
<b>Figure A.3.1.</b> Illustration of the oligo blocker strategy containing sequences and experimental results. ....	254
<b>Figure A.3.2.</b> Identified <i>E. gracilis</i> box C/D snoRNAs and their predicted target sites in the rRNA for 2'- <i>O</i> -methylation events.....	258
<b>Figure A.3.3.</b> Identified box AGAUGN snoRNAs and their predicted rRNA target sites for pseudouridine ( $\Psi$ ) formation in <i>E. gracilis</i> .....	260
<b>Figure A.3.4.</b> Newly identified <i>Euglena gracilis</i> snoRNA sequences.....	268
Description for C/D and AGAUGN RNAs appear at the start of their respective list of RNAs above .....	268
<b>Figure A.3.5.</b> Sequences of predicted orphan snoRNAs identified from a <i>Euglena</i> small RNA library.....	271
<b>Figure A.3.6.</b> Predicted secondary structures of identified <i>E. gracilis</i> box AGAUGN modification guide snoRNAs.....	274
<b>Figure A.3.7.</b> Newly identified tRNA sequences from <i>E. gracilis</i> . ....	275

## LIST OF ABBREVIATIONS

ASE	Anti-sense element
bp	Base-pairs
CM	Covariance model
CPM	Counts per million
crRNA	CRISPR RNA
EMSA	Electrophoretic mobility shift assay
ER	Endoplasmic reticulum
ETS	Externally transcribed spacer
hnRNP	Heterogenous nuclear ribonucleoprotein
ITS	Internally transcribed spacer
K-turn	Kink-turn
LECA	Last eukaryotic common ancestor
LSU	Large subunit
LUCA	Last universal common ancestor
MLO	Membraneless organelle
MRO	Mitochondrial related organelle
ncRNA	non-coding RNA
NE	Nuclear envelope
Nm	2'- <i>O</i> -methylation
nt	Nucleotides
NTP	Nucleotide triphosphate
PCR	Polymerase chain reaction
psnoRNA	processed small nucleolar RNA
RBD	RNA binding domain
RNP	Ribonucleoprotein
RPM	Reads per million
RPR	RNase P RNA
RPR	RNase P RNA
rRNA	ribosomal RNA
SAM	s-adenosyl-methionine
scaRNA	small Cajal body associated RNA
SL	Stem loop
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
sRNA	small RNA
SSU	Small subunit
TAP	Tandem affinity purification
TMG	Trimethyl guanosine
T-PK	Template and pseudoknot containing loop
TPM	Transcripts per million
TR	Telomerase RNA
tracr RNA	trans-activating CRISPR RNA
tRNA	transfer RNA

## Chapter 1: Introduction

### 1.1 Discovery of non-coding RNAs

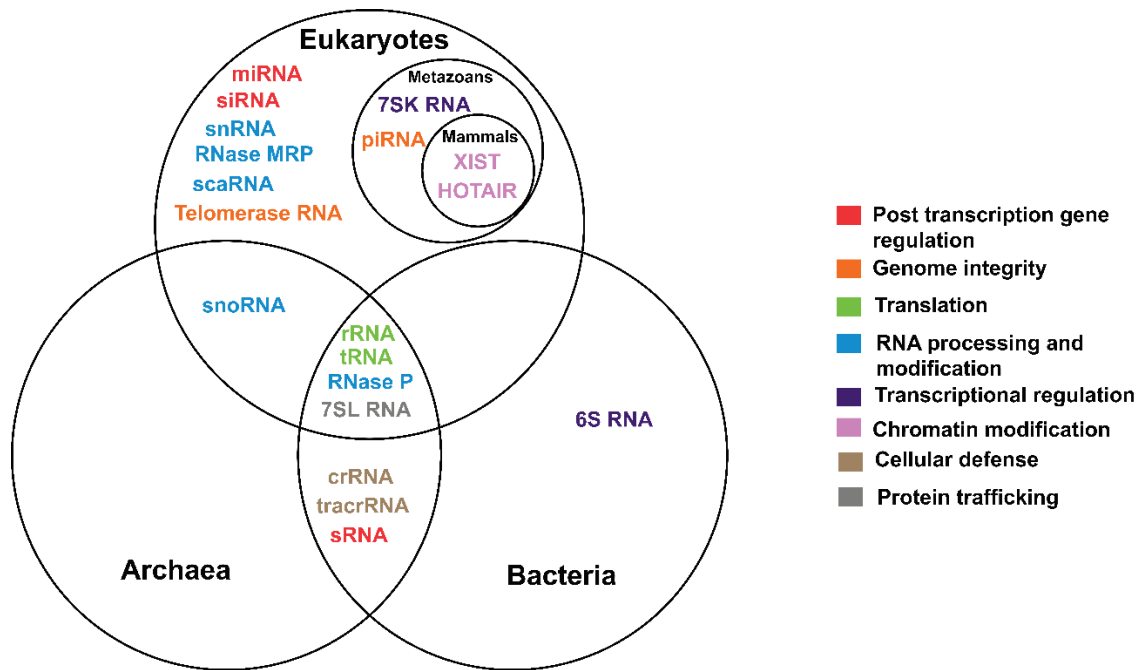
The discovery of DNA as the genetic material of inheritance by Avery, MacLeod and McCarty and subsequent unraveling of its structure by Crick, Watson and Franklin began a revolution in molecular genetics (Avery et al. 1944; Watson and Crick 1953). These discoveries generated great interest in understanding the flow of genetic information resulting in the discovery of messenger RNA (mRNA) and its central role in protein production (Crick 1958). The idea that information flows from DNA to mRNA to protein was later coined the “Central Dogma of Molecular Biology” by Francis Crick (Crick 1970). It was also during this period that the cellular machinery required for interpreting the mRNA message to produce proteins was discovered. Initially described as small granules found attached to the endoplasmic reticulum (ER) and called microsomes in the 1950s (Palade 1955), the next several decades revealed the ribosome to be a large complex comprised of RNA and proteins that is critical for the conversion of the mRNA message to a functional protein product. This was the first large ribonucleoprotein (RNP) molecule discovered and the first RNA described with a function outside of transferring genetic messages; that is, the first non-coding RNA (ncRNA). Only a few years later a second class of ncRNA called transfer RNA (tRNA) was discovered (Hoagland et al. 1958). tRNA acts as an adapter molecule, ensuring incorporation of the correct amino acid into the nascent polypeptide by the ribosome through base-pairs formed between the tRNA anti-codon and mRNA codons. The structure and function of ribosomes has since been extensively studied, revealing that not only is the ribosomal RNA (rRNA) an important structural component of the ribosome, but is in fact a ribozyme, carrying out the catalytic functions of the ribosome (Ban et al. 2000). Prior to this discovery the description of other ribozymes including RNase P and self splicing RNAs in *Tetrahymena* (Kruger

et al. 1982; Guerrier-Takada et al. 1983), led Walter Gilbert to propose the “RNA world hypothesis” in the 1980s, where he suggested that RNA, which is capable of storing genetic information, structural scaffolding, and catalysis, is the progenitor molecule to cellular life, predating both DNA and protein (Gilbert 1986). The more recent description of the catalytic role of RNAs in eukaryotic intron splicing along with the discovery of other ribozymes lends credence to Gilbert’s proposal which highlights the ancient origin of functional non-coding RNAs. While the RNA world remains a hot topic of discussion and research, there is no doubt RNA is a far more functionally diverse macromolecule than initially appreciated (Fica et al. 2013). These discoveries expand the Central Dogma, showing that the flow of information is not so linear.

Many new classes of ncRNA have been discovered and characterized since the 1950s, revealing that tRNA and rRNA are not exceptional cases. As ncRNAs were examined in a greater number of species it was determined that many ncRNA classes are conserved across large evolutionary distances in diverse phyla. This conservation can range from detection in multiple eukaryotic lineages to some ncRNAs that trace back through all three domains of life to the last universal common ancestor (LUCA). Further studies showed that ncRNAs are not limited to functions of conserved ancient origin, and unique ncRNA classes have emerged relatively recently in many lineages, performing more niche roles in small homogenous groups of species. We now know that even the most ancient classes of ncRNAs are not monolithic or entirely fixed in their function, as individual RNAs of these conserved classes have been found to perform completely unique functions, distinct from other members of the same class. These discoveries show the versatility and adaptability of ncRNAs to function as major drivers of cellular function and serve as playgrounds for valuable evolutionary innovation.

### ***1.1.1 Functions of non-coding RNAs***

Classes of ncRNAs that are conserved across species can consist of a single RNA with a conserved function that acts on one or more target (e.g. RNase P), several RNAs that each perform unique functions as part of a pathway (e.g. small nuclear RNAs), or multiple RNAs that perform the same function on many different targets (e.g. small nucleolar RNAs). Depending on the species, the number of RNAs belonging to a given ncRNA class can also vary significantly. NcRNAs are found in a broad range of sizes from <20 to thousands of nucleotides in length and display a diverse spectrum of functions. NcRNAs are present throughout the three domains of life; however, the majority of well-conserved ncRNA classes described to date are found in eukaryotes. This is likely due in part to typically higher chromosomal DNA content but a significantly lower proportion of protein coding DNA, and historically a more significant focus has been placed on the study eukaryotic ncRNAs (Figure 1.1). Regardless, current data shows that from the processing and modification of precursor rRNA (pre-rRNA) to post-transcriptional regulation of gene expression it has become increasingly clear that few conserved cellular processes are untouched by ncRNAs. Below, key features and functions for some of these important ncRNA classes are described.



**Figure 1.1. Conservation of ncRNA classes in the three domains of life.** The three major domains and relevant domain subgroups are labeled in black. ncRNA classes are labeled in colour according to their most prominent cellular function.

Some of the most well-conserved ncRNAs are those consisting of single RNAs that predominantly perform one known essential function. Conserved in all three domains of life, RNase P is an essential endonuclease that cleaves the 5' leader sequence from pre-tRNAs during maturation (Esakova and Krasilnikov 2010; Altman 2011). RNase P RNPs are made up of a single RNA and a collection of core proteins, ranging from one protein in bacteria to ten in some eukaryotes (Guerrier-Takada, et al. 1983; Altman 2011). Bacterial, archaeal, and eukaryotic RNase P RNAs (RPR) are ribozymes capable of catalysis in the absence of their respective protein partners, albeit at significantly lower specificity and catalytic efficiency (Guerrier-Takada, et al. 1983; Pannucci et al. 1999; Kikovska et al. 2007). During evolution (i.e. from bacteria to archaea to eukaryotes), RPRs became less efficient catalysts, showing increased reliance on protein components (Altman 2011). The evolutionarily-related RNase MRP is highly similar to RNase P



in both RNA structure and protein composition, sharing the same set of core proteins and two additional RNase MRP specific proteins Snm1 and Rmp1 (Schmitt and Clayton 1993; Lygerou et al. 1996; Lindahl et al. 2009; Esakova and Krasilnikov 2010; Lan et al. 2020). However, RNase MRP is specific to eukaryotes and primarily functions in the cleavage of pre-rRNA within ITS1 (cleavage site A3 in yeast) (Chu et al. 1994), though targeting of the 5' UTRs of *CLB2* (Gill et al. 2004) and *CTSI* (Aulds et al. 2012) mRNAs in yeast have also been observed, indicating a role in cell cycle regulation and potentially an even more diverse target range.

The 7SL RNA is part of the signal recognition particle (SRP), a universally conserved RNP required for co-translational targeting of nascent transmembrane and secretory proteins to the endoplasmic reticulum in eukaryotes or plasma membrane in archaea and bacteria (Akopian et al. 2013). The role of the 7SL RNA within the SRP is still not completely understood, and likely varies between bacteria, archaea, and eukaryotes. However, studies show the RNA may act as a scaffold for protein assembly and help facilitate interaction and activation of GTPases in both the SRP RNP and membrane localized receptor (Gupta et al. 2021). In contrast, the telomerase RNP is only conserved throughout eukaryotes where it is responsible for extension of telomeres to prevent loss of genetic information at the end of linear chromosomes during DNA replication (Musgrove et al. 2018). The telomerase RNA (TER) acts as a binding platform for telomerase proteins including the telomerase reverse transcriptase (TERT), but also contains the template nucleotide sequence used in directing extension of telomeric DNA (Greider and Blackburn 1989).

The clearest example of a ncRNA class consisting of a set of highly conserved RNAs acting in the same pathway is the small nuclear RNAs (snRNA). Removal of introns from eukaryotic pre-mRNA is performed by the spliceosome, a ubiquitous multi-megadalton eukaryotic RNP complex made up of a set of core snRNPs each nucleated around one of five snRNAs: U1, U2, U4, U5, and

U6 (Zhang et al. 2019; Wilkinson et al. 2020). The snRNAs each associate with many protein factors and unlike other large RNPs the spliceosome is highly dynamic, with large collections of components and even whole snRNPs coming and going throughout the processes of intron removal (Wilkinson, et al. 2020). Each snRNA performs an important role in splicing including recognition of splice sites and the branch-point sequence of the pre-mRNA (U1, U2, and U6), positioning of exons (U5), and splicing catalysis (U6) (Mount et al. 1983; Parker et al. 1987; O'Keefe and Newman 1998; Fica, et al. 2013). The U4 snRNA is unique in that its primary role is to chaperone U6 prior to its association with the intron and U2 to form the catalytic spliceosome, and it does not associate with the intron itself (Bringmann et al. 1984).

Other classes of ncRNAs are less well-conserved across species but play critical roles in the species in which they are found. These RNA classes usually consist of many RNAs that may vary in overall sequence and possess different targets but perform the same function and share common sequence elements, length, or structural features. For example, three major classes of these type of small ncRNAs involved in post-transcriptional regulation of gene expression have been described in eukaryotes. In each case the ncRNA guides a nuclease containing RNP to a target RNA molecule recognized through varying degrees of base-pairing interactions. MicroRNAs (miRNAs) (~18-21 nt) guide RNP complexes to specific mRNAs, targeting them for cleavage and degradation or binding to them to induce translational repression (Gebert and MacRae 2019). Endogenous small interfering RNAs (siRNA) (~22 nt) are derived from double stranded RNA and form RNP complexes with the same machinery as miRNA to direct cleavage of RNA targets, but require perfect complementarity in this base-pairing interaction (Okamura and Lai 2008). Finally, PIWI-interacting RNAs (piRNA) (24-35 nt) are found exclusively in metazoan

germ line cells where they repress transposable genetic elements (transposons) to help maintain genome integrity (Hirakata and Siomi 2016).

An increasing number of long non-coding RNAs (lncRNA), a broadly defined group of RNAs >200 nt in length that do not code for a protein product, have been detected in eukaryotes. Many lncRNAs have undefined functions, but roles in X-chromosome inactivation (Xist) and regulation of chromatin dynamics (HOTAIR) are characterized in mammals (Penny et al. 1996; Rinn et al. 2007). CRISPR and trans-activating CRISPR RNAs (crRNA and tracrRNA) have also been extensively studied over the last decade (Hille et al. 2018). These RNAs are part of CRISPR-Cas systems that are natively found in prokaryotes where they form a type of adaptive defense system against bacterial plasmid and viral infection. Small RNAs (sRNA) found in bacteria and archaea are a broadly defined group of ncRNAs, including anti-sense RNA (asRNA), that predominantly function in post-transcriptional gene regulation using a diverse set of mechanisms (Wagner and Romby 2015).

The presence of large amounts of non-coding DNA in many eukaryotic genomes and evidence for its transcription, along with the many organisms with unique biology whose genomes remain unexplored suggest there may yet be more ncRNA classes with novel functions to be discovered.

## **1.2 Overview of snoRNAs**

One of the most critical processes in any cell is the production of functional ribosomes, which requires the coordination of multiple processing, modification and assembly steps. Small nucleolar RNAs (snoRNAs) are a class of ncRNAs named for their discovery and localization in the nucleolus, a sub-nuclear membrane-less organelle (MLO) found in the eukaryotic nucleus. SnoRNA homologs have also since been identified in archaea indicating the ancient evolutionary

origin of these RNAs (Omer et al. 2000). SnoRNAs associate with collections of core proteins to form ribonucleoprotein (RNP) complexes that function primarily in the modification and processing of pre-rRNA (Reichow et al. 2007; Watkins and Bohnsack 2012). The role of the snoRNA is to guide the RNP complex to the correct site of the target RNA, which is achieved primarily through Watson-Crick base-pairing interactions between regions of the snoRNA and the target RNA. This mechanism allows the same conserved core proteins to specifically target many different sites depending on the particular snoRNA in the complex and its unique base-pairing (therefore unique guide potential). SnoRNAs are grouped into two distinct classes based on the presence of conserved sequence elements. In eukaryotes, box C/D snoRNAs contain evolutionarily conserved C (RUGAUGA) and D boxes (CUGA) at their terminal 5' and 3' ends respectively and often more degenerate C' and D' boxes internally (where R is a purine) (Figure 1.2A). In archaea, the C'/D' box elements are more well-conserved and the RNAs are often shorter (Gaspin et al. 2000; Omer, et al. 2000). H/ACA snoRNAs are most often composed of two stem loop structures split by a single-stranded hinge region containing an H box (ANANNA) and tailed by an ACA sequence near their 3' ends (where N is any nucleotide) (Figure 1.2B) (Ganot, Caizergues-Ferrer, et al. 1997). Archaeal H/ACA RNAs frequently only contain a single stem loop followed by the ACA motif, however two and even three-stemmed H/ACA snoRNAs are also found (Tang et al. 2002; Watkins and Bohnsack 2012).

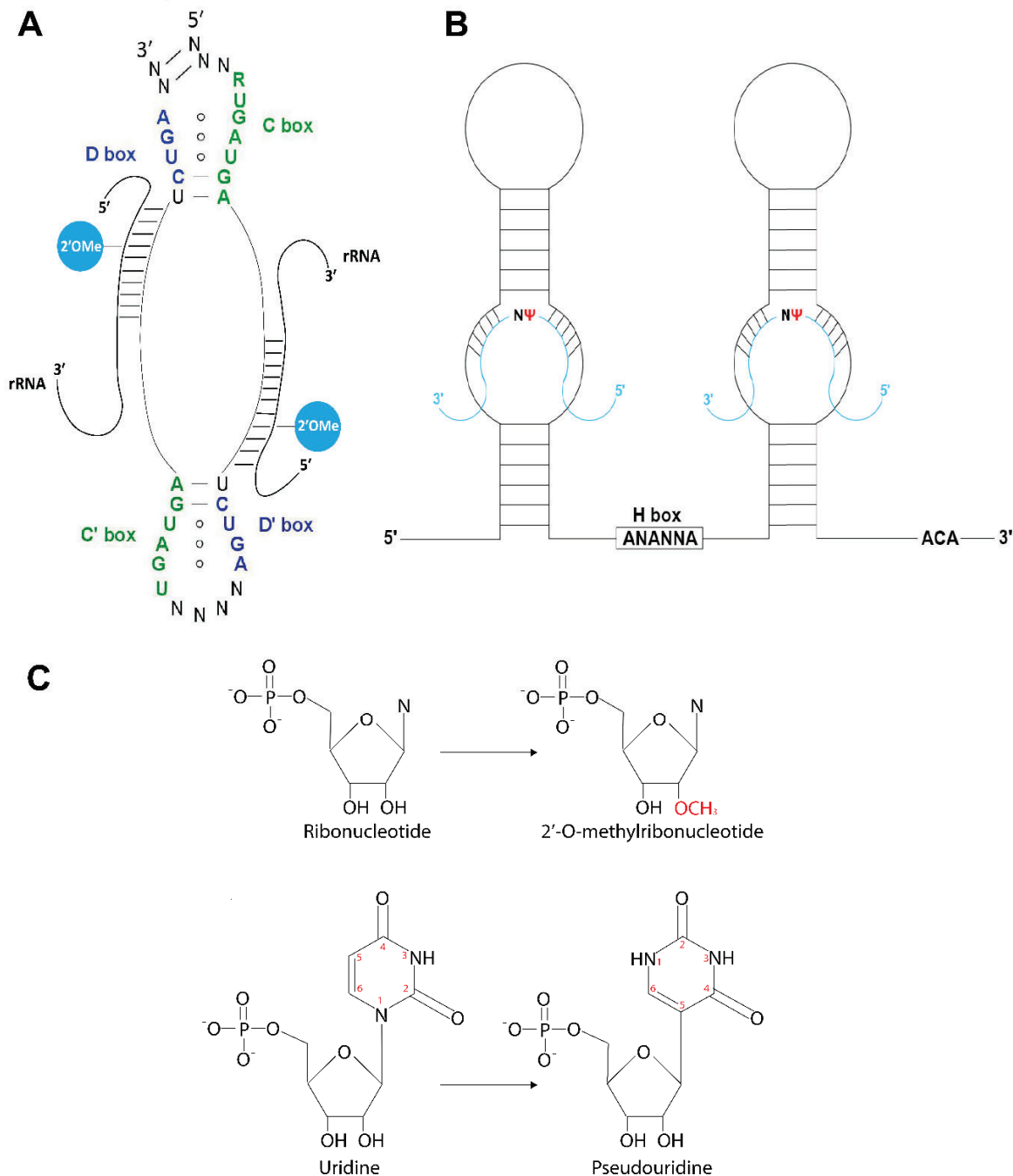
C/D and H/ACA snoRNPs are responsible for methylation at the 2'-OH group of the ribose sugar (2'-O-Me) in RNA backbones and conversion of uridine to pseudouridine ( $\Psi$ ) nucleotides respectively (Figure 1.2C) (Kiss-László et al. 1996). These modifications can be found throughout rRNA but appear at higher concentrations in functionally important regions including the peptidyl transferase centre (PTC) and mRNA decoding site, adding to ribosome stability through an

increase in base stacking and hydrogen bonding potential (Ganot, Bortolin, et al. 1997; Sloan et al. 2017). 2'-O-Me can also reduce the susceptibility of RNA to nucleolytic degradation. While loss of individual modifications often has little effect on cells, examples have been described where deletion of a snoRNA(s), leading to loss of the modification, is deleterious (Liang, Liu, et al. 2007; Baudin-Baillieu et al. 2009). Importantly, loss of several modifications, especially those that are found in clusters, have been shown to have negative effects on translational fidelity and cell viability. Intriguing recent evidence has sparked new interest in the idea of ribosome heterogeneity (Genuth and Barna 2018). Central to this concept is the idea that collections of ribosomes can differ in their protein constituents, protein isoforms, rRNA isoforms, or the modification status of individual nucleotides or residues. This can result in specialized ribosomes that may be optimized for translation: i) of certain mRNA transcripts, ii) in different cell types in multicellular organisms, or iii) under particular conditions (Sloan, et al. 2017; Genuth and Barna 2018). The extent to which ribosome heterogeneity exists in cells and their exact roles are still under investigation.

### ***1.2.1 C/D snoRNAs and snoRNPs***

As described above, the primary functions of C/D snoRNPs are the modification and processing of pre-rRNA during ribosome biogenesis (Reichow, et al. 2007). Most identified snoRNAs act as guides to direct 2'-O-Me of the nucleotide ribose sugar (Figure 1.2C) (Watkins and Bohnsack 2012). C/D RNPs in archaea contain three core proteins. The RNA binding protein L7Ae recognizes and binds the K-turn and K-loop structures formed by the box elements (Kuhn et al. 2002). Fibrillarin, the catalytic component of the complex, is an S-adenosyl-methionine (SAM) dependent methyltransferase that catalyzes transfer of the methyl group from SAM to the ribose sugar (Omer et al. 2002). Nop5 bridges protein interactions between the C/D and C'/D' boxes by forming a homodimer via its coiled-coil domains. It associates with fibrillarin using its

N-terminal domain and forms a composite binding surface between its C terminal domain and the L7Ae bound on the K-turn structure (Ye et al. 2009). Eukaryotic C/D snoRNPs possess four core proteins: Snu13p (formerly named 15.5K in humans) and fibrillarin which are homologs of archaeal L7Ae and fibrillarin and perform similar functions; and two Nop5 homologs, Nop56 and Nop58 that together form a heterodimeric complex (Tollervey et al. 1993; Watkins et al. 2000; Cahill et al. 2002; Galardi et al. 2002; Watkins and Bohnsack 2012). The presence of two distinct Nop proteins results in an asymmetric RNP structure in eukaryotes where Nop58 assembles on the C/D box motif and Nop56 on the C'/D' motif (Cahill, et al. 2002). While archaeal L7Ae can bind both K-turn and K-loop motifs, eukaryotic Snu13p only recognizes K-turn motifs (Szewczak et al. 2002).



**Figure 1.2. General structure of C/D and H/ACA snoRNA and target RNA nucleotide modification structures.** The general structure of C/D snoRNAs (**A**) bound to rRNA target regions, with the methylated nucleotide indicated with a blue circle and nucleotides of the conserved box elements in blue (D/D') and green (C/C'). H/ACA snoRNA (**B**) bound to rRNA targets (light blue) with the modified uridine depicted as Ψ. Both the C/D and H/ACA RNAs shown are dual guide snoRNAs, each capable of targeting 2 different sites. (**C**) Structural depiction of methylation of the ribose 2'-OH group and conversion of uridine to pseudouridine.

Substrate recognition is facilitated by base-pairing interactions formed between the target RNA and guide region (also called the antisense element (ASE)) of C/D snoRNAs found in regions directly upstream of either the D or D' boxes (Figure 1.2A) (Kiss-László, et al. 1996). The guide mechanism of C/D snoRNPs follows a strict rule for target nucleotide identification known as the N+5 rule in which the nucleotide of the target RNA to be modified is always the one base-paired to the fifth nucleotide upstream of the D or D' box (Kiss-László, et al. 1996). A handful of apparent exceptions to this rule exist in which single yeast snoRNAs modify two target nucleotides within the same guide region, either adjacent or one nucleotide apart (van Nues and Watkins 2016). This results from the ability of these snoRNAs to form alternative overlapping or extended D' boxes which likely form alternative base-pairs with the C' box, resulting in repositioning of Nop56 and fibrillarin and affecting modification site selection in a subset of RNPs. These guide-substrate pairings are most often approximately 10 base-pairs long in archaea and between 10 and 21 bp long in eukaryotes (Yang et al. 2016). However, analysis of guide-substrate base-pairing in both archaea and eukaryotes found that pairing beyond 10 is restricted and can even interfere with catalytic activity of the complex (Yang, et al. 2016; Yang et al. 2020). It has been suggested these predicted longer base-pairings may be used in initial substrate recognition in eukaryotes and could help prevent premature folding of the rRNA that would have to be disrupted prior to catalysis. While many box C/D snoRNAs guide a single modification site, some can act as “dual” guides targeting modifications using the ASE upstream of both their D and D' boxes.

### ***1.2.2 K-turn structure and binding***

K-turns are an RNA structural motif made up of two base-paired stems interrupted by an asymmetric bulge (Klein et al. 2001; Lilley 2012; Huang and Lilley 2016, 2018). In a standard K-turn the first stem, known as the canonical (C) stem consists of canonical Watson-Crick base-pairs,



usually G-C. The second stem contains a tandem pair of sugar edge-Hoogsteen G<sup>o</sup>A sheared base-pairs and is referred to as the non-canonical (NC) stem (Figure 1.3). Standard K-turns contain a 3+0 (nt) bulge, but structures exist with a varying number of nucleotides on either side of the bulge (Huang and Lilley 2018). The adenosines of the G<sup>o</sup>A pairs form A minor interactions with the minor groove of the canonical stem resulting in a sharp approximately 50° kink in the plane of phosphodiester backbone causing the first (L1) and second (L2) nucleotides of the loop to stack on the C and NC stems respectively and the final nucleotide of the bulge (L3) to be splayed out away from the plane of the helices, becoming solvent exposed. This also causes an opening up of the major groove, positioned on the outside of the structure. A related structure called a K-loop is structural similar to the K-turn but replaces the canonical stem with an unpaired loop (Nolivos et al. 2005). K-turns can be found as independent structures, in interactions with other RNA elements forming K-junctions, or bound by members of the L7Ae protein family, which includes homologs of Snu13p, Nhp2p, and ribosomal proteins S12e, S30, and L7Ae (Koonin et al. 1994). Based on crystal structures of various archaeal and eukaryotic L7Ae family proteins bound to K-turns, a model for binding has been proposed (Huang and Lilley 2018). A conserved basic  $\alpha$  helix binds in the opened major groove of the NC stem including sequence-specific interactions with guanines of the G<sup>o</sup>A pairs. This interaction is stabilized by non-specific electrostatic interactions with basic residues of an adjacent  $\beta$  sheet. A hydrophobic loop contacts the first two nucleotides of the loop of the K-turn via hydrogen bonding and Van der Waals interactions. Together these features allow specific and tight recognition of the kinked structure.



The C/D snoRNP assembly pathway in eukaryotes has still not been completely deciphered, but significant strides have been made in identifying the different assembly stages and proteins involved in yeast and human models. SnoRNP assembly relies on the Hsp90 R2TP chaperone/co-chaperone complex (also known as the PAQosome (Houry et al. 2018)) which in yeast consists of the proteins Rvb1, Rvb2, Tah1p, and Pih1p (RUVBL1, RUVBL2, RPAP3, and PIH1D1 respectively in humans) (Zhao et al. 2008). Rvb1/2 belong to the AAA+ ATPase family of proteins and form a heterohexameric ring structure (Gorynia et al. 2011; Rivera-Calzada et al. 2017; Muñoz-Hernández et al. 2019) involved in the assembly of many RNP complexes in yeast and humans including C/D and H/ACA snoRNPs, U4 and U5 snRNPs, Telomerase RNP and a number of others (Boulon et al. 2008; Cloutier et al. 2017; Malinová et al. 2017; Rivera-Calzada, et al. 2017; Houry, et al. 2018). Association of other C/D assembly factors rely on Rvb1/2 being in the ATP bound state and these proteins are predicted to be important for final snoRNP maturation by triggering the detachment of other assembly factors.

In addition to the R2TP complex, three more proteins have been identified as important in both yeast and human C/D snoRNP assembly: Rsa1p, Hit1p, and Bcd1p (NUFIP, ZNHIT3, and ZNHIT6 in humans (Massenet et al. 2017)). Rsa1p binds directly to Snu13p through its central PEP domain and binds Hit1p via its C terminal end, the latter interaction is important for maintaining cellular levels of Rsa1p (Boulon, et al. 2008; Rothé et al. 2013; Rothé et al. 2014). Rsa1p is also involved in bridging the interaction between Snu13p and the R2TP complex by binding the Tah1-Pih1 dimer (Quinternet et al. 2015). This interaction is thought to contribute to the transfer of Nop58 from Pih1 to Rsa1p through competition for binding, thereby loading Nop58 onto the Snu13p-Rsa1p complex. In addition to a function in modification guide snoRNPs, Rsa1p was also found to contribute to 3' end processing of the specialized U3 snoRNA and increases

affinity of Snu13p to the U3 C'/D box in yeast *in vitro* (Rothé et al. 2017). Initial analysis of Bcd1p determined its role in loading Nop58 from the R2TP complex onto the Snu13p-Rsa1p complex (Khoshnevis et al. 2019; Paul et al. 2019). Bcd1p depletion and chromatin immunoprecipitation (ChIP) studies in yeast indicate that it may also be important in recruitment of Pih1p, Rvb2, and fibrillarin to the pre-snoRNP because in its absence, only Nop56 and Snu13p are efficiently recruited to C/D snoRNA loci. The association of Bcd1p with the immature pre-snoRNP complex likely occurs early in assembly as it has been detected bound to the intron containing form of the pre-U3 snoRNA in yeast and removal of this intron is believed to occur rapidly following transcription.

Currently two models have been proposed as possible mechanisms for C/D snoRNP assembly. The first is a protein-only mechanism in which the core proteins Snu13p and Nop58 associate with the R2TP complex along with Bcd1p, Rsa1p, and Hit1p (Bizarro et al. 2014; Massenet, et al. 2017). This protein-only complex is then loaded onto the snoRNA at which point a conformational switch is induced triggering the release of most assembly factors. The second mechanism suggests that Snu13p first binds to the nascent C/D snoRNA co-transcriptionally and nucleates formation of the complex. The R2TP complex associates with Nop58 through interactions with Pih1p and is then recruited to the pre-C/D RNP. Nop58 is then transferred from Pih1p to Rsa1p through competitive binding (Quinternet, et al. 2015; Massenet, et al. 2017). This is likely facilitated by direct contacts between Bcd1p and Snu13p, with evidence that depletion of Bcd1p results in failure to recruit Nop58 efficiently to nascent snoRNAs (Khoshnevis, et al. 2019; Paul, et al. 2019). Importantly, the protein-only complex was described in humans while the later Bcd1p experiments were performed in yeast. It is possible that the assembly pathways differ between the two species. A strong and specific interaction between Nop58 and the newly

characterized C/D assembly factor protein NOPCHAP1 (formerly C12orf45) in human cells was also recently observed (Bizarro, et al. 2014; Abel et al. 2021). NOPCHAP1 is a proposed PAQosome cofactor that promotes Nop58 association with C/D snoRNPs but is apparently absent in a variety of eukaryotic lineages including Saccharomycotina fungi, insects, and Euglenozoa, indicating that there are likely species or clade-specific differences in eukaryotic assembly pathways. Currently it is still unclear how Nop56 and fibrillarin are incorporated but evidence that the depletion of Bcd1p does not prevent association of Nop56 with C/D snoRNA loci in yeast means this occurs in a manner independent of the assembly factors identified so far, and Nop56 may be recruited directly by Snu13p bound to the K-turn (Paul, et al. 2019). Together these findings suggest that while there does appear to be conserved C/D assembly factors there is likely not a single universal machinery for assembling C/D snoRNPs in eukaryotes.

Recently the first *in vitro* reconstitution of a catalytically active eukaryotic C/D snoRNP was achieved using recombinant proteins from the thermophilic yeast *Chaetomium thermophilum* (Yang, et al. 2020). The core proteins were able to assemble into RNPs capable of catalyzing site-specific methylation of target rRNA fragments in the absence of any additional factors. However, in the absence of assembly factors the proteins were unable to properly form asymmetric complexes. RNPs were formed that contained two copies of either Nop56 or Nop58 and these were significantly less efficient at modifying RNA, suggesting that part of the role of some assembly factors in eukaryotes is likely to help facilitate assembly of the asymmetric C/D snoRNP complex to ensure efficient methylation.

#### **1.2.4 H/ACA snoRNAs and snoRNPs**

H/ACA snoRNAs guide RNA modifications through bipartite base-pairing to the target RNA using nucleotides found on either side of an internal bulge in the snoRNA known as the

pseudouridylation pocket (Figure 1.2B) (Ganot, et al. 1997). This pairing leaves the target uridine and one nucleotide directly downstream “accessible” facilitating isomerization of the U to  $\Psi$ , which requires the breaking of the N1-C1' glycosidic bond, rotation of the base 180° and formation of a new C5-C1' bond (Figure 1.2C) (Yu and Meier 2014). In H/ACA snoRNAs containing multiple stems, the pseudouridylation pocket of one (single) or both (dual) stems can act as guides for modification. H/ACA snoRNAs associate with a set of four core proteins L7Ae/Nhp2, Nop10, Gar1, and Cbf5/Dyskerin in archaea and humans, respectively (Watkins and Bohnsack 2012). Archaeal H/ACA sRNAs contain a K-turn or K-loop in their upper stem a feature shared with C/D snoRNAs, which is bound by L7Ae (Rozhdestvensky et al. 2003). L7Ae is unable to bind the other H/ACA proteins in the absence of RNA but once RNA-bound, joins Cbf5 and Nop10 to form a composite surface that jointly interacts with the upper stem of the RNA (Baker et al. 2005; Li and Ye 2006). Additional protein-RNA contacts occur between the PUA domain of Cbf5 and the ACA motif adjacent to the base of the lower stem. Together these interactions contribute to proper positioning of the target uridine in the catalytic pocket of Cbf5. This positioning is constrained by the interaction of Cbf5 with the ACA motif resulting in a generally conserved distance of 14-16 nucleotides between the box motif and the target uridine (Ni et al. 1997). Gar1 is not involved in RNA binding and associates with the complex via protein-protein interactions with the thumb loop motif of Cbf5 (Duan et al. 2009).

Eukaryotic H/ACA snoRNAs lack K-turn structures and are bound in this region by Nhp2, an L7Ae family protein that does not specifically recognize K-turn or K-loop motifs (Henras et al. 2001; Li,Duan, et al. 2011). *In vitro* analysis of yeast H/ACA snoRNPs determined that while Nhp2 does bind RNA, the interaction is much weaker than for L7Ae in archaea (Caton et al. 2017). Instead, Nhp2 can form stable protein-protein interactions with the Cbf5-Nop10 dimer which

likely helps facilitate its association with the complex (Wang and Meier 2004; Li, et al. 2011). Despite this weaker binding Nhp2, like L7Ae in archaea, is required for efficient rates of pseudouridylation *in vitro* and accumulation of H/ACA snoRNAs *in vivo* (Henras, et al. 2001; Baker, et al. 2005; Liang et al. 2009; Caton, et al. 2017). Strikingly, the Nop10-Gar1-Cbf5 complex is able to bind RNA with sub-nanomolar affinities, but this interaction is non-specific. This led to the proposal that this trimer is the primary means of RNA binding, with Nhp2 instead playing a role in altering the structure of the complex for proper positioning of the target uridine (Caton, et al. 2017). In addition to its role in catalysis, the Cbf5 thumb loop motif binds the RNA duplex formed between the snoRNA and target (Duan, et al. 2009; Liang, et al. 2009; Li, et al. 2011; Yang et al. 2012). Gar1 binds to this Cbf5 motif, an interaction that is important for substrate turnover in both archaea and eukaryotes.

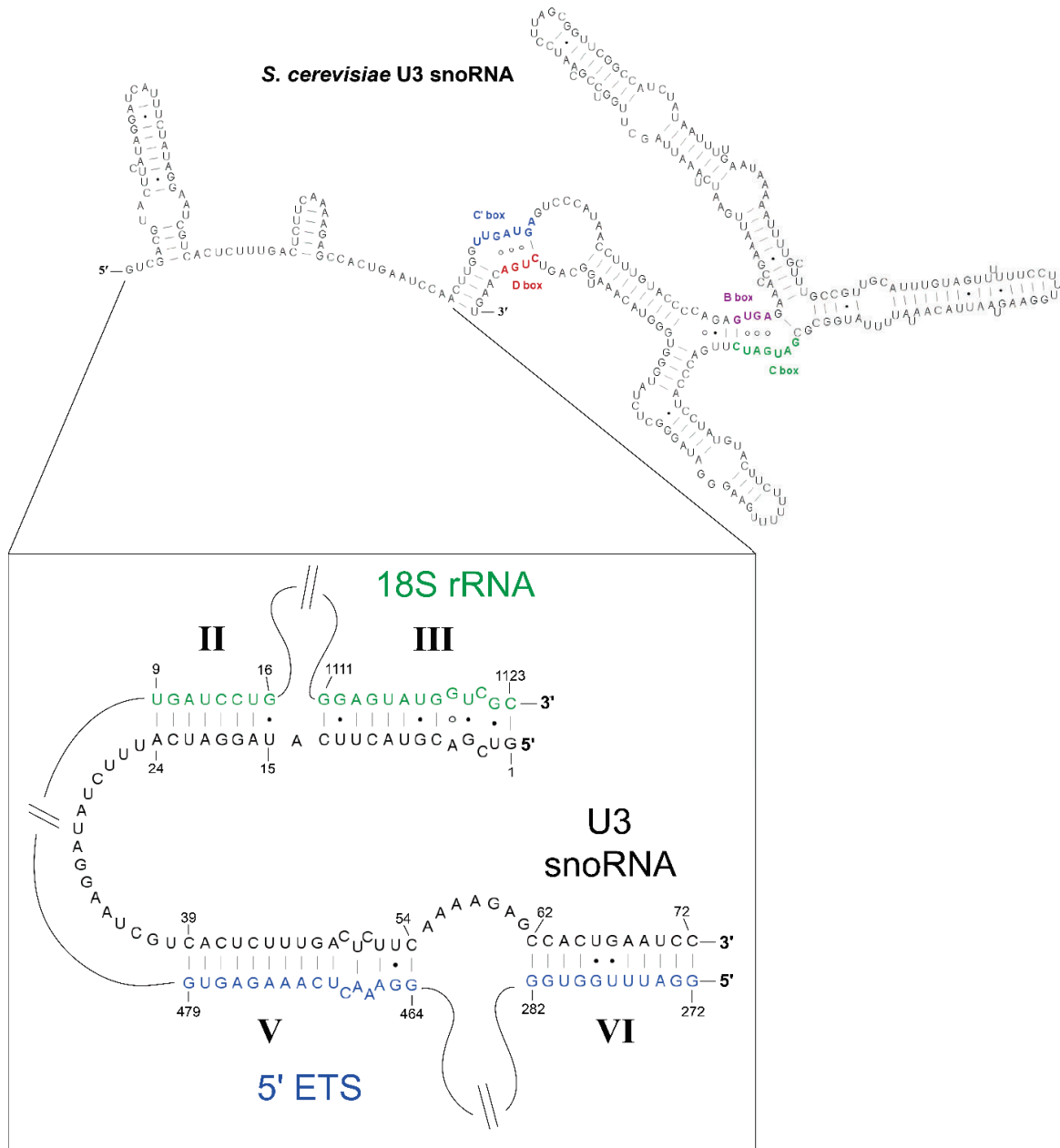
The ability of eukaryotic H/ACA snoRNPs to assemble *in vitro* in the absence of *in vivo* assembly factors contrasts with the apparent dependence on assembly factors of the C/D snoRNPs *in vitro*. Despite sharing components of the assembly machinery, the purpose for this machinery appears to differ for the two RNPs. C/D snoRNPs rely on assembly factors to properly assemble asymmetric complexes for efficient modification, while the H/ACA assembly factors function to mediate the intrinsic affinity of the core components for RNA. Additionally, in both cases assembly factors may also be involved in preventing the premature activation of the pre-RNPs to ensure site-specific nucleotide recognition.

### ***1.2.5 snoRNPs in pre-rRNA processing***

Most identified snoRNAs guide nucleotide modifications but several have also been described that play important roles in pre-rRNA processing and ribosome biogenesis. The most well-studied of these is the U3 C/D box snoRNA which is ubiquitous in eukaryotes (Marz and

Stadler 2009). U3 RNA is part of an RNP complex essential for cleavage of the pre-rRNA at positions A<sub>0</sub>, A<sub>1</sub>, and A<sub>2</sub> (following the yeast naming convention) (Kass et al. 1990; Savino and Gerbi 1990; Hughes and Ares 1991; Beltrame and Tollervey 1995; Borovjagin and Gerbi 1999; Marmier-Gourrier et al. 2011). The U3 snoRNP contains several unique features that distinguish it from other C/D snoRNP complexes. First, in place of the conventional C/D and C'/D' boxes, U3 contains a set of stem-closing C'/D and internal B/C boxes both of which form K-turns but differ in consensus sequence from conventional C/D elements (Figure 1.4) (Hughes et al. 1987; Méreau et al. 1997; Samarsky and Fournier 1998). U3 also contains the unique box elements A and A' at its 5' end which form critical base-pairs with the 18S rRNA and 5' external transcribed spacer (ETS) in the 35S pre-rRNA, generating four helices (II, III, V, and VI) (Figure 1.4 insert) (Barandun et al. 2017; Sun et al. 2017; Clerget et al. 2020). In addition to the four usual core C/D snoRNP proteins the B/C box also recruits the U3 specific protein Rrp9 (Venema et al. 2000). Finally, unlike modification guide snoRNPs, U3 does not act as an independent complex, instead associating with a large collection of U-three associated proteins (UTP) and additional protein factors to form the small subunit (SSU) processome, a large multi-complex assembly required for 18S rRNA maturation and 40S subunit formation (Dragon et al. 2002).





**Figure 1.4. Secondary structure of U3 snoRNA from *Saccharomyces cerevisiae* alone and base-pairing to the 35S pre-rRNA.** The secondary structure of the U3 snoRNA from yeast when not bound to the 18S rRNA is shown on top. Nucleotides of the conserved box elements are coloured. Depicted in the inset is the experimentally confirmed base-pairing of the 5' domain of the U3 snoRNA from yeast to the 5' ETS and 18S rRNA portion of the 35S pre-rRNA. Nucleotides belonging to the 5' ETS are blue, belonging to the 18S rRNA are green and belonging to the U3 snoRNA are black. Numbers indicate nucleotide positions in the 35S rRNA and U3 snoRNA. Line cartoons with slashes represent stretches of 35S secondary structure between rRNA-snoRNA interactions not involved in base-pairing to U3 snoRNA.

Other snoRNAs important for guiding steps in pre-rRNA processing have been discovered in various eukaryotes but are less well studied than U3. These include U14 and U17/snR30 which are essential in yeast (Li et al. 1990; Fayet-Lebaron et al. 2009) and U8 which is essential in *Xenopus* (Peculis and Steitz 1993). U14 is found throughout eukaryotes but is unique among C/D snoRNAs as in many species it both guides a 2'-O-Me modification and facilitates 18S rRNA processing (Li, et al. 1990; Liang and Fournier 1995; Dunbar and Baserga 1998). However, examples of homologs that lack either the modification (Yuan et al. 2003; Moore and Russell 2012) or processing guide regions (Yang et al. 2005) have been described. The H/ACA snoRNA U17/snR30 is found in metazoans and some other species but is not present in all eukaryotes (Atzorn et al. 2004; Vos and Kothe 2020). U17/snR30 does not guide a modification, but instead binds two regions of the 18S rRNA denoted rm1 and rm2 using the pseudouridylation pocket of its 3' stem. These interactions are important for pre-rRNA processing and deletion of snR30 from yeasts causes defects in cleavage at the A<sub>0</sub>, A<sub>1</sub>, and A<sub>2</sub> sites, but the mechanism by which snR30/U17 facilitates this cleavage remains unknown (Morrissey and Tollervey 1993; Lemay et al. 2011; Vos and Kothe 2020). U8, another C/D snoRNA found only in vertebrates, is required for correct processing of the 28S and 5.8S rRNAs (LSU rRNA species) (Peculis and Steitz 1993). Like snR30/U17, the detailed mechanism of action for U8 processing is not known, but the 5' leader sequence of U8 forms conserved base-pairs to the 3' end of the 28S rRNA (Peculis 1997; Lu et al. 2016) and loss of U8 results in incomplete processing of pre-rRNA, primarily affecting maturation in ITS2 and the 3' ETS in both humans and *Xenopus* (Peculis and Steitz 1993; Langhendries et al. 2016).

A unique case of snoRNA driven pre-rRNA processing is found in trypanosomes, a collection of protozoan parasites. The Trypanosoma LSU rRNA is fragmented into six discrete

RNA species with each fragment region initially separated by an ITS region (Cordingley and Turner 1980; Hasan et al. 1984; Hashem et al. 2013). In addition to U3, snR30, and RNase MRP homologs, knock-down experiments implicated an additional 15 snoRNAs in the processing of this particularly complex pre-rRNA primary transcript in *Trypanosoma brucei* (Gupta et al. 2010; Michaeli et al. 2012; Chikne et al. 2019). Current hypotheses propose that some of these snoRNAs regulate processing via the modifications they guide, while others may recruit as of yet unidentified proteins to facilitate cleavage events at unique sites (Rajan et al. 2019).

### ***1.2.6 Additional snoRNA functions***

As described above the majority of snoRNAs target modifications or regulate pre-rRNA processing events but there are additional snoRNAs with no known cellular targets. These RNAs are referred to as “orphans” and while they match other snoRNAs in both conserved sequence and structural features, their biological functions are poorly or completely not elucidated. In humans, target predictions estimate that between 20-50% of C/D and 15-20% of H/ACA box snoRNAs currently remain categorized as orphans, though some bioinformatic target predictions reduce the total proportion of orphans to around 16% when combining the classes (Dupuis-Sandoval et al. 2015; Jorjani et al. 2016; Falaleeva et al. 2017). This number is much lower in yeast where only three snoRNAs are classified as orphans (Sloan, et al. 2017). Nevertheless, the presence of orphan snoRNAs in the majority of examined eukaryotes suggests there may be many non-canonical snoRNA functions yet to be discovered. Indeed, a variety of functions have now been identified for snoRNAs initially classified as orphans, predominantly in humans and yeast.

Perhaps the least surprising function determined for previously classified orphan snoRNAs is the targeting of modifications outside of rRNA. Indeed, various snoRNAs have been implicated in modification of mRNA and other snoRNAs in both yeast and humans (Carlile et al. 2014;

Schwartz et al. 2014; Elliott et al. 2019). Several tRNA modifications are guided by snoRNAs in archaea, and a single example of snoRNA-mediated tRNA modification in humans was recently identified (d'Orval et al. 2001; Joardar et al. 2011; Vitali and Kiss 2019). A unique case in *T. brucei* where the spliced leader RNA is modified by an H/ACA snoRNA has also been described (Liang et al. 2002; Zamudio et al. 2009). While the targeting mechanism is apparently the same these findings are important as they reveal that snoRNAs can evolve to modify diverse RNA classes, albeit less frequently.

Several snoRNAs have been found to guide completely different types of RNA modifications altogether. Two previously orphan C/D snoRNAs in yeast, snR4 and snR45, are required for acetylation of cytosines in the SSU 18S rRNA (Sharma et al. 2017). Both RNAs associate with canonical C/D snoRNP proteins, along with the acetyltransferase enzyme Kre33. Base-pairing does not occur using the conventional D/D' guide regions, instead utilizing two alternative regions of the snoRNA to form bipartite pairing with the rRNA, resulting in a short single stranded loop in the rRNA containing the target nucleotide. The presence of a functional snR45 homolog in humans (designated U13) (Sharma et al. 2015), and homologous cytosine acetylation in plants suggests this mechanism is not limited to yeast (Sharma, et al. 2017). These currently remain the only known examples of snoRNAs guiding a modification type other than 2'-O-Me or  $\Psi$ , but also suggest it is possible that other types could exist.

In humans several snoRNAs regulate alternative pre-mRNA splicing. C/D snoRNA SNORD27 masks an alternative splice site in the transcription factor E2F7 pre-mRNA by blocking U1 snRNP binding, resulting in expression of several distinct E2F7 isoforms (Falaleeva et al. 2016). E2F7 is involved in regulating cell proliferation but the role of individual isoforms has not been determined and the importance of SNORD27 based regulation is unclear (Dimova and Dyson

2005). Base-pairing between E2F7 pre-mRNA and SNORD27 occurs outside the normal guide regions, instead using the C' and D box regions which would presumably prevent association of the canonical RNP proteins. This suggests that SNORD27 could form an alternative RNP in certain circumstances to regulate E2F7 expression. SNORD27 also has a canonical function in targeting modification of A27 in the 18S rRNA, but how involvement in the two pathways is regulated is unknown. A second C/D snoRNA, SNORD115, is also involved in alternative splicing, targeting exon V of the serotonin receptor 2C pre-mRNA (Kishore and Stamm 2006). The functional form of the mouse SNORD115 homolog is a shortened 73 nucleotide processed snoRNA (psnoRNA) which, like SNORD27, lacks the ability to associate with canonical snoRNP factors, instead interacting with heterogeneous nuclear RNPs (hnRNP) (Kishore et al. 2010). Deletion of the chromosomal region encoding SNORD115 is responsible for the human neurodegenerative disease Prader-Willi syndrome resulting from dysregulation of serotonin receptor 2C isoform expression. Additional examples of snoRNAs involved in alternative splicing have been proposed, but to date no H/ACA snoRNA has been implicated in the process.

In addition to the discovery of the shortened psnoRNAs there is growing evidence for involvement of even smaller snoRNA fragments in gene regulation. Examples of functional miRNAs in humans have been shown to be derived from precursor RNAs that contain H/ACA and C/D snoRNA-like sequence and structural features (sno-miRNAs) but have no known modification or processing guide activity (Scott et al. 2009; Ono et al. 2011). The reciprocal has also been observed in which snoRNAs with known canonical modification activity can generate snoRNA-derived small RNAs (sdRNAs) that facilitate regulation of mRNA through miRNA-like activity (Ender et al. 2008). Observation of these dual function snoRNAs is not human-specific, and they have been documented in organisms ranging from *Arabidopsis thaliana* to mouse (Taft

et al. 2009), and are even found in the parasite *Giardia lamblia* (Saraiya and Wang 2008). A human piRNA was also shown to be derived from a snoRNA-like precursor and proposed to regulate interleukin-4 mRNA levels, providing further evidence for the potential diversity of sdRNA functions (Zhong et al. 2015). Investigation of other potential cellular roles for sdRNAs is an ongoing area of inquiry as these RNAs are found to be stably expressed in more species.

Analysis of chromatin-associated RNAs responsible for generating open chromatin structure in *Drosophila* and humans has identified snoRNAs as an important fraction of this RNA pool. The effect of these snoRNAs is thought to rely on interactions with Df31, a chromatin factor responsible for generating open chromatin (Schubert et al. 2012). As is the case with psno/sdRNAs, these snoRNAs appear to associate with a novel set of proteins to perform their function. Roles for C/D snoRNPs in oxidative stress response and cholesterol trafficking have also been described in mouse and human (Michel et al. 2011; Brandis et al. 2013). Neither of these roles relies on modification activity but loss of associated RNAs was found to reduce stress response and trafficking. These findings are intriguing support for additional unique snoRNA functions but are currently only isolated cases and require further analysis regarding the specifics of their functional mechanisms and the prevalence of these types of functions. Finally, a collection of human snoRNAs were found to associate with mRNA 3' processing machinery in the absence of conventional snoRNP proteins (Huang et al. 2017). One of these, SNORD50A, was analyzed and found to have a role in modulation of alternative polyadenylation in part through competitive binding to the cleavage and polyadenylation specificity factor protein Fip1.

### ***1.2.7 ScaRNA***

SnoRNA features have also been found in a related class of ncRNAs found primarily within the Cajal bodies (CB). CBs are MLOs found within the nucleus that are rich in proteins important

for snoRNP/snRNP biogenesis and modification. Like snoRNAs, small CB-specific RNAs (scaRNAs) possess C/D or H/ACA box elements, are bound by the same core proteins and guide 2'-O-Me and  $\Psi$ , but target RNA polymerase II transcribed snRNAs of the spliceosome rather than pre-rRNA (Darzacq et al. 2002; Deryusheva and Gall 2013; Meier 2017). Unlike snoRNAs, scaRNAs can consist of tandem C/D or H/ACA sequences or hybrids of the two classes; with an H/ACA snoRNA embedded in the loop region of a box C/D RNA for example (Jády and Kiss 2001; Kiss et al. 2002). ScaRNAs also possess specialized CAB box sequence elements (H/ACA) and wobble stem (C/D) elements required for CB localization (Richard et al. 2003; Marnef et al. 2014).

Other RNP complexes not involved in RNA chemical modification are also found in Cajal bodies and can also be categorized as scaRNAs. This includes the telomerase RNA which in vertebrates contains H/ACA box elements, associates with all four core H/ACA proteins, and possesses a CAB box which targets telomerase to CBs (Mitchell et al. 1999; Pogacíc et al. 2000; Lukowiak et al. 2001; Jády et al. 2004; Collins 2006; Fu and Collins 2007). Localization to CBs was believed to be important in RNP assembly but evidence now shows that deletion of the H/ACA domain in human TR prevents trafficking to the CBs but does not disrupt telomere extension (Vogan et al. 2016). Additionally, the H/ACA RNP region may play a role in the rate of telomerase assembly and RNP stability, but the true purpose of these elements is currently poorly understood.

The discovery of new novel functions for snoRNAs and related species is likely to continue as we look deeper into the orphan snoRNAs already identified and as more is done to examine a greater collection of diverse species

### **1.3 Understanding eukaryotic RNP diversity through studying protists**

The positioning of taxa within the eukaryotic tree of life is an ever-evolving task with regular “reworking” of taxonomic groups being required. From the earliest classification of species based on morphological features, to rRNA sequence analysis, and then into the phylogenomics era, the accurate grouping of like species has proven to be complex. A relic of earlier periods of classification in which the characterization of microorganisms was significantly more restricted and based primarily on morphological features is the classification called protists, a large and diverse collection of eukaryotic organisms. Rather than pertaining to a single evolutionarily-related group, the term protist refers to any unicellular eukaryotic organism that cannot be classified as an animal, plant, or fungi (Pawlowski et al. 2012). Protists are found in all major eukaryotic taxonomic supergroups and represent the majority of eukaryotic diversity (Sibbald and Archibald 2017). Despite their prevalence, and due to the comparative ease of working with other “conventional” model organisms prior to the “-omics” era, protists have remained relatively poorly studied (del Campo et al. 2014). With the development of high throughput technologies for studying RNA, DNA and protein, analysis of a greater number of protists has allowed for significant advances in our understanding of eukaryotic molecular biology, biochemistry, and evolution. As research reveals more about these diverse lineages it has become increasingly clear that there is substantial novel variation in the structure and function of many cellular complexes, both those ubiquitous in eukaryotes and other complexes unique to individual clades. Through the increased study of protist lineages, we can further examine these evolutionary innovations granting us unique opportunities to better observe and understand the variety that exists in these complexes and how they have diversified to solve unique cellular challenges. This analysis can be a powerful companion to more traditional types of experimentation in model systems which explores how



alteration or disruption of native complexes impacts cells. These organismal comparisons can also reveal the features of RNPs that have remained unchanged, therefore defining their most “rigid”, or core features. In this PhD thesis, I explore some of the diversity of ncRNAs and ncRNPs in protists primarily by examining *Giardia lamblia*, *Giardia muris* and *Euglena gracilis*.

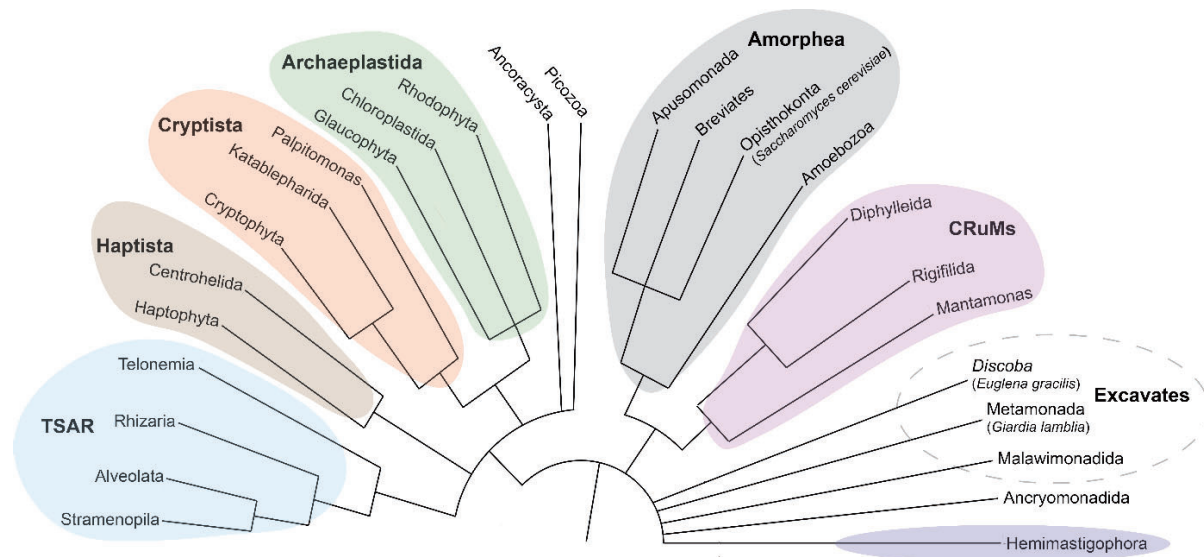
### **1.3.1 *Giardia lamblia* and the metamonads**

*Giardia lamblia* (synonym: *Giardia intestinalis*, *Giardia duodenalis*) is a protist intestinal parasite and the causative agent of Giardiasis (“Beaver fever”). *G. lamblia* is the most common cause of gastrointestinal infection worldwide, responsible for over 180 million infections each year (Torgerson et al. 2015). Initially described by Antonie Van Leeuwenhoek in 1681, *G. lamblia* was long believed to be one of the most ancient eukaryotic lineages due to its apparent lack of typically conserved eukaryotic structures such as mitochondria and other cellular features, like golgi and nucleoli (Adam 2001). *G. lamblia* is the most well-studied member of the diplomonads, a eukaryotic order made up of anaerobic and micro-aerophilic organisms possessing two nuclei (binucleated). Both nuclei are transcriptionally active and are generally thought to be genetically identical, though evidence has shown that at least some isolates of diplomonad species possess chromosomal variations between nuclei (Tůmová et al. 2007; Tůmová et al. 2016).

More recent analysis has revealed the presence of highly reduced mitochondrial-related organelles (MRO) termed mitosomes, that lack essentially all conventional mitochondrial features and maintain only the ability to perform Fe-S cluster biosynthesis (Tovar et al. 2003). Other earlier branching metamonads also contain unique MROs known as hydrogenosomes (for example, *Spironucleus spp.*) (Jerlström-Hultqvist et al. 2013) or have altogether lost their MROs leaving them as functional amitochondriates (for example, *Monocercomonoides exilis*) (Karnkowska et al. 2016). These examples show that the apparent absence of mitochondria in the metamonads is due

to secondary loss of mitochondrial components rather than an ancestral amitochondriate state.

Nucleoli have also since been discovered in *G. lamblia* further indicating that many hallmark eukaryotic features were likely present in the progenitor of the diplomonad lineage (Jiménez-García et al. 2008). There is ongoing debate concerning the placement of *Giardia* within the eukaryotic tree of life. The diplomonads belong to the phylum Metamonada which until recently was classified as part of the eukaryotic supergroup Excavata, along with the Malawimonadidae (a small poorly understood group) and the phylum Discoba, which consists of groups including Kinetoplastids, Euglenozoa and Jakobids (Adl et al. 2019). Modern phylogenomic analyses question whether the Excavates actually represent a genuine supergroup; and if in fact they are a true clade, the exact members are not clear (Figure 1.5) (Burki et al. 2020). As a result, the placement of the previous excavate groups within the eukaryotic tree remains unresolved and the classification of *G. lamblia* as an “ancient” eukaryote has been questioned. Several factors have contributed to the difficulty in placing these particular organisms including the apparent rapid evolution of their genomes, presence of derived cellular characteristics and lateral gene transfer events, all of which can lead to analysis complications including long branch attraction artifacts in phylogenetic trees.



**Figure 1.5. Phylogenetic tree representing the current proposed distribution of eukaryotic supergroups.** Defined supergroups are grouped by colour with supergroup names in bold. Relevant species from certain phyla within groups are indicated in brackets. Groups previously believed to make up the supergroup Excavata are enclosed in a dashed circle. Adapted from Burki et al. 2020 (Burki, et al. 2020).

### 1.3.2 *G. lamblia* ncRNAs

The molecular study of *G. lamblia* continues to reveal intriguing biological features resulting from its unique evolutionary history. One of the largest appeals of using *G. lamblia* as a system to study RNP complexes relates to its highly-reduced genome, a result of its evolution as a parasite. A consequence of this move towards genomic minimalization is that many cellular complexes and pathways have been streamlined, losing accessory components that are not essential to their core functions. In other cases, novel features have evolved to potentially compensate for a reduction in the number of accessory proteins and/or contribute to adaptation to its particular parasitic niche. These pressures have driven innovation resulting in *G. lamblia* possessing unique ncRNA and RNA processing features. Examples include *trans*-splicing of mRNAs whereby two independently transcribed pre-mRNAs are spliced together to form a mature

transcript (Roy et al. 2011). Additionally, a highly conserved 12 nucleotide motif is found at the 3' end of all non-tRNA/rRNA ncRNAs in *G. lamblia* and in the 5' halves of *trans*-spliced introns (Hudson et al. 2012). This motif is cleaved at a conserved position and is important for maturation of the 3' ends of these RNAs (Hudson 2014). *G. lamblia* was also the first species in which sdRNA resembling miRNAs were identified (Saraiya and Wang 2008), leading to significant subsequent searches for similar ncRNA-derived snoRNAs in other species. Analysis of *G. lamblia* sdRNA and other miRNAs found that they may be involved in regulating variable surface protein (VSP) expression in *G. lamblia* (Li, Saraiya, et al. 2011; Saraiya et al. 2011). The first structure of a Dicer protein (RNA interference pathway) was obtained using the *G. lamblia* homolog, exploiting the reduction in size and domain complexity of the protein, a feature common to many *G. lamblia* proteins (MacRae et al. 2006). Finally, the *G. lamblia* spliceosome is highly-reduced, retaining homologs of only 63 splicing proteins. *G. lamblia* snRNAs are also unusual, possessing features of both major and minor snRNAs (Hudson, et al. 2012; Hudson et al. 2015).

The potential power of comparative analysis has been expanded as genomes for additional metamonads have been completed (Andersson et al. 2007; Karnkowska et al. 2019; Xu et al. 2020). This allows for deeper exploration of the unique function and evolution of these RNPs in *G. lamblia* and related species.

### **1.3.3 *G. lamblia* snoRNAs.**

A total of 19 C/D and 12 H/ACA snoRNAs have been described in *G. lamblia* to date (Yang, et al. 2005; Luo et al. 2006; Chen et al. 2007; Hudson, et al. 2012). Predicted target nucleotides in the *G. lamblia* rRNA have been identified for 13 of the C/D and 7 of the H/ACA snoRNAs, but only a subset of the 2'-O-Me and none of the  $\Psi$  modifications have been experimentally verified. One of the C/D snoRNAs, GlsR2, is predicted to be a homolog of U14

but lacks the processing domain. Another, GlsR1, was suggested to be a U3 homolog but it lacks essentially all characterized U3 features and guides a 2'-O-Me in the SSU rRNA, a function not observed for any other known U3 snoRNA (Marz and Stadler 2009). RNAs of both classes tend to be short and often possess more degenerate box elements than in other eukaryotes (Chen, et al. 2007). Additionally, no snoRNA of either class has been shown to function as a dual guide. Most *G. lamblia* snoRNAs are found encoded between protein-coding ORFs and are expressed individually from their own promoters. However, two cases have been described in which two adjacent H/ACA snoRNAs are expressed dicistronically, with each RNA retaining a 3' end processing motif (Hudson, et al. 2012). Comparative analysis of the more degenerate *G. lamblia* snoRNAs has not been performed as no snoRNAs have yet been identified in other diplomonad species, and have only been examined in one other metamonad (*T. vaginalis*) (Chen et al. 2009; Chen et al. 2011). An RNase MRP RNA homolog was also identified in *G. lamblia* which is predicted to target cleavage at a position homologous to the A3 position in yeast.

### **1.3.4 *Euglena gracilis* and ncRNAs**

*Euglena gracilis* is a unicellular photosynthetic protist and member of the phylum Euglenozoa which, like *G. lamblia* and the metamonads, was previously included in the supergroup Excavata but now has a less certain position within the eukaryotic tree (Burki, et al. 2020). *E. gracilis* possesses chloroplasts gained through secondary endosymbiosis of a green algae and is capable of both autotrophy and heterotrophy (Gibbs 1978; Ogbonna et al. 2002). *E. gracilis* is one of the most well-studied Euglenozoans along with some kinetoplastids including *Trypanosoma brucei* and *Leishmania major*. A recently completed draft genome of *E. gracilis* determined a predicted genome size of at least 500 Mb (Ebenezer et al. 2019). This may be an underestimate as detailed analysis of the genomic structure of *E. gracilis* has been exceedingly

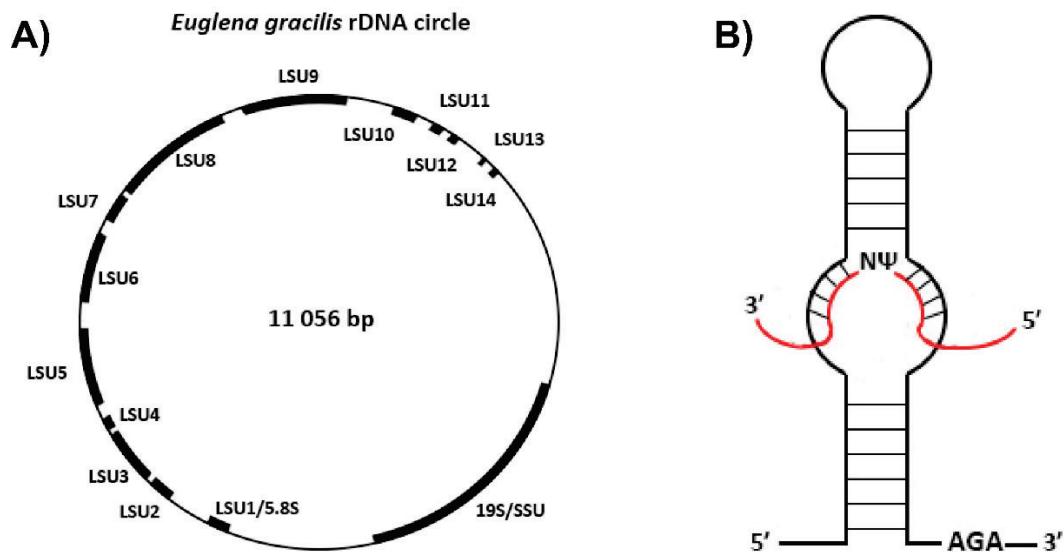
difficult due to a high degree of sequence repetition making genome assembly in many regions difficult, leaving only 20% of the genome completely assembled.

Analysis of ncRNAs and their processing in *E. gracilis* has revealed many novel features that have significantly expanded our understanding of these classes of RNA, including several unusual splicing strategies. Approximately 56% of *E. gracilis* nuclear mRNAs contain a short 5' leader sequence as their first exon (Yoshida et al. 2016). This leader sequence is added via spliced leader *trans*-splicing, in which an independent ncRNA termed the spliced leader RNA (SL RNA) containing the 5' leader exon and a short intronic sequence with a 5' splice site is spliced onto many different pre-mRNAs during maturation, in a spliceosome-dependent manner (McWatters and Russell 2017). Additionally, while *E. gracilis* has protein-coding genes containing many conventional spliceosomal introns and conventional snRNAs (Breckenridge et al. 1999; Charette and Gray 2009), it also contains unique so called “non-conventional” mRNA introns that lack clear splice site and branch point sequences (Muchhal and Schwartzbach 1994; Tessier et al. 1995). These non-conventional introns can often form stable secondary structures which bring the exon/intron junctions into close proximity (Milanowski et al. 2014). However, the only well-conserved structural feature appears to be base-pairing potential between positions +4 to +6 and -8 to -6 of the introns (positions numbered relative to splice boundaries). Non-conventional introns are predicted to be removed in a spliceosome-independent manner, but experimental validation of this hypothesis is lacking and an alternative mechanism for their removal has yet to be determined (Milanowski, et al. 2014; Gumińska et al. 2018).

### **1.3.5 *E. gracilis* snoRNAs**

One of the most striking novelties of *E. gracilis* RNA biology is its rRNA. Like the trypanosome rRNA described above, the *E. gracilis* LSU rRNA is present as multiple fragments.

However, in *E. gracilis* this has been taken to an even greater extreme where the LSU coding sequence is fragmented into 14 pieces each separated by an ITS region. Most *E. gracilis* rRNA species are encoded on an extrachromosomal plasmid (Figure 1.6A) (Schnare et al. 1990; Schnare and Gray 1990). The *E. gracilis* rRNA is also the most highly modified of any rRNA currently known. The SSU and LSU rRNA contain 88 and 262 modifications respectively; the significantly higher LSU modification density is predicted to help stabilize the assembled fragments (Schnare and Gray 2011). Of these modifications, 211 are 2'-*O*-Me and 116 are  $\Psi$ . This large number of modifications suggests many guide RNAs are likely required to guide their addition. Indeed, guide snoRNAs have previously been identified for 99 2'-*O*-Me and 12  $\Psi$  sites, most of which possess several isoforms, indicating significant redundancy for modification guides (Russell et al. 2004, 2006; Moore and Russell 2012). However, this leaves many modifications without an identified guide RNA, especially for  $\Psi$  sites. Additionally, a U14 homolog has been identified which like *G. lamblia* lacks a processing domain. A U3 snoRNA has also been characterized in *E. gracilis* indicating at least part of the SSU pre-rRNA processing pathways from other eukaryotes is retained (Greenwood et al. 1996). Similar to *T. brucei*, additional novel snoRNAs may be involved in LSU rRNA processing events in *E. gracilis*, but none have yet been determined to function in this role.



**Figure 1.6. Unique pre-rRNA processing features of *Euglena gracilis*.** (A) The 14 LSU rRNA fragments are transcribed within a single transcript from an extrachromosomal rDNA circle also containing the SSU rRNA. Relative locations of individual LSU fragments and the SSU rRNA are indicated and labeled. (B) All  $\Psi$ -guide snoRNAs detected in *E. gracilis* thus far consist of a single stem flanked by a conserved AGA sequence near the 3' end.  $\Psi$ -guide RNA is in black, target rRNA in red with the two unpaired nucleotides (N $\Psi$ ) shown.

C/D snoRNAs in *E. gracilis* are shorter than in other eukaryotes and often contain more divergent C' and D' boxes. They also act almost exclusively as single guides, each snoRNA targeting only one rRNA modification site. The 12  $\Psi$ -guide snoRNAs detected in *E. gracilis* to date differ significantly from conventional eukaryotic H/ACA snoRNAs. In place of two stem-loops and the H and ACA boxes, they instead consist of a single stem flanked on the 3' end by an AGA sequence (Figure 1.6B) (Russell, et al. 2004; Moore and Russell 2012). With so few snoRNAs identified relative to the many  $\Psi$  sites found in *Euglena* rRNA it is unclear whether this represents the predominant form of these RNAs. Noteworthy, however, is the identification of similar single-stem  $\Psi$  guide snoRNAs in the closely related kinetoplastids (Uliel et al. 2004; Liang, Hury, et al. 2007).



*E. gracilis* snoRNAs are often encoded in genomic clusters containing several consecutive snoRNAs of either C/D only, or mixed clusters with both C/D and  $\Psi$ -guide types (Moore and Russell 2012). The U3 snoRNA is a special case and is found in several genomic loci linked to either U5 snRNA or tRNA<sup>Arg</sup> genes (Charette and Gray 2009). Clusters are found as tandem repeats in the genome and in at least some cases can span tens of kilobases (Moore 2015). This repetition is a source of new C/D snoRNA evolution (Moore and Russell 2012). These snoRNAs are transcribed as polycistronic transcripts which are then processed into mature RNAs by an unknown mechanism. The presence of far more verified rRNA modifications than snoRNAs suggests there are likely many more that remain to be discovered and with the degree of redundancy present in the snoRNA isoforms that have been characterized, it is also plausible that unique snoRNA functions may also await discovery in *E. gracilis*, similar to what has been observed in other eukaryotic species.

### **1.3.6 Genomic expansion and reduction**

*G. lamblia* and the other diplomonads contrasted with *E. gracilis* represent relative extremes for genome size in unicellular eukaryotes. Both dramatic expansion and reduction can lead to significant evolutionary innovation, via necessity generated through loss or from opportunity created through excess. A thorough understanding of eukaryotic RNP structure-function relationships and discovery of novel features can come from broadening the scope of organisms from which they are studied. These protists species and others like them therefore present a valuable opportunity to gain insight into this diversity.

## **1.4 Objectives**

The majority of research on ncRNPs to date has focused on a relatively small number of model organisms; while protists, which make up the bulk of eukaryotic evolutionary diversity,

remain largely unexplored. Open questions remain about the degree to which our current knowledge of these cellular complexes is indicative of eukaryotes as a whole or whether substantial variation exists in both form and function of these ubiquitous RNPs in different eukaryotic lineages. To this end, the encompassing goal of my research project was to examine how eukaryotic ncRNAs, with a primary focus on snoRNAs, and their associated RNPs have evolved in cases of significant genomic expansion and reduction by studying members of the diplomonads and the Euglenozoan *E. gracilis*.

At the onset of my research there had been little success with the *in vitro* assembly of eukaryotic C/D snoRNP complexes which restricted their more detailed study. The first objective of my thesis work was to utilize bioinformatic and experimental techniques to characterize the components and features of *G. lamblia* C/D snoRNPs and determine how they may differ from the more widely-studied snoRNPs of higher eukaryotes (Chapter 3). Earlier studies in the Russell lab had discovered unusual snRNAs in *G. lamblia* and our later work found further snRNA structural divergence in several *Spironucleus* species (Hudson et al. 2019). Thus, a second focus of my research was to examine different ncRNA classes in another *Giardia* species, *Giardia muris*; to investigate the extent of ncRNA variation within the genus and to determine if the *G. lamblia* 3' end processing motif is conserved across species (Chapter 2). Finally, to explore the impact of significant genomic expansion and sequence repetition on snoRNA evolution and function, I used a novel RNA-seq method developed in our lab to explore snoRNA function, structure, and evolution in *E. gracilis* (Chapter 4).

## **Chapter 2: Analysis of *Giardia muris* ncRNAs reveals highly divergent spliceosomal, telomerase, and U3 RNAs and uncovers an RNase P-snoRNA diRNP**

### **2.1 Introduction**

Small non-coding RNAs (ncRNAs) have a vast array of cellular roles across the domains of life. These include maintenance of chromosome ends by telomerase RNA (TER) (Musgrove, et al. 2018), post-transcriptional RNA modification targeted by small nucleolar RNAs (snoRNAs) (Henras et al. 2017), tRNA processing by RNase P RNA (Jarrous 2017), and RNA splicing mediated by the spliceosomal small nuclear RNAs (snRNAs) (Zhang, et al. 2019). Advances in high throughput -omics technologies have allowed for the identification of RNAs in a broader sampling of organismal diversity. This has generated new opportunities to perform comparative genomics and RNA-omic analysis, thus resulting in a better understanding of the diverse cellular roles and mechanisms of action of ncRNAs in different species, along with their evolution.

Protists represent the largest proportion of eukaryotic evolutionary diversity but due to factors including difficult cultivation and/or lack of detailed genomic sequence information, many of these species remain poorly characterized. Metamonada is a eukaryotic phylum composed of anaerobic and microaerophilic protists (metamonads) exhibiting high rates of genomic sequence evolution and possessing unique mitochondrial-related organelles (MROs) in place of traditional mitochondria (Burki, et al. 2020). The metamonads are composed of three major groups: Preaxostyla, Parabasalia, and Fornicata which contain both free-living and parasitic species with significant differences in genome size and complexity. The fornicate group Diplomonadida contains species responsible for disease in animals including the most well studied metamonad, *Giardia lamblia*, a common human enteric pathogen responsible for an estimated 300 million cases of diarrheal disease worldwide each year (Einarsson et al. 2016).

Parasitism is commonly associated with significant changes to genome structure, including the streamlining of cellular processes through genomic reduction and the evolution of novel mechanisms to facilitate the adaptation to a new lifestyle (Jackson et al. 2016; Xu, et al. 2020). This parasite-driven minimalization is apparent in the *G. lamblia* genome where studies have found many examples of significant reduction in the number of components in metabolic pathways, the kinome, transcription initiation, and RNA processing machinery (Best et al. 2004; Morrison et al. 2007; Manning et al. 2011; Xu, et al. 2020). This streamlining is also seen in ribonucleoprotein (RNP) complexes such as the spliceosome where only 62 of 167 human spliceosomal proteins are identifiable in *G. lamblia*, compared to 115 in the oxymonad *Monocercomonoides exilis* (Hudson, et al. 2019).

The study of evolutionarily conserved eukaryotic cellular complexes in species undergoing genomic minimalization grants the unique opportunity to identify the most rigid or core features of these streamlined complexes, that is, those that have been retained and are essential to perform their function. These features may also be complemented with novel “innovations” that have evolved in genus like *Giardia* during the transition to its unique parasitic lifestyle. *G. lamblia* in particular has been the source of a number of important discoveries regarding ncRNAs; including solving the first structures of the miRNA processing protein DICER (MacRae, et al. 2006) and the discovery of snoRNA-derived miRNAs (sdRNAs) (Saraiya and Wang 2008). Analysis of ncRNAs in *G. lamblia* has also revealed novel features such as structurally-divergent snRNAs and a unique general ncRNA processing pathway in this organism (Hudson, et al. 2012).

Comparative genomics using a variety of *G. lamblia* isolates has led to valuable new insights into intraspecies variation (Morrison, et al. 2007; Franzén et al. 2009; Jerlström-Hultqvist et al. 2010; Adam et al. 2013; Ankarklev et al. 2015). The genomes of a number of other

metamonads have also been sequenced including the oxymonad *Monocercomonoides exilis* (Karnkowska, et al. 2019), the parabasalid *Trichomonas vaginalis* (Carlton et al. 2007), and the diplomonad *Spironucleus salmonicida* (Andersson, et al. 2007). However, most analyses performed so far have focused on the protein-coding repertoire of these genomes with the only detailed description of different ncRNA classes reported for *G. lamblia*. This lack of information has restricted any comparative analysis of ncRNAs from other metamonad species, including the diplomonad *Giardia muris*, an intestinal parasite found infecting a variety of rodent species.

Identification of ncRNAs in non-model organisms can be challenging as many ncRNA classes only conserve short sequence elements while their overall structure and size can vary significantly. As a result, RNA expression data is often crucial for accurate ncRNA detection and annotation. The lack of ncRNA data from *G. muris* has in part been due to the inability to cultivate it axenically, preventing ready access to sufficient cellular material for the characterization of relatively non-abundant ncRNAs using lower throughput techniques (Cacciò et al. 2018; Dann et al. 2018). The recent assembly of the *G. muris* genome (Xu, et al. 2020) provides a first step towards ncRNA prediction in this species, but additional experimental data is required for robust identification. The *G. muris* genome is even smaller than *G. lamblia* (9.8 Mbp vs 12.8 Mbp) containing shorter intergenic regions, with mean/median intergenic lengths of 264 bp/37 bp between protein-coding sequences in *G. muris*, among the smallest identified in any eukaryote (Xu, et al. 2020). The protein-coding gene density leaves little space for ncRNAs whose mature sizes typically range from approximately 20-200 bp in *G. lamblia*. ncRNA prediction is further complicated by a lack of global synteny in the two *Giardia* species due to many small-scale genomic rearrangements. The two genomes do however retain stretches of local synteny.

Our previous work uncovered a 12 nt sequence involved in the maturation pathway of all previously identified ncRNAs in *G. lamblia*, excluding ribosomal RNA (rRNA) and transfer RNA (tRNA) species (Hudson, et al. 2012). This sequence motif is also found at the 3' extremity of 5' *trans*-spliced intron halves, indicating a broader role in RNA maturation. The same motif was identified at corresponding positions in the *G. muris trans*-spliced introns (Xu, et al. 2020), but the presence and role of the 3' processing motif in RNA maturation in this species remained unexplored as very few *G. muris* ncRNAs have been identified. In this study, we have analyzed the genome of the diplomonad *Giardia muris* and generated a small ncRNA library to investigate the complement of ncRNAs in this species and compared these findings to ncRNA data from *G. lamblia* and other metamonads.

Our analysis revealed a compact collection of snoRNAs seemingly conserved between reduced genomes with different evolutionary histories, corrects the previously misidentified *Giardia* U3 snoRNA leading to new evolutionary insights about metamonad pre-rRNA processing machinery, and identified an RNase P–snoRNA fused diRNP capable of targeting tRNA<sup>Met</sup> for methylation at a nucleotide conserved throughout the three domains of life. This work highlights the importance of comparative genomic and RNA-omic analysis in diverse eukaryotes and how it can reveal important information about both highly conserved ncRNAs and novel innovations in ncRNA structure and function.

## **2.2 Results**

### ***2.2.1 G. muris ncRNA candidate prediction***

The conserved 3' processing motif sequence at the end of *G. lamblia* ncRNAs provides a valuable tool for *de novo* identification of ncRNAs. The discovery of a similar motif residing in the *trans*-spliced introns of *G. muris* suggested that this may be a conserved mechanism for ncRNA

3' end processing in additional *Giardia* species. We could therefore employ this predicted feature as a tool to search the genome and generate a set of ncRNA candidate sequences to be further analyzed for features of distinct RNA classes. To explore this possibility, we first performed BLASTN searches to the *G. muris* genome using 40 *G. lamblia* ncRNAs. This uncovered candidate homologs for 10 *G. lamblia* ncRNAs in *G. muris* which we then aligned to inspect their 3' ends for sequences resembling the processing motif. This revealed strong conservation of a 5'-CCTTYNHTNAA-3' (Y = pyrimidine, H = A, C or T) motif at the 3' end of predicted *G. muris* RNA encoding regions, a near identical match to the *G. lamblia* 5'-CCTTYNHHTHAA-3' motif consensus sequence. We next searched the *G. muris* genome for matches to this motif and collected matching sequence hits along with approximately 100 nucleotides (nt) upstream of the motif or enough upstream sequence to reach an AT rich putative promoter. We also searched for potential ncRNA candidates that may lack the processing motif but have conserved features diagnostic of various ncRNA classes using additional software (see Materials and Methods). In total, these combined searches generated 219 ncRNA candidates. Three of these were the previously identified *trans*-spliced introns, leaving 216 total ncRNA candidates which were used as the starting point for more detailed analyses.

To examine the cellular expression of these ncRNA candidates, we generated a stranded small ncRNA library from mRNA depleted *G. muris* RNA. Raw 50 nucleotide Illumina reads were mapped to the *G. muris* genome, then assigned to genomic features to quantify RNA abundance. Reads tended to be 3' biased, likely due to the short read length and significant secondary structure of many ncRNAs. We therefore used read count as the metric for relative abundance rather than FPKM, RPKM, or TPM. Of the 21,915,561 reads mapping to the *G. muris* genome: 11,355,012 were assigned to rRNA; 2,359,788 to tRNA; and 950,697 to our 216 new ncRNA candidates. The

majority of the new ncRNA reads mapped to 104 of the candidates, presumably the most abundant RNA species.

### ***2.2.2 Identification of C/D and H/ACA box snoRNAs in G. muris***

Small nucleolar RNAs (snoRNAs) are split into two classes, C/D and H/ACA, both of which act as guides for RNP complexes targeting various RNA species for nucleotide modification via base-pairing interactions (See Chapter 1 for a detailed description). Previous research on small ncRNAs in *G. lamblia* identified a total of 31 snoRNAs: 19 Box C/D and 12 box H/ACA (Yang, et al. 2005; Luo, et al. 2006; Chen, et al. 2007; Hudson, et al. 2012). Of these, 13 C/D and 7 H/ACA snoRNAs have convincing predicted interactions to target rRNA modification sites. The remaining snoRNAs are classified as “orphan”, with no currently identified modification targets in other RNA species. Using a combination of a covariance model (CM), Patscan, and BLASTN searches along with the Snoscan webserver, we examined our ncRNA candidates and searched the *G. muris* genome for homologs of *G. lamblia* snoRNAs. These searches identified homologs of 14 C/D snoRNAs (including a U3 snoRNA discussed below) and 7 H/ACA snoRNAs in *G. muris* (Table 2.1 and Figure A.1.1). We refer to these RNAs as GmsR# for *Giardia muris* small RNA, following the naming convention established in *G. lamblia*, with homologs numbered accordingly (i.e. the GlsR1 homolog is named GmsR1).



**Table 2.1. List of C/D and H/ACA box snoRNAs identified in *G. muris* and their homologs in other species.**

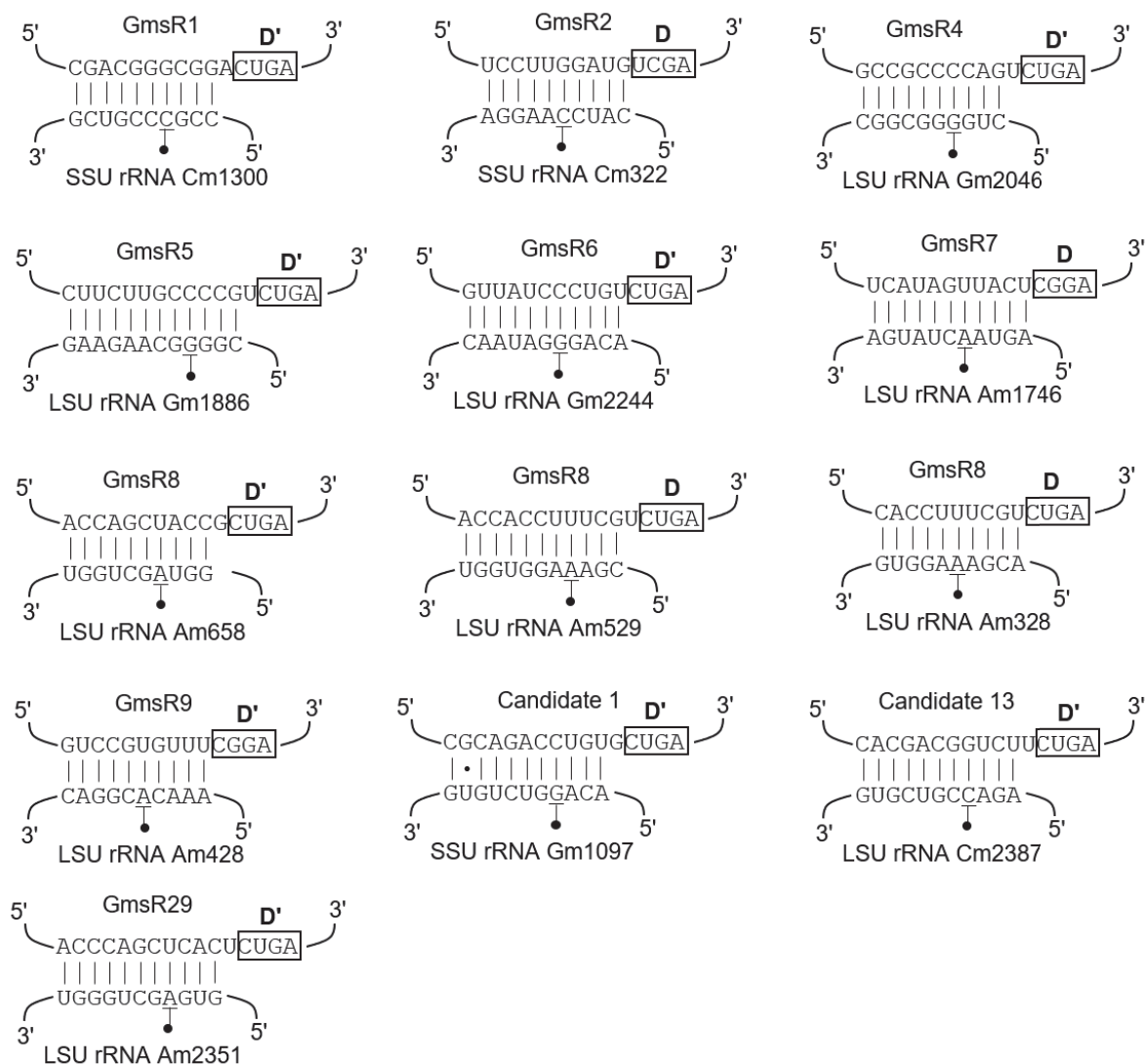
C/D snoRNAs									
<i>G. muris</i> RNA	Predicted target nucleotide	Guide box	Number of reads	<i>G. lamblia</i> homolog	Conserved target position	Found in syntenic position	<i>G. theta</i> nucleomorph targeting homologous position	Human/Yeast snoRNA targeting homologous position	Pyrobaculum sRNA targeting homologous position
GmsR1	18S rRNA C1300	D' box	61152/61087	GlsR1	Yes	Yes	-	SNORD43/snr70	-
GmsR2	18S rRNA C322	D box	5741	GlsR2	Yes	Yes	GINM-R9	SNORD14/U14	-
GmsR4	28S rRNA G2046	D' box	26578	GlsR4	Yes	Yes	-	SNORD31/snr67	-
GmsR5	28S rRNA G1886 tRNA <sup>Ser</sup> U28	D' box	65646	GlsR5	Yes	No	-	-	-
		D box			No		-	SNORD12/snr190	-
GmsR6	28S rRNA G2244	D' box	13534	GlsR6	Yes	Yes	GINM-R10	SNORD1A/snr38	-
GmsR7	28S rRNA A1746	D' box	54452	GlsR7	Yes	Yes	GINM-R8	SNORD46/snr63	sR063D'
GmsR8	28S rRNA A529	D box	25031	GlsR8	Yes	No	GINM-R13	SNORD51/snr59	-
	28S rRNA A658	D' box			Yes				-
	28SrRNA A328	D box			Yes				-
GmsR9	28S rRNA A428	D' box	83495	GlsR9	Yes	Yes	-	-	-
GmsR13	U4 snRNA C63	D' box	15910	GlsR13	No	No	-	-	-
GmsR14	28S rRNA U1112 <sup>a</sup>	D' box	7954	GlsR14	No	Yes	-	-	-
GmsR15	tRNA <sup>Met</sup> C34	D box	25538	GlsR15	Yes	No	-	SNORD97(SCARNA97)/-	-
Candidate 1	18S rRNA C1097	D' box	72954	Candidate 1	Yes	Yes	GINM-R12	SNORD25/snr56	sR032D
Candidate 13	28S rRNA C2387	D' box	12144	Candidate 13	Yes	Yes	GINM-R7	SNORD35/snr73	sR030D'
GmsR29	28S A2374	D' box	78672	-	-	-	GINM-R11	SNORD29/snr71	sR115D

H/ACA snoRNAs						
G. muris RNA	Predicted target nucleotide	Pseudouridylation on pocket	Number of reads	G. lamblia homolog	Conserved target position	Found in syntenic position
GmsR17	-	-	709	Yes	-	Yes
GmsR18	28S rRNA U2352	2	2538	Yes	Yes	Yes
GmsR19	28S rRNA U1805	2	1875	Yes	Yes	Yes
GmsR20	-	-	927	Yes	No	Yes
GmsR22	-	-	6725	Yes	-	Yes
GmsR27	28S rRNA U1749	2	236	Yes	Yes	Yes
Candidate 16	28S rRNA U2309	2	2121	Yes	Yes	No
GmsR30	28S rRNA U2675	2	824	-	-	-

-: indicates N/A

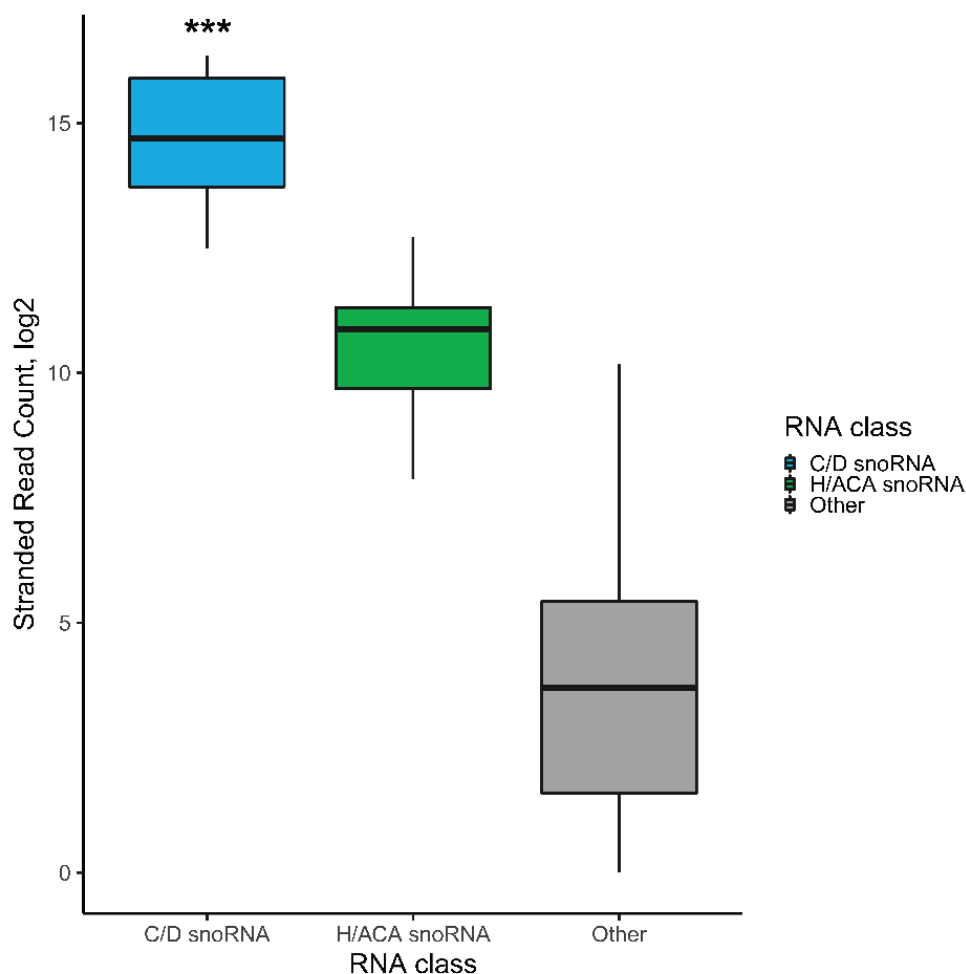
a: indicates that targeting this nucleotide requires a degenerate box element.

Pseudouridylation pocket: indicates which of the two stems contains the guide pocket



**Figure 2.1. *G. muris* C/D snoRNAs predominantly target rRNA positions also targeted in *G. lamblia*.** Predicted base-pairing for *G. muris* C/D snoRNA antisense elements (ASE) (top) to rRNA target sequences (bottom). All three putative GmsR8 targets are shown. The D/D' box element downstream of the guide for each snoRNA is boxed and labeled based on which ASE functions as the guide for that interaction. The nucleotide targeted for 2'-O-methylation in the rRNA is underlined and indicated by a line connected to a •. Non-Watson-Crick GU pairs are indicated as •. The target rRNA subunit and nucleotide position modified are labeled underneath the respective base-pairing figure. A minimum of 10 base consecutive pairs was required, and extended base-pairing is shown when predicted to occur.

We found *G. muris* homologs for 11/13 *G. lamblia* C/D modification guide snoRNAs based on sequence similarity and conserved rRNA targets (Table 2.1 and Figure 2.1). Intriguingly these snoRNAs form a distinct grouping among our ncRNAs based on read abundances in our small RNA library (Figure 2.2). These are the most abundant RNAs, making up just over 65% of all ncRNA reads (Supplemental file 1). Using the Snoscan webserver (Lowe and Eddy 1999) we searched for unique *G. muris* C/D snoRNA modification target sites. We included *G. muris* C/D snoRNAs homologous to *G. lamblia* snoRNAs with known target nucleotides in these searches as some snoRNAs are able to act as dual guides, using both their D and D' box anti-sense elements (ASE) to guide modification. This approach was successful in identifying novel interactions for two C/D snoRNAs to target sites apparently not being targeted for modification in *G. lamblia*. The first is an ASE guide upstream of the D' box in GmsR13 that is predicted to target position C63 of our predicted *G. muris* U4 snRNA homolog (Figure A.1.2B). GlsR13 and GmsR13 maintain a high level of overall sequence similarity with the notable exception of the guide sequences upstream of the D and D' boxes (Figure A.1.1). This is in stark contrast to nearly all other box C/D RNA homologs in the two *Giardia* species which maintain perfect or near perfect conservation in at least one of the two ASE guide regions while the rest of the RNA sequence is divergent (Figure A.1.1). These changes are significant enough that GlsR13 in *G. lamblia* is not predicted to target U4 for modification. Box C/D snoRNA/scaRNA-guided modification of snRNAs is seen in many eukaryotic lineages including vertebrates and fungi, but no snRNA targeted modifications have been previously determined in metamonads.

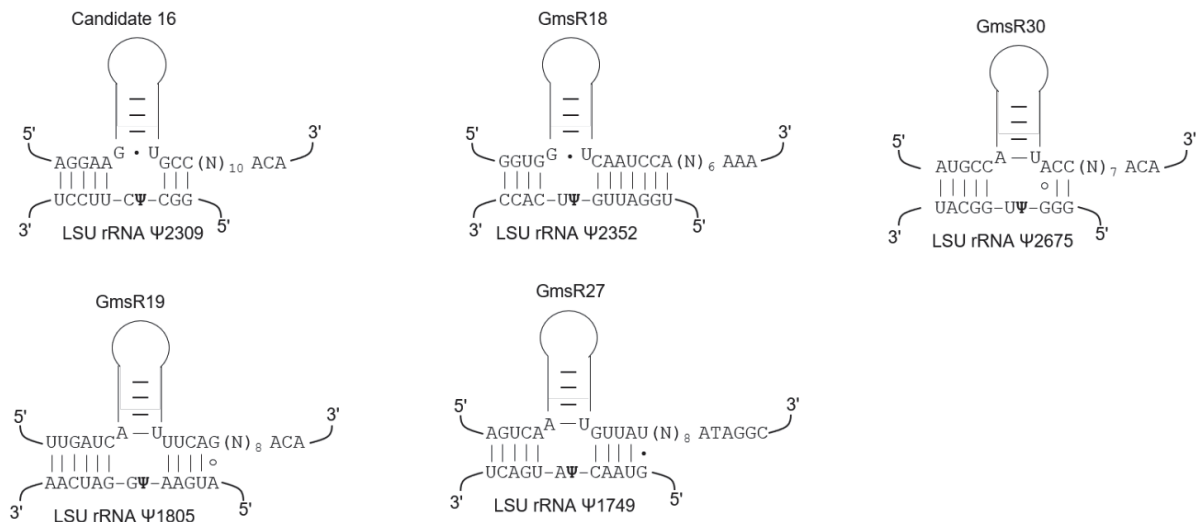


**Figure 2.2. *G. muris* snoRNAs form distinct abundance clusters among ncRNA candidates.** Our predicted C/D and H/ACA snoRNAs form clusters separate from all other ncRNAs in the RNA-seq data based on read counts. C/D snoRNAs are the most abundant ncRNA class, followed by H/ACA RNAs. ‘Other’ refers to all remaining RNA candidates from our 216 predicted candidates, excluding the 63 repeated rDNA operon RNAs. Colours correspond to colours associated with RNA classes in Supplemental file 1, C/D snoRNAs in blue, H/ACA snoRNAs in green, and Other RNAs in gray. \*\*\*  $p < 1 \times 10^{-14}$  One-way ANOVA, Tukey Post-hoc test for C/D snoRNAs against H/ACA snoRNAs

The *G. lamblia* GlsR8 C/D snoRNA has a known methylation target at position A507 in the 28S rRNA. We found that GmsR8 is also predicted to target a homologous position in *G. muris*, supported by a compensatory change in the snoRNA-rRNA interaction to maintain base-pairing (Figure 2.1). However, we also detected an additional two targets for GmsR8 in the 28S rRNA at nearby rRNA positions. One of the new targets is guided by the same D box ASE as the A507

modification, the other by the D' box ASE. All three of these guide-target interactions are predicted to be at least 10 contiguous Watson-Crick base-pairs in length (Figure 2.1). Upon re-examination of the *G. lamblia* GlsR8 snoRNA we found that the two additional snoRNA-rRNA interactions are conserved in this species (Figure A.1.2C). If these additional targets are in fact all modified, then GlsR8 and GmsR8 would be the first identified multi-site guide snoRNAs in *Giardia* and first C/D snoRNAs to guide multiple sites for modification using the same guide region to be identified in the metamonad lineage.

Our searches in *G. muris* also identified homologs of seven *G. lamblia* H/ACA snoRNAs. Of these, four have predicted pseudouridylation target sites in *G. muris*, all in the LSU rRNA (Figure 2.3). As with the C/D snoRNAs, we found that all our predicted H/ACA snoRNAs had a comparable number of reads in our small RNA library thus forming a group separate from most other RNA classes based on read count (Figure 2.2. And Supplemental file 1). GmsR17 and GmsR22 snoRNAs were both identified via sequence conservation with their *G. lamblia* homologs and syntenic positioning in the two species' genomes but searches for targets in *G. lamblia* and *G. muris* using both snoGPS software (Schattner et al. 2004) and manual inspection approaches have not as yet revealed any plausible targets.



**Figure 2.3. Newly identified H/ACA snoRNAs from *G. muris*.** Predicted bipartite base-pairing for *G. muris* H/ACA snoRNAs to rRNA target segments. Each H/ACA snoRNA (top) is labeled and bipartite pairing to the rRNA is depicted with the predicted 5' half of the pseudouridylation pocket on the left and 3' half on the right. The first nucleotide pair of the upper H/ACA snoRNA stem is shown with remainder of the upper stem indicated as a cartoon stem loop. The distance in nucleotides to the ACA or H box elements following the second half of the bipartite pairing is indicated as (N)<sub>x</sub>. The nucleotide targeted for pseudouridylation is indicated as a Ψ. Non-Watson-Crick base-pairs are indicated as • for GU or ° for GA.

We detected two snoRNAs, one C/D (GmsR29) and one H/ACA (GmsR30) unique to *G. muris* (Table 2.1). Based on read count, GmsR29 groups with the other C/D snoRNAs in our small RNA library data and is the second most abundant ncRNA detected. The D, D', and C' boxes are perfect matches to the conventional snoRNA consensus sequences and the C box differs by only a single nucleotide (AUGACGA). We identified a perfect 11 base-pair interaction between the D' ASE of GmsR29 and the *G. muris* 28S rRNA homolog targeting nucleotide position A2351 of the peptidyl transferase center (PTC) (Figure 2.1). This position is homologous to the nucleotide targeted by C/D snoRNAs found in humans, yeast, and some archaeal species (Table 2.1). GmsR30 contains the conventional H/ACA sequence and structural features (Figure A.1.3), and manual inspection detected a possible target for the pseudouridylation pocket of stem 2 in the 28S rRNA

(Figure 2.3). However, one side of the bipartite pairing consists of only three pairs including a non-Watson-Crick GA pair making it unclear if the interaction would be stable enough to allow for modification.

15 of the 16 most abundant RNAs in our sequencing data are C/D snoRNAs, with the only exception being a single RNA of unknown function (nc151), which contains some C/D snoRNA-like features and may represent a more evolutionarily divergent RNA of this class. Of the next 14 most abundant RNAs in the library, 8 are H/ACA snoRNAs creating a second snoRNA-rich read cluster. Read count between RNAs can vary significantly, but the close clustering (by abundance) of RNAs of the same class further supports our classification of these ncRNAs.

### ***2.2.3 C/D snoRNA-guided modification sites are conserved between *G. muris* and the *Gu. theta* nucleomorph rRNAs***

The *G. muris* genome is over 20% smaller than the *G. lamblia* genome yet has conserved nearly all C/D snoRNAs with known rRNA targets. This led us to pose the question: are a similar set of nucleotide positions also being targeted for modification in other distantly related eukaryotic species possessing reduced genomes. The cryptomonad *Guillardia theta* is a eukaryotic species that acquired a photosynthetic plastid through secondary endosymbiosis (i.e. through endosymbiotic uptake of a eukaryotic algae rather than a cyanobacteria). Even more unusual is the fact that in this species the plastid still retains a highly reduced eukaryotic nuclear genome termed a nucleomorph (NM), that is the remnant of the algal endosymbiont genome (Zimorski et al. 2014). The *Gu. theta* nucleomorph genome is derived from a red algae but has been reduced to just 551 kbp in size (Archibald 2007). A recent study on sRNAs in the *Gu. theta* NM uncovered a collection of C/D snoRNAs (designated GtNM-R#) including two that are either homologs or functional analogs of *G. lamblia* GlsR6 (GtNM-R10) and GlsR7 (GtNM-R8) based on CM searches (Åsman

et al. 2019) suggesting the existence of a selective pressure to maintain snoRNA guided modification at these positions. Comparison of small RNA complements in *Giardia* species with the nucleomorph of *Gu. theta* presents a valuable opportunity to study functional commonalities of highly reduced but evolutionarily unrelated genomes that were subject to different selective pressures towards size compaction. We used our new data on *G. muris* C/D snoRNAs to compare predicted snoRNA guided nucleotide modifications in the *Gu. theta* nucleomorph and *G. muris*. Aligning the rRNA from the two species revealed that 7 of the 11 *Gu. theta* NM modified rRNA nucleotides targeted by C/D snoRNAs are also targeted for modification in *G. muris* (Table 2.1 and Figure A.1.4). Alignment of the snoRNAs targeting homologous positions in their respective rRNAs shows little conservation outside of the ASE guide regions and their associated C/D boxes (Figure A.1.5) preventing a definitive conclusion as to whether these snoRNAs share common ancestry. However, all seven of these rRNA positions are modified in both yeast and humans, and four of the sites are targeted by C/D sRNAs in a variety of archaeal lineages (Table 2.1) (Dennis et al. 2015; Lui et al. 2018). These four modified positions are among the sites targeted by snoRNAs which were previously predicted to trace back to the last eukaryotic common ancestor (LECA) (Hoeppner and Poole 2012). Conservation of these modifications over such large evolutionary distances and among species that otherwise maintain only a small collection of snoRNAs suggests significant functional importance and selective pressure to preserve these particular rRNA modifications and the snoRNA-mediated mechanism of targeting them.

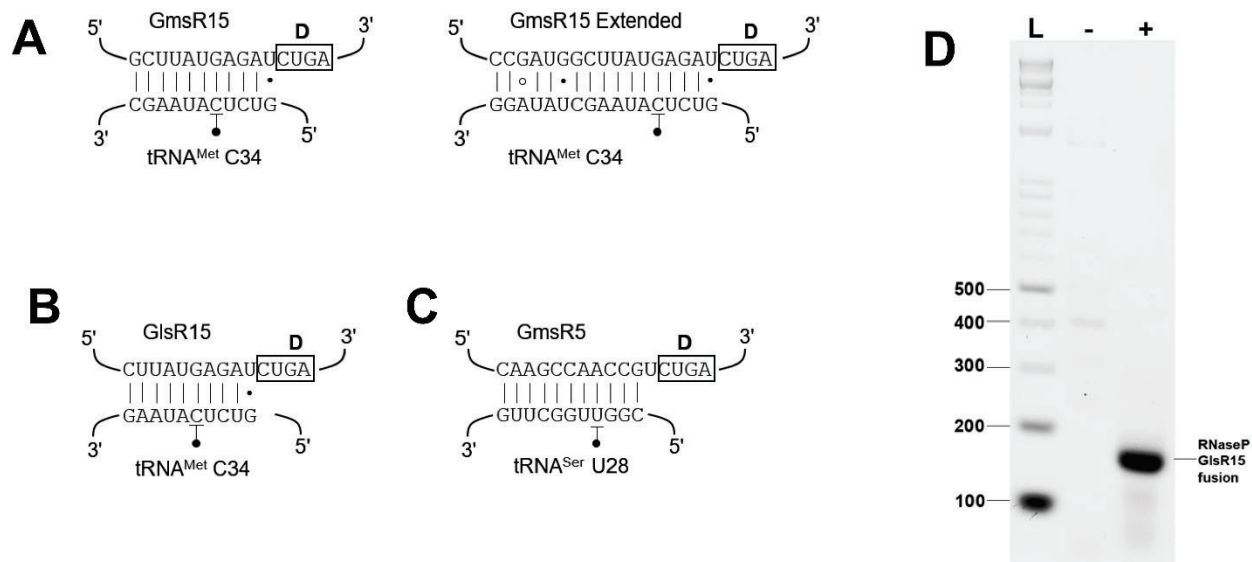
Mapping the *G. muris* and *Gu. theta* NM rRNA modification sites onto the human rRNA structure shows that many occur in functionally important regions of the rRNA. The majority of the conserved modification positions are found in regions homologous to the peptidyl transferase center (PTC) and helix 69 (H69) of the B2a inter-subunit bridge in the 28S rRNA as well as helix



34 (H34) of the decoding site in the 18S rRNA (Figure A.1.6). Each of these snoRNA-guided modifications has been found to play an important role in translation fidelity in yeast (also see Discussion).

#### **2.2.4 An RNase P-snoRNA diRNP targets *tRNA*<sup>Met</sup> 2'-O-methylation in *Giardia* species**

Among the C/D snoRNAs that are conserved between *G. muris* and *G. lamblia* are several that do not appear to have rRNA targets. Two of these, GmsR14/GlsR14 and GmsR15/GlsR15, maintain near perfect sequence conservation in the 10 nucleotides upstream of their D box elements for the homologs between species (Figure A.1.1). This conservation suggests they are functional guide elements and share a common RNA modification target in the two species, which prompted us to search for potential targets in other classes of ncRNA. We did not detect any plausible targets for GmsR14/GlsR14 in currently known ncRNAs; remarkably, both the GmsR15 and GlsR15 snoRNAs can base-pair extensively with *tRNA*<sup>Met</sup> to target position C34, located in the anticodon loop for modification. In both cases, the snoRNA-tRNA interaction consists of at least 10 contiguous base-pairs, with the GmsR15-*tRNA*<sup>Met</sup> pairing able to extend to 17 consecutive pairs (Figure 2.4A and B).



**Figure 2.4. An RNase P-snoRNA fusion RNA and GmsR5 have predicted tRNA targets for 2'-O-methylation.** Predicated base-pairing interactions between *G. muris* C/D snoRNAs (**A** and **C**) or *G. lamblia* GlsR15 snoRNA (**B**) and predicted tRNA targets. For the GmsR15-tRNA<sup>Met</sup> interaction, the minimal 10 base-pairing (left) and possible extended base-pairing potential (right) are both shown. All three predicted snoRNA-tRNA interactions use the anti-sense guide region upstream of the D box (boxed and labeled). The nucleotide targeted for 2'-O-methylation is underlined and indicated by a line connected to a •. Non-Watson-Crick base-pairs are indicated as • for GU or ° for GA. (**D**) The 3' RACE product mapping the 3' end of the *G. lamblia* RNase P transcript is resolved on a 3% agarose gel. The fused RNase P-GlsR15 transcript is indicated. Lanes with products from the reaction containing reverse transcriptase (RT) (+) or lacking RT as a control (-) are indicated above the lane. The included DNA size ladder lane (L) has bands labeled in base-pairs.

Previous studies in *G. lamblia* found that the proposed 5' end of the GlsR15 overlaps the 3' end of the RNase P RNA by approximately 23 nucleotides (Chen, et al. 2007). The two genes were found to be transcribed in at least some cases initially as a single fusion transcript but were predicted to then be internally cleaved to form two distinct mature RNAs, however this was never experimentally examined. Due to the nature of the overlap, such a processing event would likely render one of the two RNAs non-functional. Additionally, examination of all the currently identified *G. lamblia* ncRNAs reveals that the RNase P RNA is the only known ncRNA from *G. lamblia* that lacks the conserved 3' end processing motif in the region of its previously predicted

mature 3' end. To further explore this, we performed 3' rapid amplification of cDNA ends (RACE) to determine if the mature *G. lamblia* RNase P transcripts retain the GlsR15 sequence (at the 3' end) or are instead cleaved internally utilizing a site-determination mechanism independent of the normal ncRNA 3' end processing motif, as previously proposed (Chen, et al. 2007). RACE experiments detected a single prominent band at the predicted size for the RNase P-GlsR15 “fusion” transcript, which was then verified by DNA sequencing of the RACE product (Figure 2.4D and Figure A.1.7A). This result indicates that the vast majority (if not all) of the transcripts from this genomic region remain as an RNase P–GlsR15 fusion without internal processing, instead relying solely on the processing motif located at the 3' end of GlsR15 for 3' end maturation. Additional work from our lab found that co-precipitation of RNA and proteins in complex with *G. lamblia* tagged homologs of the C/D snoRNA binding protein Snul3p significantly enriches the RNase P RNA and precipitated protein components of the RNase P RNP (see Chapter 3 for details). Together these data strongly support formation of a hetero-diRNP containing both an RNase P and snoRNA domain assembled on the ‘fused’ RNA. Analysis of the genomic region surrounding the newly identified RNase P RNA sequence in *G. muris* revealed a similar dicistronic fused organization for the RNase P–GmsR15 genes in which only GmsR15 possesses the conserved 3' processing motif (Figure A.1.7B). This shows that the diRNP structure is conserved in these *Giardia* species. We examined the sequences downstream of RNase P RNA genes in other metamonad species to look for potential snoRNA features or antisense elements for pairing to tRNA<sup>Met</sup> but found no evidence of either, in any genus outside of *Giardia*. The RNase P–snoRNA fusion is therefore likely *Giardia* specific.

During our analysis of *G. muris* snoRNAs we also found that GmsR5, which targets G1886 of the LSU rRNA for methylation using its D' box guide, can pair with tRNA<sup>Ser</sup> using its D box

ASE to target position U28, with a base-pairing length similar to most ASE interactions of C/D snoRNAs in *Giardia* (Figure 2.4C). The GmsR5 D box ASE is nearly 100% conserved in GlsR5, again suggesting a common RNA target in the two species. However, unlike GmsR15/GlsR15, this predicted tRNA interaction is not conserved in *G. lamblia* due to changes in the G1 tRNA<sup>Ser</sup> sequence preventing proper snoRNA-tRNA pairing. Additionally, no tRNA in *Giardia* or other species has yet been experimentally validated to possess a 2'-O-methylation at position 28. It will require further biochemical investigation to determine the modification status of U28 in *G. muris*.

### **2.2.5 Identification and analysis of Metamonad U3 snoRNAs**

As described in Chapter 1 the U3 snoRNA is critical for the processing and cleavage of pre-rRNA in eukaryotes. Previous analyses of U3 snoRNAs from diverse eukaryotes only examined a single member of the metamonad lineage, *G. lamblia*. These studies suggested the *G. lamblia* C/D snoRNA GlsR1 as a possible *G. lamblia* U3 candidate, however this RNA lacks all but one of the conserved helices found in other eukaryotic U3 RNAs and possesses a very short 5' extension with no clear base-pairing potential to the pre-18S rRNA (Figure A.1.8A) (Marz and Stadler 2009). This would also be the only known U3 snoRNA to target a modification site, as GlsR1 guides methylation of C1325 in the *G. lamblia* 18S rRNA, a site homologous to the modification performed by snR70 in yeast (Yang, et al. 2005). To determine if this predicted but structurally unusual U3 could be a conserved feature of U3 snoRNAs in the metamonads, we used the Infernal software package to perform CM searches using seven metamonad genomes including: five fornicates (*Kipferlia bialata* and four additional diplomonad species: *Giardia lamblia*, *Giardia muris*, *Spironucleus salmonicida* and *Spironucleus vortens*), a parabasalid (*Trichomonas vaginalis*), and two oxymonads (*Monocercomonoides exilis* and *Streblomastix strix*). Initial searches using the consensus U3 structure obtained from the rFam database yielded

plausible U3 snoRNA candidates for *S. salmonicida*, *S. vortens*, *M. exilis*, and *S. strix*. The absence of a putative U3 homolog for both *G. lamblia* or *G. muris* suggests that if a U3 homolog is present, either as GlsR1/GmsR1 or another RNA, it has diverged significantly from previously described U3 structures.

We searched for possible alternative U3 candidates using the RNA co-precipitation data obtained in our lab from co-precipitation with the core C/D snoRNP RNA-binding protein Snu13p in *G. lamblia* (see Chapter 3). Snu13p is a component of the U3 snoRNP in other species and would be predicted to bind any U3 snoRNA present in *G. lamblia*. We found that the ncRNA ‘Candidate 17’, an ncRNA whose function was not previously determined, was significantly enriched in this data. A search of our ncRNA candidates in *G. muris* revealed a highly expressed RNA (nc195) with significant sequence similarity to Candidate 17. Alignment of these two *Giardia* RNAs with the U3 candidates from the other metamonad species highlighted regions corresponding to the conserved U3 box elements, C', B, C, and D and the immediate surrounding nucleotides (Figure 2.5A).



**Figure 2.5. Metamonad U3 snoRNA candidates maintain critical U3 features.** (A) Alignment of the 6 predicted metamonad U3 snoRNAs, aligned using ClustalOmega (<https://www.ebi.ac.uk/Tools/msa/clustalo/>) with manual curation. Conserved sequence elements are indicated with coloured text: GAC sequence (orange), and box C' (blue), B (purple), C (green), and D (red). \* indicates an identical nucleotide present at this position in all 6 RNAs. (B) Base-pairing interactions formed between the U3 snoRNA and 18S rRNA homologs for *S. cerevisiae* and the 6 metamonad U3 snoRNAs based on the intermolecular helix II and III that was detected recently in yeast SSU processome cryo-EM structures. For *S. vortens* and *M. exilis*, only helix II and III respectively are shown due to missing rRNA sequences in genomic databases. Roman numerals II and III refer to the names of U3-18S intermolecular helices. First and last nucleotides of the base-pairing regions are numbered for positions within each RNA. Purple text is 18S rRNA sequence, black text is U3 snoRNA sequence. Metamonad U3 secondary structures are shown in Figure A.1.9.

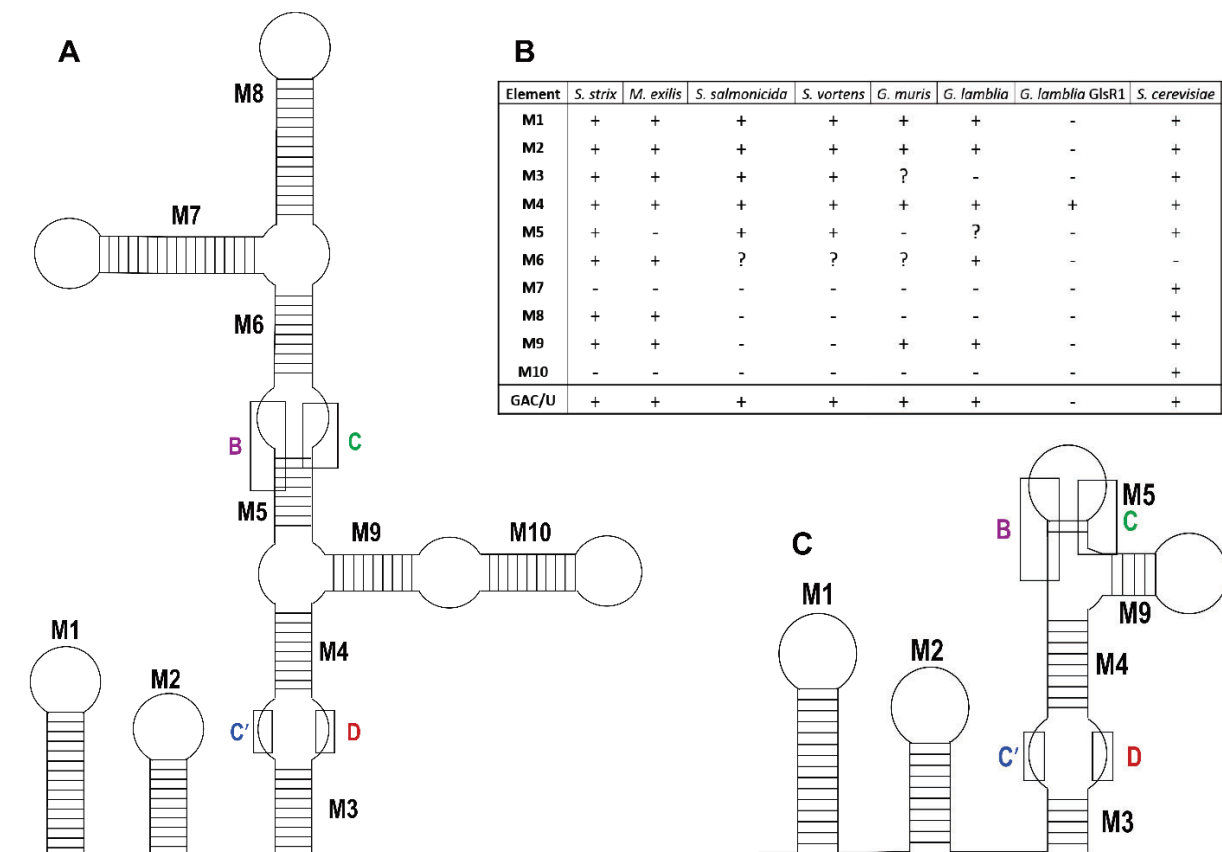
The 5' domain of U3 snoRNAs contain a collection of sequences that form critical base-pairing interactions with the 18S rRNA sequence and 5' ETS of the pre-rRNA. Previous biochemical analysis and genetic studies in yeast and humans suggested the formation of five helices between the rRNA and U3, designated helix I, II, III, V, and VI, where I-III form within the 18S, and V and VI form with the 5' ETS (Beltrame and Tollervey 1995; Hughes 1996; Méreau, et al. 1997). However, recent cryo-EM structures of the yeast SSU processome and mutational analysis of the yeast U3 determined that an alternative helix III forms between U3 and a different region of the 18S rRNA, and helix I is absent entirely (Barandun, et al. 2017; Sun, et al. 2017; Clerget, et al. 2020). Alignment of the metamonad and yeast 18S rRNAs shows conservation of nucleotides involved in helix II and the newly identified helix III. Using sequence alignment and manual inspection we found that the U3 snoRNAs from *S. salmonicida*, *S. strix* and both *Giardia* species are able to form helix II and III with their respective 18S rRNA homologs (Figure 2.5B). Only incomplete rRNA sequence data is available for *S. vortens* and *M. exilis* but from this we were able to predict formation of helix III in *M. exilis* and helix II in *S. vortens*. This data shows that the newly identified rRNA-U3 helices are conserved in the metamonads, supporting the

evolutionary importance of these interactions and our classification of these RNAs. Plausible formation of helix V and VI for *G. lamblia* and *G. muris* can also be predicted with the 5' ETS and U3 homologs but high G/C content and poor conservation in this rRNA region between species makes accurate prediction of helices difficult and will likely require biochemical verification to accurately annotate (data not shown). Based on the conservation of the internal U3 box elements, predicted formation of critical base-pairing interactions with the respective 18S rRNAs, and significant enrichment of *G. lamblia* Candidate 17 in Snu13p co-precipitations we conclude that these RNAs rather than GlR1/GmR1 are the *Giardia* U3 snoRNAs. Going forward we refer to Candidate 17 and nc195 as Gl-U3 and Gm-U3 respectively.

We next examined the evolution of U3 in the metamonad lineage and compared their features to U3 snoRNAs in other taxonomic groups. 97% of U3 snoRNA sequences in the rFAM database contain a GAY (Y = C or U) motif near their 5' end. All six metamonads conserve this sequence (4/6 "GAC", 2/6 "GAU"). The distance between the B and C box elements is significantly larger in the oxymonad U3s than the diplomonads and sequences are not maintained between metamonad groups (Figure 2.5A). Predicted secondary structures for the U3 snoRNAs produced using MFOLD and manual curation show this region forms two extended stem loops in the oxymonad RNAs that are absent in both *Spironucleus* and *Giardia* species (Figure A.1.9). These helices correspond to helix M8 and M9, following the terminology of Marz and Stadler (Marz and Stadler 2009), which are found in U3 snoRNAs from representatives in all other major eukaryotic domains (Figure 2.6A and B). The absence of these and other evolutionarily conserved helices in *Spironucleus* and *Giardia* result in compact U3 RNA structures. Together this data suggests that the ancestor of metamonads possessed a larger, more complex U3 structure that has undergone significant size reduction within the diplomonad lineage. Including our newly-



identified sequences with those analyzed by Marz and Stadler point to a U3 snoRNA with most of the additional helices found in higher eukaryotes being present in LECA (Marz and Stadler 2009). Loss of many U3 structural features during diplomonad evolution suggests that they are not strictly required for the core function of the U3 snoRNP as only helices M1, M2, and M4 are predicted to stably form in all four examined diplomonad U3 homologs (Figure 2.6B and C). Other helices may form in diplomonad U3s *in vivo*, including a short, mostly non-Watson-Crick-paired M5, an M9 helix in *Giardia* and M3 helix in *Spironucleus*, but these would still be highly reduced compared to those present in higher eukaryotes (Figure 2.6B and Figure A.1.9).



**Figure 2.6. Size reduction of the U3 snoRNA in diplomonads.** Organizational schematic for the different intramolecular helices in U3 snoRNA secondary structures across various eukaryotic lineages (A) and in the diplomonads (C). Helix names and the secondary structure in (A) are based on the schematic from Marz and Stadler, 2009. (B) Table of the helices predicted to form in the U3 snoRNAs from the six examined metamonads species, GlsR1 snoRNA from *G. lamblia* and the *S. cerevisiae* U3 snoRNA. Presence of the GAC sequence motif is also indicated. (+) indicates that the helix is present, (-) absent, or (?) if a helix does form it would be short, contain many non-Watson-Crick pairs and likely be relatively unstable.

### 2.2.6 Small nuclear RNA candidates from *G. muris*

Each of the five conserved snRNAs of the spliceosome (Chapter 1): U1, U2, U4, U5, and U6 possess highly conserved sequence and secondary structural features important for their function. Previous searches for snRNAs in the diplomonads *G. lamblia*, *S. salmonicida*, and *S. vortens* have described varying incomplete sets of these RNAs (Hudson, et al. 2012; Hudson, et al. 2019). Identified diplomonad snRNAs possess important core spliceosomal RNA elements but they are divergent enough to escape detection using conventional search strategies based on snRNA features from other eukaryotic snRNAs. Significant sequence and structural differences are even found between snRNA homologs of different diplomonad species. This rapid rate of divergence in normally well-conserved structures make diplomonads an intriguing lineage for the study of snRNAs and their potentially diversified role in splicing.

We performed BLASTN and CM searches to the *G. muris* genome using the previously identified *G. lamblia* snRNAs as queries. These searches detected strong candidates for U2, U4 and U6 which show significant sequence conservation with the *G. lamblia* snRNAs (Figure A.1.10). Each snRNA candidate possesses a 3' end processing motif and is well represented in our small RNA library, with similar read counts for each of the three RNAs (Supplemental file 1). Each candidate was examined for important sequence elements and the ability to form conserved inter- and intramolecular helices. In other eukaryotic species, the U4 and U6 snRNAs form an extensive network of base-pairs during the early stages of splicing as part of the U4/U6•U5 tri-snRNP (Bringmann, et al. 1984; Hashimoto and Steitz 1984; Brow and Guthrie 1988; Wilkinson, et al. 2020). Our *G. muris* U4/U6 candidates can pair to form the evolutionarily conserved intermolecular helix I and II which meet to form a tri-helical junction with the U4 intramolecular

5' stem loop (SL) (Figure 2.7A). The *G. muris* U4 5' SL can also form the ubiquitous K-turn structure critical for U4 snRNP assembly in other eukaryotes (Nottrott et al. 1999).

Following recruitment of the tri-snRNP to the intron, the intermolecular helices formed between U4 and U6 are unwound by helicases allowing U6 and U2 to pair, generating structures important for the catalytically active spliceosome (Madhani and Guthrie 1992; Zhang, et al. 2019). U6 also base-pairs to the 5' splice site (5'ss) of the intron via the evolutionarily invariant 'ACAGAGA' sequence while U2 pairs to the intronic branch point sequence (BPS) causing the catalytic branch point A to bulge (Parker, et al. 1987; Valadkhan et al. 2009; Zhang, et al. 2019; Wilkinson, et al. 2020). U6 snRNAs also contain an 'AGC' sequence that forms helix Ib in the U2/U6 complex and is key in facilitating splicing (Yu et al. 1995; Mefford and Staley 2009). Intermolecular pairing between *G. muris* U2/U6 forms the conserved helices I-III which facilitates formation of the U6 intramolecular stem loop (ISL) (Figure 2.7C). The U6 ISL is a critical structure involved in magnesium binding during the catalytic stages of splicing (Valadkhan, et al. 2009; Zhang, et al. 2019). Both a canonical 'ACAGAGA' and 'AGC' triad sequence are found in the *G. muris* U6, and U2 possesses a stretch of nucleotides between helices I and III capable of base-pairing with the conserved *G. muris* BPS 'ACUAACACGCAG'. Interestingly, the interaction between the *G. muris* U2 and BPS can result in either of two adjacent intronic adenosines to bulge. This feature is also observed for the *G. lamblia* and *Spiroplasma* spp. U2 snRNAs (Hudson, et al. 2012). The intramolecular structures SL I-IV of U2 are also conserved in our candidate (Figure 2.7C).



**Figure 2.7. Divergent *G. muris* spliceosomal snRNAs.** Predicted intermolecular and intramolecular base-pairing for the *G. muris* U1, U2, U4, and U6 snRNA candidates. **(A)** U4/U6 pairing. The U4 kink-turn in the 5' stem loop and Sm binding site are boxed. The conserved ACAGAGA sequence of U6 is also boxed. **(B)** U1 is depicted base-pairing to the 5' splice site of the outer arm dynein heavy chain  $\gamma$  (OADHC $\gamma$ ) *trans*-spliced intron. The U11-like sequence and conserved helices of the 'cloverleaf' structure are labeled. **(C)** U2/U6 pairing. The ACAGAGA sequence of U6 is shown pairing to the 5' splice site of OADHC $\gamma$ . The intron branch point sequence is depicted pairing to U2 and bulging the catalytic adenosine. The putative U5 structures are found in Figure A.1.11.

We next extended our search for more divergent candidates of U1 and U5 since CM searches of the *G. muris* genome were unsuccessful in detecting convincing homologs (Nawrocki and Eddy 2013). In other species, the U1 snRNA binds to the 5'ss during the early stages of splicing via a sequence near the snRNAs 5' end (Mount, et al. 1983; Wilkinson, et al. 2020). A candidate U1 snRNA in *G. muris* should therefore possess a sequence complementary to the *G. muris* 5'ss. Using the reverse complement of the extended 5'ss of *G. muris* consensus sequence (GU[A/U]UGUG) we queried our identified ncRNA candidates and their upstream regions (+50 bp). The upstream region (+22-16) of the nc038 candidate RNA contains a stretch of seven nucleotides (CACAUAC) complementary to the *G. muris* 5'ss consensus and structural predictions for this RNA using MFOLD show it can form the conserved U1 'cloverleaf' structure (Figure 2.7B). Like the *G. lamblia* U1, nc038 contains a minor spliceosomal U11 snRNA-like SL III sequence and lacks both an SL IV structure and U1-70K protein-binding site in SL I (Hudson, et al. 2012). In contrast, unlike *G. lamblia*, both the U1A binding site in the SL II loop and the Sm binding site are also absent in our *G. muris* candidate revealing a considerable degree of sequence divergence in normally well conserved regions. The read abundance for nc038 in our small RNA library is in slight excess of what is observed for the other three snRNA candidates; consistent with findings in humans where U1 is present in approximately 10-fold excess of other snRNAs (Baserga and Steitz 1993; Dvinge et al. 2019).

U5 snRNA is the only snRNA yet to be definitively identified in *G. lamblia*. As with U1 we manually searched uncharacterized ncRNAs in the two *Giardia* species for conserved U5 features. We found that *G. lamblia* ‘Candidate 5’ RNA and *G. muris* nc158 can form stem structures containing a U rich loop, resembling the U5 loop I sequence (Figure A.1.11) (Wilkinson, et al. 2020). Loop I of U5 is critical in splicing as it pairs to the end of the adjacent upstream exon during splicing, tethering it to the spliceosome following the breakage of the RNAs phosphodiester backbone caused by the first transesterification reaction (O’Keefe and Newman 1998). The two RNAs are found in syntenic positions in the two *Giardia* genomes but have divergent sequences and vary considerably in stem length. Additionally, nc158 does not group with the other four snRNAs based on read count and is poorly represented with only 66 reads mapping to the coding strand. In the absence of other candidates, it is plausible these RNAs could act as U5 homologs, but we cannot conclude based on our analysis that Candidate 5 and nc158 are definitively the *Giardia* U5 snRNAs without additional experimental verification.

#### **2.2.7 A highly-reduced Telomerase RNA (TER) component in *Giardia* spp.**

Telomeres are long stretches of a short repetitive sequence found at the end of eukaryotic linear chromosomes that are critical in preventing loss of genetic information during DNA replication. Telomeres are generated by the Telomerase RNP complex, composed of a set of proteins including a telomerase reverse transcriptase (TERT), and a telomerase RNA component (TER) (Musgrove, et al. 2018). TERs can vary dramatically in size from approximately 150 to over 2000 nt, but minimally contain four core sequence and structural features: a template sequence, template boundary element (TBE), stem terminus element (STE), and pseudoknot (Podlevsky and Chen 2016). Template sequences are most often 1.5-2X repeats of the complementary sequence to the telomeric repeat (eg. in humans the telomeric repeat is 5’-

(GGTTAG)<sub>n</sub>-3' and the template region of the TER is 5'-(CUAACCCUAAC)-3' ) (Podlevsky and Chen 2016). A TERT homolog has been identified in *G. lamblia* and telomere extension has been characterized but no TER has been detected in any *Giardia* species or other metamonad indicating that they could be considerably evolutionarily divergent (Malik et al. 2000; Uzlíková et al. 2017).

Following our analysis of other ncRNA classes only five ncRNA candidates with >200 reads remained unassigned. Additionally, in *G. lamblia* all but three predicted ncRNAs have been assigned functions (Hudson, et al. 2012) (Chapter 3 and this work). Pairwise alignment of these uncharacterized RNAs from the two *Giardia* species revealed a high degree of sequence similarity between GlsR28 from *G. lamblia* and nc136 in our *G. muris* ncRNA library (Figure A.1.12). Secondary structure predictions for GlsR28 and nc136 using MFOLD and manual curation found that the RNAs fold into nearly identical structures with only minor sequence variation, mostly limited to single stranded regions (Figure 2.8A and B). These *Giardia* TER candidates both contain a 10 base-pair long stem with a conserved apical loop capable of pairing with a nearby upstream sequence to form the TER pseudoknot. *G. lamblia* and *G. muris* share the same telomeric repeat sequence 5'-(TAGGG)<sub>n</sub>-3' (Morrison, et al. 2007; Xu, et al. 2020). We found that both candidate TERs contain a short sequence complementary to the telomeric repeat in a single stranded region within the pseudoknot-containing loop, 'UACCCUA' and 'UACCCU' for *G. lamblia* and *G. muris* respectively which could act as the template sequence. Intriguingly, both *Giardia* TER candidates are similar in length to and resemble the secondary structure of the ciliate *Tetrahymena thermophila* TER, with the template and pseudoknot elements residing in similar relative positions within the RNA. The TBE (stem II) found in *T. thermophila* and many other eukaryotic TERs is missing from the *Giardia* TERs (Figure 2.8A-C). We instead propose that the helix enclosing the template sequence and pseudoknot region (or T-PK) could act as the TBE in *Giardia* (Figure 2.8A

and B, stem labeled TBE). This type of TBE is strikingly similar to those found in a majority of vertebrate and Angiosperm TERs (Figure 2.8E and F) (Chen and Greider 2003; Song et al. 2019; Logeswaran et al. 2020). Use of this mechanism is dependent on the distance between the 5' end of the template sequence and the T-PK closing helix for extension boundary definition in studied eukaryotes. Intriguingly, this distance is six nucleotides in both *Giardia* candidates and the *Arabidopsis thaliana* TER (Figure 2.8A, B and F). It therefore appears plausible that this helix could function in a similar manner in *Giardia* as it does in these other eukaryotes. Finally, the terminal 5' and 3' ends of the *Giardia* TER candidates form a long discontinuous stem and helical junction. The STE of vertebrate, fungal, and land plant TERs are composed of variations of a similar helical junction, proposed to be an ancestral feature of TER, though at a somewhat different relative position in the RNA. Given this similarity and small size of the *Giardia* TERs, this region likely also acts as the *Giardia* STE.



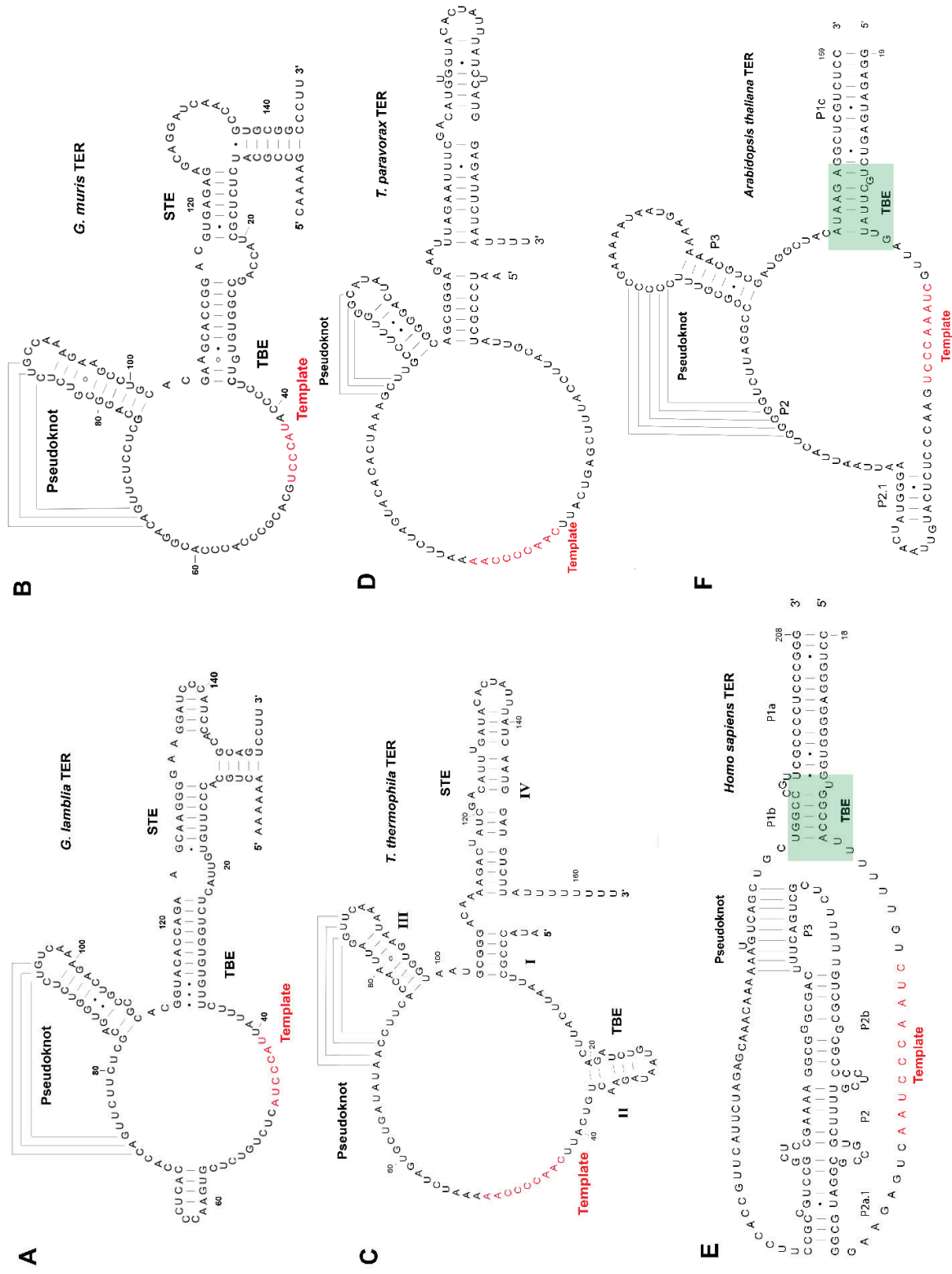


Figure 2.8. Comparison of *Giardia* telomerase RNA (TER) structures with other eukaryotic TERs.

**Figure 2.8. Comparison of *Giardia* telomerase RNA (TER) structures with other eukaryotic TERs.** TER secondary structures are based on MFOLD predictions and manual curation for *G. lamblia* (A), *G. muris* (B), and the previously characterized *T. thermophila* (C), *T. praravorax* (D), *H. sapiens* (T-PK domain) (E), and *A. thaliana* (T-PK domain) (F) telomerase RNAs. Gray lines indicate base-paired nucleotides that are part of the pseudoknot structure. Template regions predicted to be used to produce telomeric repeats are labeled and indicated in red. The four evolutionarily conserved core TER features are labeled: template, template boundary element (TBE), pseudoknot, and stem terminus element (STE) for A-C. Identified helices from *T. thermophila* TER are labeled with roman numerals I-IV and with conventional P# nomenclature for *H. sapiens* and *A. thaliana*. Experimentally validated TBE regions of the RNA are highlighted in green boxes and labeled for *H. sapiens* and *A. thaliana*.

## 2.3 Discussion

### 2.3.1 Conservation of a ncRNA processing motif in *G. muris*

Following our ncRNA analysis, we re-examined the consensus sequence of the *G. muris* 3' processing motif, now including all identified ncRNAs and the previously characterized *trans*-spliced introns. This consensus closely matches the one identified from the first 10 ncRNAs in *G. muris*: 5'-CCTTYDHTNAA-3' (Figure A.1.13). The only notable difference between the motif in the two *Giardia* species is an apparent bias for a T at the position directly upstream of the conserved 'CCTT' at the start of the motif in *G. lamblia*. No nucleotide preference was observed for this position in *G. muris* (Position 1, Figure A.1.13). The presence of a nearly identical 11 nt 3' processing motif at the end of all the identified ncRNAs in *G. lamblia* and *G. muris* shows the processing mechanism is highly conserved in the *Giardia* genus. The cellular machinery responsible for recognition and cleavage of the motif has not yet been identified but is therefore likely also conserved. Analysis of the limited number of currently identified *S. salmonicida* ncRNAs failed to detect similar conservation at their 3' termini, indicating that the motif likely evolved after the evolutionary branching of *Spironucleus* and *Giardia*. What could have driven the

evolution of such a well conserved and apparently critical processing signal in all non-rRNA/tRNA ncRNAs in such a relatively short period remains unknown but warrants further experimental investigation.

Only five *G. muris* ncRNA candidates with >200 reads remain unclassified. This includes two RNAs whose sequences partially resemble GlrR24 and GlrR26 H/ACA snoRNAs but lack sufficient evidence for definitive classification (data not shown). Another of these RNAs contains elements resembling C and D boxes (nc151) and groups with the other C/D snoRNAs, based on read count in our library. But nc151 lacks a predicted modification target and contains more degenerate box elements than other C/D snoRNAs in *G. muris*. Only one of our predicted ncRNAs, the possible U5 snRNA, has below 200 reads. Therefore, our RNA library was robust in detecting ncRNA expression and our prediction criteria were appropriately stringent, as RNAs predicted based on conserved sequence and structural features made up nearly all well-represented RNAs in our library.

Of the 216 ncRNA candidates included in our analysis, a collection of 63 predicted from the genome were found to have identical sequences. These map downstream of the mature 28S rRNA homolog sequence of the various rDNA operons found throughout the genome (data not shown). They were originally predicted as ncRNAs based on the presence of the 3' processing motif. The motif in these regions is flanked on either side by long stretches of sequence that are also highly conserved between different genomic loci, suggesting a functional importance. This likely does not represent a small RNA but instead could indicate a role for the motif in rRNA processing. A search for the motif sequence downstream of the 28S rDNA in *G. lamblia* did not detect a similar arrangement.

### 2.3.2 *snoRNA repertoires are conserved between Giardia species*

We detected *G. muris* homologs for nearly all *G. lamblia* snoRNAs of both classes with known nucleotide modification targets. We also detected a *G. muris* GlsR13 homolog (GmsR13) targets the U4 snRNA in *G. muris*, which is not seen in *G. lamblia*. In higher eukaryotes some snoRNA-like RNAs localize to the Cajal bodies (scaRNAs) and guide modification of snRNAs (Jády et al. 2003; Richard, et al. 2003). Identification of GmsR13 as a potential guide for U4 modification is the first example of a snoRNA-guided snRNA modification in a metamonad. SnRNA modification occurs in the Cajal bodies rather than the nucleolus in higher eukaryotes; however, Cajal bodies have not been identified in *Giardia*. The presence of the 3' processing motif in both snoRNAs and snRNAs indicates a shared biogenesis pathway. This could suggest the existence of a central subnuclear hub, perhaps the nucleolus, for RNP biogenesis that would allow for GmsR13 to interact with U4 without the requirement for special localization signals.

The additional targets identified for GmsR8/GlsR8 represent the only current example of a snoRNA targeting multiple positions in the diplomonads. The ability to target multiple positions using a single snoRNA may be advantageous with such a small complement of guide RNAs but appears not to be a common feature of *Giardia* snoRNAs. The modification status of the additional target nucleotides predicted for GmsR8/GlsR8 is unknown, and it is possible that only a subset of the three positions are actually modified. It is unclear what mechanism would be responsible for target site selection in this case, as the guide RNAs have equivalent length pairing potential to all three sites.

### ***2.3.3 A conserved set of snoRNA-guided modifications present in functionally important positions of ribosomal RNA in species with reduced genomes***

Comparative analysis of species that have undergone genomic reduction with distinct evolutionary histories provides an opportunity to examine the common features of cellular complexes that are being selectively maintained. We identified a set of shared snoRNA-guided modifications in *G. muris* and the distantly-related *Guillardia theta* nucleomorph rRNAs that are predominately found in functionally important regions of the rRNA. Clustering of nucleotide modifications in functionally important regions of the ribosome is prevalent in eukaryotes and prokaryotes and occurs regardless of the overall number of modifications in any particular species (Sergiev et al. 2018). Many of these positions have been studied in yeast via snoRNA knockouts to eliminate modifications and been determined to play roles in translational fidelity. For example, helix 69 (H69) of the LSU rRNA forms part of the B2a intersubunit bridge and contacts tRNA in the A and P sites during translation (Yusupov et al. 2001). In yeast this helix contains 5 modifications: four  $\Psi$  and one 2'-O-methylation (Nm). *G. muris* and *Gu. theta* rRNA contain the corresponding H69 Nm at the same relative position, guided by GmsR7 and GtNM-R8 respectively (Figure A.1.4 and A.1.6). Knockouts of the yeast snoRNAs guiding these five modifications caused significant growth defects and reduced rRNA stability when >3 modifications were removed, especially when growth temperatures were elevated (Liang, et al. 2007). Knockout of the H69 Nm alone also resulted in changes to both mRNA readthrough and frameshifting frequencies (Baudin-Baillieu, et al. 2009). Knockouts of the yeast homolog of Candidate 1/GtNM-R12 targeting H34 of the decoding region in the 18S rRNA also altered frameshifting, but only affected readthrough in combination with removal of other modifications in the same region of the ribosomes 3D structure. Finally,

knockout of the yeast GmsR29/GtNM-R11 homolog which targets a nucleotide in the peptidyl transferase center causes significant growth defects under a variety of extra-ribosomal growth challenge conditions, with particularly significant negative effects occurring when grown in the presence of chemicals causing redox imbalance (eg. DTT and *tert*-butyl-HOOH) (Esguerra et al. 2008). These modifications may play an even larger role in species like *G. muris* or *Gu. theta* as they often lack the additional modifications found in the same 3D region of the ribosome that are present in these higher eukaryotes.

The predicted rRNA targets of nearly all C/D snoRNAs not shared between the *Gu. theta* NM and *G. muris* are modified in both yeast and humans (Table 2.1). This overall pattern of conservation of snoRNA-guided modification suggests a model in which a set of common “core” snoRNA-guided modifications are preferentially maintained at the most functionally critical rRNA sites during genomic reduction (eg. those shared between *G. muris*, *Gu. theta* NM, yeast and humans) and “supported” by a small number of additional beneficial but not as critical rRNA modifications (eg. conserved sites in yeast, humans and either *G. muris* or *Gu. theta* NM).

Nearby the GmsR7 guided methylation of H69 in *Giardia* there is a snoRNA-guided  $\Psi$  modification. The importance of an additional modification in H69 is consistent with the cumulative effects of multiple snoRNA deletions in yeast (Liang, et al. 2007). The *Gu. theta* NM genome appears devoid of H/ACA snoRNAs (Åsman, et al. 2019) preventing comparison of any snoRNA-guided pseudouridylation sites with *G. muris*. It is possible that the 2'-*O*-methylation is sufficient to maintain translational fidelity in *Gu. theta* NM ribosomes. Indeed, different combinations of snoRNA deletions had varying effects on ribosome function and translation in yeast, with removal of more modifications through

additional snoRNA deletions not always resulting in larger translational effects (Liang, et al. 2007; Baudin-Baillieu, et al. 2009). The reduced set of modifications found in *Gu. theta* NM and *G. muris* could represent a balance in maintaining a modification distribution sufficient for efficient core ribosome function while missing other modifications that provide additional translational control in higher eukaryotes. We also note that H/ACA snoRNAs or stand-alone pseudouridine synthases (protein-only enzymes) imported from the *Gu. theta* nucleus could modify NM rRNA; there has been no experimental mapping of rRNA modification to confirm a complete lack of  $\Psi$ . Determining the full complement of modifications in both *G. muris* and the *Gu. theta* NM rRNA will be important in understanding which modifications are truly key for accurate translation in these genomically reduced species.

#### ***2.3.4 Cm34 tRNA modification targeted by the Giardia RNase P-snoRNA diRNP is conserved throughout eukaryotes and archaea***

tRNAs are heavily post-transcriptionally modified in species throughout the tree of life. Only a handful of these modifications are 2'-O-methylations and only three of these (positions 18, 32, and 34) are conserved across bacterial, archaeal and eukaryotic species. (Hori 2014; Ayadi et al. 2019). The N34 modification is located at the wobble position of the anticodon loop of some tRNA species and is suggested to be important in stabilizing the codon-anticodon interaction during decoding. In most tRNA species in many bacteria, archaea, and eukaryotes, N34 2'-O-methylation has been assigned to stand-alone enzymes or protein-only complexes (Ayadi, et al. 2019). However, in some archaeal species C/D s(no)RNA-guided tRNA modifications have been identified. *Pyrococcus abyssi* and *Haloferax volcanii* each possess a C/D sRNA that guides Nm modification of C34 and U39

in tRNA<sup>Trp</sup> (d'Orval, et al. 2001; Joardar, et al. 2011). An additional four C/D sRNA-guided tRNA modifications have been proposed in *P. abyssi*, three of which target the N34 position of a tRNAs, including one in elongator tRNA<sup>Met</sup>. *H. volcanii* also contains a C/D sRNA that guides Cm34 in elongator tRNA<sup>Met</sup>, and homologs of these sRNAs have been predicted in species throughout Euryarchaea. The first example of a snoRNA-guided tRNA modification in eukaryotes was only recently discovered; in an identification of the cooperative roles of SNORD97 and SCARNA97 in C34 2'-*O*-methylation of the human elongator tRNA<sup>Met</sup> (Vitali and Kiss 2019). Potential homologs for SNORD97 and SCARNA97 were predicted throughout vertebrates; and SNORD97 homologs were bioinformatically detected in some invertebrates such as the tunicate *Ciona intestinalis*, and the plants *A. thaliana* and *Brassica napus*. The overall sparsity of identified snoRNA-guided tRNA modifications in archaea and eukaryotes suggests a barrier to the evolution of these interactions, but a shocking proportion of these modifications occur at position N34, with a particularly strong apparent bias for tRNA<sup>Met</sup>. The predicted ability of GmsR15/GlsR15 to target tRNA<sup>Met</sup> C34 in *Giardia* adds strong support for the wide evolutionary distribution and conservation of this snoRNA-dependent modification mechanism in eukaryotes. This is among the first lines of evidence suggesting an important role for snoRNAs in tRNA modification in eukaryotes and retention of this interaction in both studied *Giardia* species (which each retain such a small repertoire of snoRNAs) further emphasizes its functional importance. Identification of snoRNA candidates capable of guiding modification of C34 in tRNA<sup>Met</sup> within the eukaryotic supergroups Obazoa, Archaeplastida, and now Metamonada, as well as in archaeal species, indicates that this tRNA-modifying guide RNA was present in LECA and has since been lost in some eukaryotic lineages or remains to be identified.



To our knowledge, the fusion of these *Giardia* tRNA-targeting snoRNAs to the catalytic RNase P RNA is the first case described of any RNase P RNA containing a domain belonging to another ncRNA class (Guerrier-Takada, et al. 1983; Kikovska, et al. 2007). Our data showing the association of the fused RNA with protein components of both RNP classes (described in Chapter 3) strongly supports the hypothesis that both the RNase P and snoRNA domains of the diRNP retain their functions, constituting a multi-functional hybrid tRNA processing RNP. Fused snoRNP domains have been observed in various RNPs in some species; including vertebrate and *Trypanosoma brucei* telomerase RNPs which contain H/ACA and C/D snoRNP domains respectively. However, in these instances the snoRNA domain is involved in RNP processing and assembly rather than guiding RNA modification (Gupta et al. 2013; Nguyen et al. 2018). This makes the *Giardia* RNase P-snoRNA diRNP the first snoRNA-ncRNA fusion to retain its modification-guide function. The fused arrangement of RNase P and GlsR15/GmsR15 provides a very plausible explanation for how this individual snoRNA could localize to a tRNA target rather than the site of pre-rRNA processing in *Giardia* cells. The mechanism by which tRNA<sup>Met</sup> colocalizes with SNORD97/SCARNA97 for modification in humans remains unresolved. If this snoRNA is evolutionary conserved in eukaryotes multiple mechanisms have likely evolved to facilitate colocalization of the snoRNA guide and tRNA target.

### ***2.3.5 Reduction in U3 snoRNA size during metamonad evolution***

Our identification of U3 snoRNA homologs from six metamonad species corrects the previous erroneous classification of GlsR1 as the *G. lamblia* U3 homolog and gives significant insight into the sequence and structural evolution of U3 in this lineage. U3 possess two major domains, the 5' domain which base-pairs with the pre-rRNA (M1 and

M2) and the 3' domain which contains the snoRNA box elements and extended helices (M3-M10) (Samarsky and Fournier 1998). Cryo-EM structures of the SSU processome in yeast show that the core U3 snoRNP proteins (Snu13p, Nop56, Nop58, Nop1 (fibrillarin), and Rrp9) bind to the 3' domain of U3, contacting the RNA at the four box elements (C'/D and B/C) and the M4/M5 helices that separate the two sets of boxes (Barandun, et al. 2017; Sun, et al. 2017). Additional contacts were also detected between the M9 helix and Nop56/Nop1 through UV-crosslinking (Granneman et al. 2009). Despite the loss of many structural elements found in other eukaryotes, the four diplomonad U3 homologs retain the four box elements and M4 helix. Gl-U3/Gm-U3 are also predicted to form a short M9 helix which could help facilitate the additional interactions with Nop56 and fibrillarin (Nop1p). It is likely the *Giardia* homologs are near the lower size and complexity limit for functional U3 snoRNAs (Gl-U3 ~180 nt and Gm-U3 ~172 nt) as any further reduction in the 3' domain would be predicted to disrupt binding of the core U3 snoRNP proteins. Similarly small U3 snoRNAs found in *Trypanosoma brucei* (143 nt) (Mottram et al. 1989; Hartshorne and Toyofuku 1999) and *Euglena gracilis* (180 nt) (Greenwood, et al. 1996) have also been characterized and support this lower size range.

Conservation of the recently identified helix III that forms between U3 and the 18S rRNA in the metamonad U3 snoRNAs adds further evidence for its formation in place of the previously predicted helix III structure. Intermolecular helix I of the U3-18S rRNA interaction was not observed in the yeast cryo-EM structures and was subsequently found to be dispensable (Sun, et al. 2017; Clerget, et al. 2020). The 'UUUCU' sequence of U3, previously thought to form part of helix I, is conserved in all metamonads except *G. lamblia*; but only *S. salmonicida* and *S. strix* would be capable of forming a stable helix I

due to non-compensatory changes in the rRNA sequences of most metamonads (Figure 2.4A and A.1.14). Conservation of this U3 sequence in the metamonads and across eukaryotes without conservation of a paired rRNA sequence to form helix I implies this U3 sequence performs some other conserved function. Mutation of this sequence in yeast U3 did not have an observable impact on cell growth or pre-rRNA processing but has been suggested to play a role in U3 RNA stability (Clerget, et al. 2020). The change in G1-U3 to ‘GUUUU’ could mean that some sequence flexibility is tolerated in this region or that this mechanism of RNA stabilization is absent in *G. lamblia*.

The similarity in overall structure between the oxymonad U3 snoRNAs and those found in higher eukaryotes adds new evidence supporting the hypothesis that a U3 containing many of the elongated helices found in higher eukaryotes was present in LECA. The diplomonad U3 snoRNAs reveal yet another example of RNA structure mirroring overall genomic minimalization, an intriguing feature now apparent in a number of diplomonad ncRNAs

### **2.3.6 Rapid divergence of snRNAs in the diplomonads**

Both the sequence and structures of the U2, U4, and U6 snRNAs are similar in *G. muris* and *G. lamblia*. The most divergent of the three is the U4 snRNA which shares 66/107 nucleotides between the *Giardia* species. Most of the sequence variation occurs in the K-turn containing 5' SL with compensatory mutations having occurred to maintain base-pairing in the stem (Figure A.1.10). Our previous studies of *Spironucleus* snRNAs only identified candidates for U2 and U5 RNAs (Hudson, et al. 2019), but a complete complement of all five snRNAs can be detected using CM searches (with conventional snRNA models) in the oxymonad *M. exilis*. This indicates that the metamonad ancestor

likely had more structurally conventional snRNAs. The difficulty in identifying U1, U4 or U6 candidates in *Spironucleus spp.* along with the unusual *G. muris* U1 and U5 candidates shows that rapid sequence and structural evolution has occurred within the diplomonads, even in normally well-conserved regions. U1 appears to be the most rapidly evolving as considerable differences are observed even between *G. muris* and *G. lamblia* U1 snRNAs. Intriguingly, a recent analysis of *G. lamblia* and *S. salmonicida* spliceosomal proteins found no core U1 snRNP proteins when searching for encoding genes in either genome (Hudson, et al. 2019). This absence of U1 snRNP protein factors in the diplomonads could have led to the observed loss of the typically conserved corresponding protein binding sequences and the divergence in structural elements in the *G. muris* U1 snRNA as they would no longer be evolutionarily constrained to maintain protein binding. Similar arguments can also be applied to the as yet unidentified *Spironucleus* U1. It also remains possible that these unique U1 features are being driven by the emergence of lineage-specific U1-associated proteins or rapid evolution of conserved U1 proteins that makes their detection by bioinformatic strategies problematic.

Previous work speculated that the *G. lamblia* snRNAs may be a hybrid of the U2 and U12-type snRNAs based on observed sequence and structural features. The newly identified *G. muris* snRNAs retain few of these U12 features. This finding, along with the more conventional major spliceosome-like oxymonad snRNAs of *M. exilis* indicate that the U12-type features found in *G. lamblia* are more likely a result of convergent snRNA evolution. None-the-less, the similarities observed between *G. lamblia* snRNAs and those of the minor spliceosome along with other similarities in splicing including close proximity of intronic BPS and 3' splice sites in diplomonads, raise interesting questions about the

ways in which the splicing mechanisms in these species may compare to those of the minor spliceosome.

### ***2.3.7 Reduced Giardia telomerase RNAs highlight convergent evolution of TBE features***

Like the *Giardia* U3 homologs, the *Giardia* TERs are significantly shorter than the TERs found in most eukaryotic species, placing them alongside the ciliate TERs as the shortest ever described. Both *Giardia* and ciliate TERs consist essentially of only the TBE, STE, pseudoknot, and template with little to no additional sequence. A peculiar feature of the proposed *Giardia* TER is the short three base-pair pseudoknot. The ciliate pseudoknot is closest in size, generally consisting of four base-pairs, but a TER has been described from *Tetrahymena paravorax* that also forms a three base-pair pseudoknot (Figure 2.8D) (McCormick-Graham and Romero 1995). The *Tetrahymena* pseudoknot is important for catalytic activity but requires the presence of telomerase RNP proteins to correctly form (Mihalusova et al. 2011). Three base-pairs may therefore be sufficient to form the TER pseudoknot in *Giardia*, but additional protein factors could be required to help form and/or stabilize the structure. A single example of a TER lacking a pseudoknot entirely has also been described in the trypanosomes (Gupta, et al. 2013; Podlevsky et al. 2016). The presence of a putative pseudoknot in *Giardia* TERs mean the absence of a pseudoknot in trypanosomes is most likely to be secondary loss rather than an ancestral feature as previously suggested. It will be necessary to determine if the *Giardia* pseudoknot does form *in vivo* and if this is found in other metamonad TERs to better resolve its evolutionary history.

The similarity of the TBE in *Giardia* spp. to the vertebrate and angiosperm TERs (Figure 2.8) is intriguing as it suggests this mechanism of regulating the extension of

telomeric repeats arose independently at least three times during eukaryotic evolution. The similarity is particularly striking in *Giardia* and angiosperms where the precise distance between the 5' most nucleotide of the template and the TBE element is conserved. A tendency to utilize this type of TBE mechanism cannot be explained by the reduction in size of the TERs as the vertebrate and angiosperm RNAs are much larger than in *Giardia*, and other small TERs use the more common template adjacent stem-loop TBE mechanism. Intriguingly, while most ciliates utilize a stem loop II TBE, *T. paravorax* and a number of other related ciliates including *Euplotes* species also lack the stem loop II TBE element (Lingner et al. 1994). Analysis of the *Giardia* TBE and associated proteins will be key in determining how closely the regulatory mechanism resembles telomerase RNPs lacking Stem II TBE elements in other species. This will hopefully shed light on possible reasons for the recurring evolution of this type TBE.

Presence of the *Giardia* specific 3' end processing motif in our TER candidates also adds to the already diverse set of biogenesis pathways described for telomerase RNAs. Metazoan TERs contain an H/ACA snoRNA/scaRNA domain and mature using the snoRNA/scaRNA processing pathway. Ciliate TERs are RNA pol III transcripts and terminate transcription with a poly-U stretch to produce mature 3' ends. Trypanosome TERs have been predicted to share the C/D snoRNA processing pathway, and many fungi use the snRNA biogenesis pathway for maturation where they are bound by the septameric Sm complex (Podlevsky and Chen 2016). It is intriguing that like in these other groups, the *Giardia* TERs have co-opted a processing pathway used to mature other ncRNA classes. As *Giardia* appears to be the only metamonad genus with a conserved ncRNA processing motif, other metamonad TERs likely utilize yet another pathway for biogenesis.

## 2.4 Conclusions

The findings presented here highlight the impact of comparative -omics on our understanding of widespread eukaryotic features and the novel innovations that have arisen during evolution under a diverse range of selection pressures. The *Giardia* U3 and telomerase RNAs are among the smallest identified in any eukaryote and comparison to other species, including other metamonads, shows this to be the result of a reduction in RNA sequence length and compaction or elimination of structural features. These RNAs show an intriguing parallel between the reduction in protein content and size, and the length of the RNAs in RNP complexes in organisms with compact genomes, refining the list of potential core RNP components and features. Retention of select snoRNA species between *G. muris* and *Gu. theta* indicates the important role of the targeted modifications in ribosome function, again uncovered through comparative analysis. The highly diverse diplomonad snRNAs are an example of how features that are generally considered to be highly conserved or even thought to be functionally critical in eukaryotes when examining a limited number of model organisms may not be required across all species.

The discovery of a *Giardia* RNase P-snoRNA diRNP represents the only fusion (hybrid) RNase P RNP ever described and reveals some of the first evidence for snoRNA-guided tRNA modification in eukaryotes. This complex is one of many lineage-specific innovations that have been uncovered through the study of diverse protist species and further analysis will likely lead to a deeper understanding of these RNA classes and their cellular roles.

## 2.5 Materials and Methods

### 2.5.1 RNA extraction and sequencing

The protocol for *G. muris* RNA extraction was adapted from Xu et al. 2020. Briefly, *Giardia muris* Roberts Thomson trophozoites were harvested from the small intestines of three experimentally infected C57 mice (day 7 of infection). The cells were lysed in TRIzol® and total RNA was extracted by the standard protocol. The RNA was used to prepare a strand-specific sequencing library using the TruSeq® Small RNA Sample Prep Kit. The stranded small RNA library was sequenced at GATC Biotech (Konstanz, Germany) with the Illumina HiSeq 2000 system, yielding 171,734,000 reads of 50 bp.

### 2.5.2 Annotation of ncRNAs

We performed BLASTN searches with -task blastn (which is more sensitive for distant matches) of the 40 *G. lamblia* ncRNAs against the *G. muris* genome to search for conserved homologs. This revealed 10 ncRNAs with E value <0.1 that did not overlap other genes. Alignment of the 10 ncRNAs using MUSCLE v3.8.31 showed a clear motif on the 3' end, CCTTYNHTNAA, which is similar to what was found in *G. lamblia* (CCTTYNHHTHAA). This motif was used to search for other potential ncRNAs in *G. muris* using scan\_for\_matches (<http://blog.theseed.org/servers/2010/07/scan-for-matches.html>).

We then searched the genome for other potential ncRNA candidates using a collection of different methods. First, we performed BLASTN searches against miRBase. Next, cmscan from the Infernal software package (v1.1.2) was used to search the *G. muris* Robert Thompson genome against the Rfam database to look for RNAs highly conserved across eukaryotes (Nawrocki and Eddy 2013). Infernal was additionally used to search the genome for *G. lamblia* specific RNA homologs using CM models built from individual *G.*



*lamblia* ncRNAs with cmsearch. We predicted additional small RNAs based on our small RNA-Seq data using Shortstack v1.2.4. Sequences for all search and prediction results were combined following the rule that ncRNAs do not overlap (or overlap little) with other annotated genes or with themselves. We also utilized manual curation efforts to include potential ncRNA candidates containing features of various ncRNA classes. These approaches combined to generate 219 ncRNA candidates: 16 from similarity searches against the *G. lamblia* genome, 8 from Rfam searches, 17 from Shortstack predictions and 178 from motif searches. Sequence conservation for candidate ncRNAs with homologs in *G. lamblia* were assessed by pairwise alignments using ClustalOmega (<https://www.ebi.ac.uk/Tools/msa/clustalo/>) as a first step before assigning them to an RNA class. These strategies identified a large collection of C/D and H/ACA snoRNAs, the U2, U4, and U6 snRNAs, and RNase P RNA. New candidate RNAs with predicted classes were verified by inspection for conserved structural features and sequences elements (e.g. snoRNA box elements, ‘ACAGAGA’ and ‘AGC’ sequences in U6) and the ability to form intermolecular helices for U2/U6 and U4/U6.

The snoScan webserver (Lowe and Eddy 1999) was used to predict target sequences for candidate C/D snoRNAs in rRNA, tRNAs, snRNAs, and snoRNAs. snoRNA homology between *G. lamblia* and *G. muris* was assessed based on sequence and structural similarity between RNAs along with target site prediction and genomic synteny. Similar searches were performed using *G. lamblia* C/D snoRNAs targeting the same modified positions as predicted for the newly identified *G. muris* snoRNA homologs. We also searched for RNA targets using unclassified ncRNA candidates to look for novel interactions specific to *G. muris*.

Searches for the more divergent U1 and U5 snRNAs required additional efforts. The reverse complement of the extended 5'ss of *G. muris* intron sequences (GU[A/U]UGUG) was used to query the 216 identified ncRNAs and their upstream regions (+50 bp). A similar search was done to look for the conserved U5 loop I sequence 'UGCCUUUUACY'. Potential candidate snRNAs were manually inspected for conserved sequence elements, and MFOLD was used to predict secondary structures. Candidate U1 snRNAs were chosen based on their ability to base-pair with the 5'ss of *G. muris* introns and form the conserved 'cloverleaf' structure. We required that any U5 candidate contained a putative loop I sequence closed by formation of a helix of at least 6 base-pairs, allowing for bulges.

To search for divergent telomerase RNAs, we examined ncRNA candidates not yet assigned to a class in both *G. muris* and *G. lamblia* by individually folding them using MFOLD from the UNAFold webserver (<http://www.unafold.org/>) to look for conserved telomerase RNA secondary structure features. Likely candidates were manually inspected for sequences complementary to the 'TAGGG' *Giardia* telomeric repeat and potential structures corresponding to the conserved template boundary element (TBE), stem terminus element (STE) and the pseudoknot adjacent to the template sequence based on structures found in various eukaryotic lineages.

A final consensus sequence for the 3' processing motif was determined using the sequences from all identified ncRNA genes and the *trans*-spliced introns by generating a frequency plot (n = 40 sequences) using WebLogo software (Crooks et al. 2004).

### **2.5.3 Quantification of ncRNA candidates**

A genome annotation file for our 216 ncRNA candidates in gff3 format was generated. The annotation file was used to quantify reads mapping to each of our predicted ncRNA candidates using the featureCounts component of the Rsubread R package (v2.2.6) (Liao et al. 2019) (settings were default with the following exceptions: largestOverlap = TRUE, minOverlap = 5, countMultiMappingReads = TRUE, fraction = TRUE, strandSpecific = 1). Reads mapping to both strands were quantified using the same command but with default strandSpecific settings. These values were used to calculate percentage of reads mapping to the correct strand.

### **2.5.4 Assessing homologous snoRNA target sites**

18S and 28S ribosomal RNA sequences from *G. muris*, *Gu. theta* nucleomorph, *S. cerevisiae*, and human were aligned using MUSCLE webserver (<https://www.ebi.ac.uk/Tools/msa/muscle/>) and snoRNA target nucleotides were mapped onto the sequences. Modified nucleotides found at same position within the alignment for two species were considered homologous. Modifications were mapped onto the secondary structure of the human rRNAs downloaded from RNA Central (<https://rnacentral.org/>) and compared to 3D ribosome structures to assess positioning of modifications in functionally important regions of the ribosome.

### **2.5.5 Synteny analysis**

MUMmer (v3.23) was used to align the draft genomes of *G. muris* and *G. lamblia* (Kurtz et al. 2004). Promer was used to align the genome at the protein level. Show-coords was used to view a summary of all the alignments produced by Promer. MUMmer\_toolkit

was used to convert the Promer .coords files into .crunch format compatible with Artemis Comparison Tool (ACT). The synteny information between *G. muris* and *G. lamblia* displayed in ACT was used to manually evaluate if ncRNAs have conserved gene order.

#### **2.5.6 Bioinformatics identification of diplomonad U3 snoRNA homologs**

Genomes for *Giardia lamblia*, *Giardia muris*, *Spironucleus salmonicida*, and *Monocercomonodies exilis* were downloaded from Giardiadb release 46 (<https://giardiadb.org/giardiadb/>). The *Spironucleus vortens* draft genome was downloaded from the Joint Genome Institute (<https://jgi.doe.gov/>). Genomes for *Kipferlia bialata* and *Streblomastix strix* were obtained from NCBI. Covariance models (CMs) were generated using Infernal (version 1.1.2) software and U3 snoRNA or GlsR1 stockholm files downloaded from the rFam database (<https://rfam.xfam.org/>). CMs were used to search genomes using the cmsearch function with default settings to identify sequences with conserved features for the U3 snoRNA. Hits were manually inspected for conserved box elements and structural elements. Alignments were generated using ClustalOmega webserver (<https://www.ebi.ac.uk/Tools/msa/clustalo/>). U3-18S helix forming regions for metamonads were identified by aligning *S. cerevisiae* and metamonad U3 and 18S RNAs and manually inspecting for conserved sequences. The metamonad sequences that aligned to the regions of U3 and the 18S involved in forming helices I, II, III, V, and VI in *S. cerevisiae* were obtained and assessed for the ability to base-pair.

#### **2.5.7 3' rapid amplification of cDNA ends (RACE) of the RNase P-GlsR15 fusion transcript**

For 3' RACE experiments, poly A tails were added to approximately 70 ng of total *Giardia lamblia* WB C6 RNA using *E. coli* Poly A polymerase (1 mM ATP, 1 X Poly A

reaction buffer (NEB), 5 Units *E. coli* Poly A polymerase (NEB)). Reactions were incubated at 37°C for 30 minutes then inactivated with EDTA. Reverse transcriptase (RT) reactions for first strand cDNA synthesis were performed as 20 µL reactions containing: 2 µL of Poly A tailing reaction, 500 µM each dNTP, 40 pmol oP-94 poly dT reverse primer (Table A.4.1), 1X SSIV reaction buffer (Thermo Scientific), 5 mM DTT, 200 U SuperScript IV™ reverse transcriptase (Thermo Scientific) according to the manufacturer's instructions. A control reaction was also performed lacking the reverse transcriptase enzyme. PCR reactions were performed on tailed first strand cDNA (and minus RT controls) with Phusion polymerase (Thermo Scientific) using the oP-94 reverse primer and an RNase P specific forward primer (DM129) (Table A.4.1). 3' RACE amplified products were resolved on 3% agarose gels and stained with ethidium bromide to visualize DNA bands. PCR products for plus RT reactions were cloned into pJET1.2 vector, following the manufacturer's instructions and cloned DNA was then sequenced via Sanger sequencing (Psomagen Inc) to confirm the overall sequence, and the location of the mature 3' end of the RACE products.

## Chapter 3: Analysis of two Snu13p homologs and their associated complexes from the diplomonad *Giardia lamblia*

### 3.1 Introduction

As described in Chapter 1, C/D snoRNAs are important for modification and processing of rRNA. The intestinal parasite *Giardia lamblia* is the most well studied of the metamonads, and one of only a few in which ncRNAs have been examined. C/D snoRNAs were among the first classes of ncRNAs described in *G. lamblia*, predominantly identified based on the presence of the conserved sequence box elements (Yang, et al. 2005; Chen, et al. 2007). Many of these RNAs have predicted targets in rRNA, but only a handful have been experimentally validated as modified nucleotide sites. Additionally, the box elements of *G. lamblia* C/D snoRNAs are often less well conserved than in other species, making definitive classification difficult based on primary sequence alone. These early analyses also led to a number of important discoveries, including *G. lamblia* being the first species in which snoRNA fragments with apparent microRNA properties were described (Saraiya and Wang 2008).

Eukaryotic C/D snoRNAs associate with the core C/D snoRNP proteins Snu13p, Nop56, Nop58 and the SAM dependent methyltransferase fibrillarin. Snu13p functions as the primary RNA binding protein through recognition of the K-turn, a conserved RNA structure formed in C/D snoRNAs by interactions between the C and D box motifs. Snu13p is a member of the L7Ae protein family and in addition to C/D snoRNAs has been shown to bind a K-turn present in the U4 snRNA in all previously examined eukaryotes (Koonin, et al. 1994; Nottrott, et al. 1999; Watkins, et al. 2000). Other protein members of the family include the ribosomal proteins S30, S12, and L7A, the H/ACA snoRNP protein Nhp2p

(Henras et al. 1998; Watkins et al. 1998), and the RNase P protein Pop3/Rpp38 (Wu et al. 2018). The specialized U3 C/D snoRNA is also bound by Snu13p, both at its conventional C'/D box K-turn as well as the U3 specific B/C box K-turn. Snu13p utilizes unique sequence features present in the bound K-turn to differentiate between RNAs and dictate which additional proteins to recruit, therefore regulating which RNP complex is assembled. In yeast the C/D snoRNP proteins Nop56 and Nop58 are recruited to the C/D and C'/D' motifs respectively (Cahill, et al. 2002; Watkins et al. 2002). Prp31 is recruited when Snu13p is bound to the U4 snRNA of the spliceosome (Schultz et al. 2006), and in the U3 snoRNP the Snu13p bound to the B/C box recruits the U3-specific protein Rrp9 (Cléry et al. 2007). The ability of Snu13p to differentiate between complexes based on RNA features and recruit the appropriate complex-specific proteins is a critical step in the assembly of these RNPs.

To date only a single study examining a Snu13p homolog from *G. lamblia* has been published (Biswas et al. 2011). This work qualitatively demonstrated a weak binding to an exogenous C/D snoRNA and intriguingly also found that the *G. lamblia* Snu13p, but not mouse Snu13p, could be substituted for the archaeal L7Ae protein to reconstitute catalytically active C/D s(no)RNP complexes with the archaeal proteins Nop5 and fibrillarin. This work highlighted the novel protein binding properties of the *G. lamblia* Snu13p compared to other eukaryotes, but no additional analysis has been performed. Our recent description of an RNase P-snoRNA diRNP in *G. lamblia* (Chapter 2) also raised new questions about the potential role of Snu13p in this novel RNP complex. Finally, Rrp9 and Prp31 are apparently absent from the *G. lamblia* genome (Feng et al. 2013; Hudson, et al.

2019). As both are direct Snu13p binding partners their absence could significantly affect how Snu13p nucleates the formation of RNP complexes compared to other eukaryotes.

In this study, we detected two distinct Snu13p homologs in the *G. lamblia* genome and characterized their associated RNP complexes. This led to the first experimental analysis of small subunit processome proteins in a metamonad and validation of our predicted G1-U3 snoRNA (Chapter 2). *In vitro* analysis also found an important role for Nop56, Nop58, and fibrillarin in snoRNA binding in *G. lamblia*. Additionally, we obtained experimental evidence supporting the assembly of a *Giardia* RNase P–GlsR15 diRNP *in vivo*. RNA and protein co-precipitation data suggests that the *G. lamblia* U4 snRNP does not contain either Snu13p homolog, the first documented instance of a U4 lacking Snu13p in any examined eukaryote. Finally, we detected and analyzed three previously uncharacterized proteins, unique to *Giardia*, that appear to associate with the ribosome. This new information forms a much clearer picture of the unique properties of many key Snu13p containing ncRNPs in *G. lamblia* and highlights the diversity that exists for these complexes within eukaryotes.

## **3.2 Results**

### ***3.2.1 Diplomonads possess two distinct Snu13p homologs***

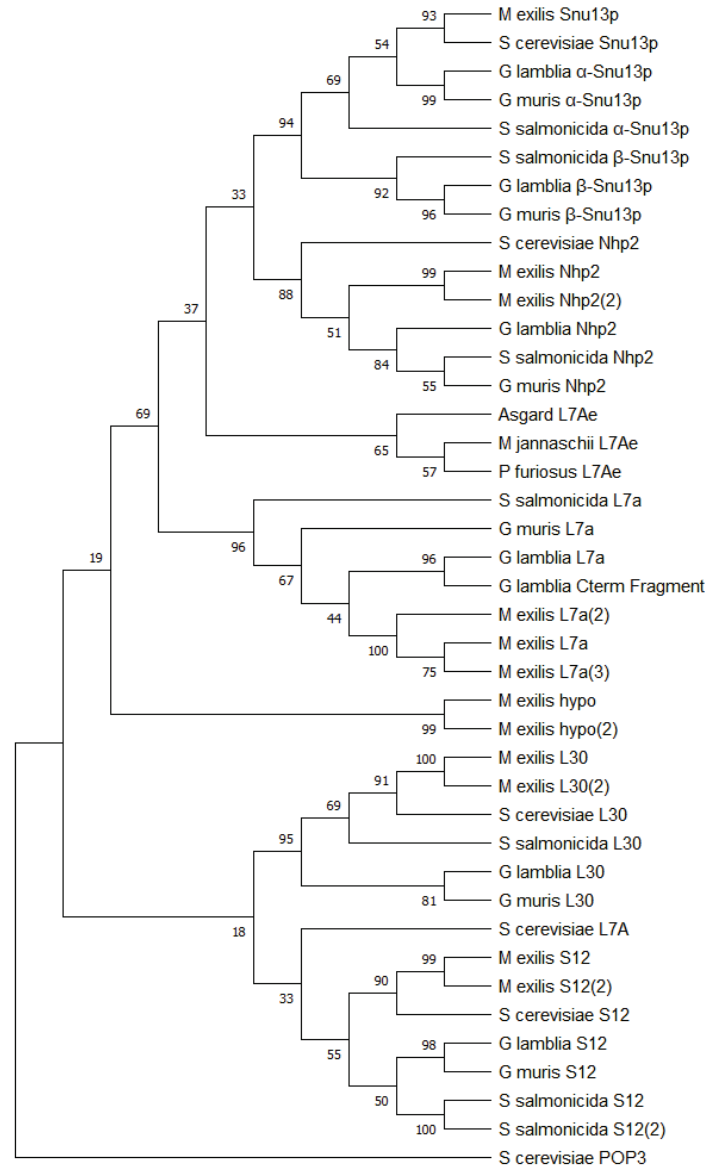
Homologs of the four core C/D snoRNP proteins have previously been described in *G. lamblia* (Narcisi et al. 1998; Russell et al. 2005); however, a search of the genome revealed an additional uncharacterized L7Ae family protein with features closely resembling Snu13p. To confirm the classification of this protein we collected sequences for all L7Ae domain-containing proteins from *G. lamblia* as well as two diplomonads *Giardia muris*, *Spironucleus salmonicida*, and an oxymonad, *Monocercomonoides exilis*.



Proteins were analyzed by sequence alignment to examine key conserved amino acid residues and phylogenetic analysis was performed using MEGAX. The uncharacterized *G. lamblia* L7Ae protein clustered with other Snu13p homologs, with strong bootstrap support for its placement (Figure 3.1A and B). The presence of a second distinct Snu13p homolog is conserved in the two other diplomonad species examined, but not for the oxymonad *M. exilis* (Figure 3.1A and A.2.1)). We now designate the previously described homolog as  $\alpha$ -Snu13p and newly identified homolog as  $\beta$ -Snu13p.

The diplomonad Snu13p proteins also group together when performing a phylogenetic analysis of eukaryotic Snu13p and archaeal L7Ae proteins; a pattern suggesting that the second homolog was gained through a gene duplication event that occurred at some point within the diplomonad lineage rather than horizontal gene transfer (Figure 3.2). Sequence alignments revealed that the diplomonad homologs differ at several residues which are otherwise well-conserved in eukaryotes (Figure A.2.2). This includes a string of four “signature” Snu13p/L7Ae residues (VSRP in many eukaryotes) that are important for RNA binding, located in loop 9 of the proteins structure (Gagnon et al. 2010). All six diplomonad Snu13p homologs deviate from the conserved eukaryotic loop 9 sequence by at least one residue and at another normally highly conserved upstream glutamine (Q34 in humans). Retention of two distinct Snu13p homologs in these otherwise highly streamlined genomes suggests important functional roles for both proteins. The sequence variation in key regions of both Snu13p homologs could suggest that they have unique RNA binding properties and may potentially be part of different (novel) RNP complexes.

**A**

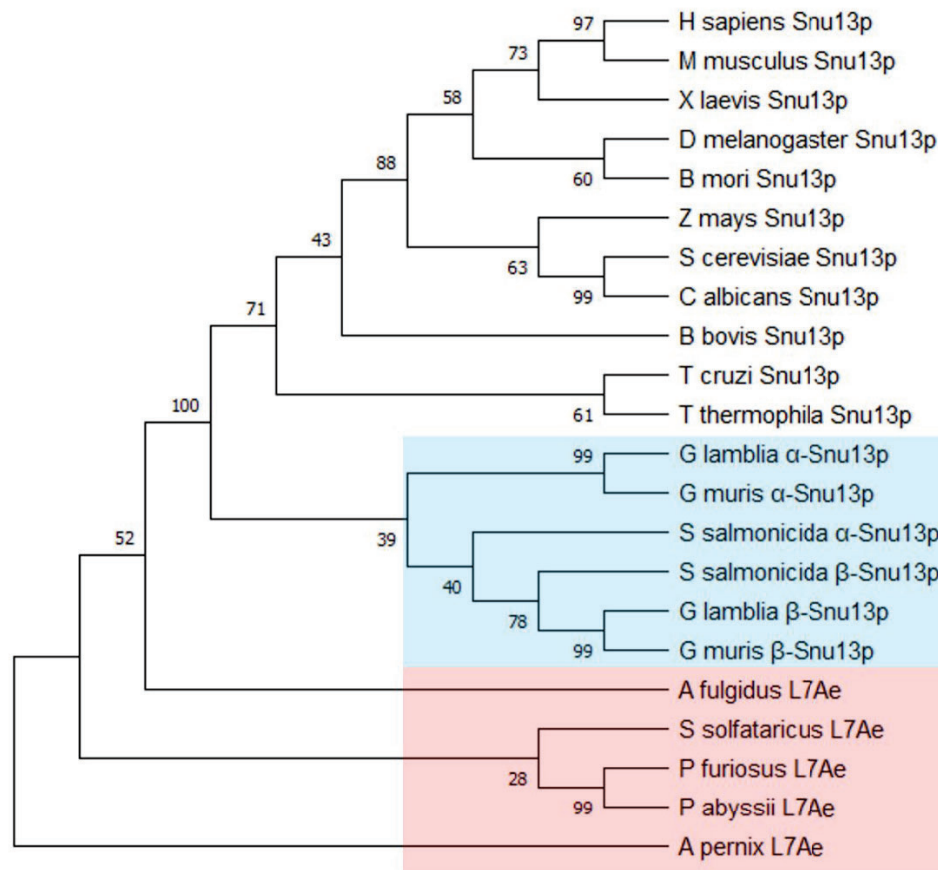


**B**

G. lamblia $\alpha$ Snu13p	MQIDPRAIPFANEELSLELLNLVKHGASLQAIKRGANEALKQVNRGKAEIVIIAADADPI	60
G. lamblia $\beta$ Snu13p	-MPDARAVPLASEAQSKRIYELVDLAKNSRSISRGMNEVTKALNKGKARLVVLSADALPL	59
	* **:*:*.* * .: :*. . . :*:.* **.* * :*:***.***:*** **	
G. lamblia $\alpha$ Snu13p	EIVLHLPLACEDKGVYPYVFIGSKNALGRACNVSVPTIVASIGKHD---ALGN---VVAEI	114
G. lamblia $\beta$ Snu13p	ELVLHLPEVCEDKGIAYIIFVPSRQELGRSVGISRQAVAVAIAKAPRQGTALDDKLNIFLTE	119
	*:***** .*****: *:*: *:*: *:*: .:* :*:*:.* **.: .:	
G. lamblia $\alpha$ Snu13p	VGKVEALV	122
G. lamblia $\beta$ Snu13p	LGH-----	122
	:*:	

**Figure 3.1. Diplomonads possess two distinct Snu13p homologs.**

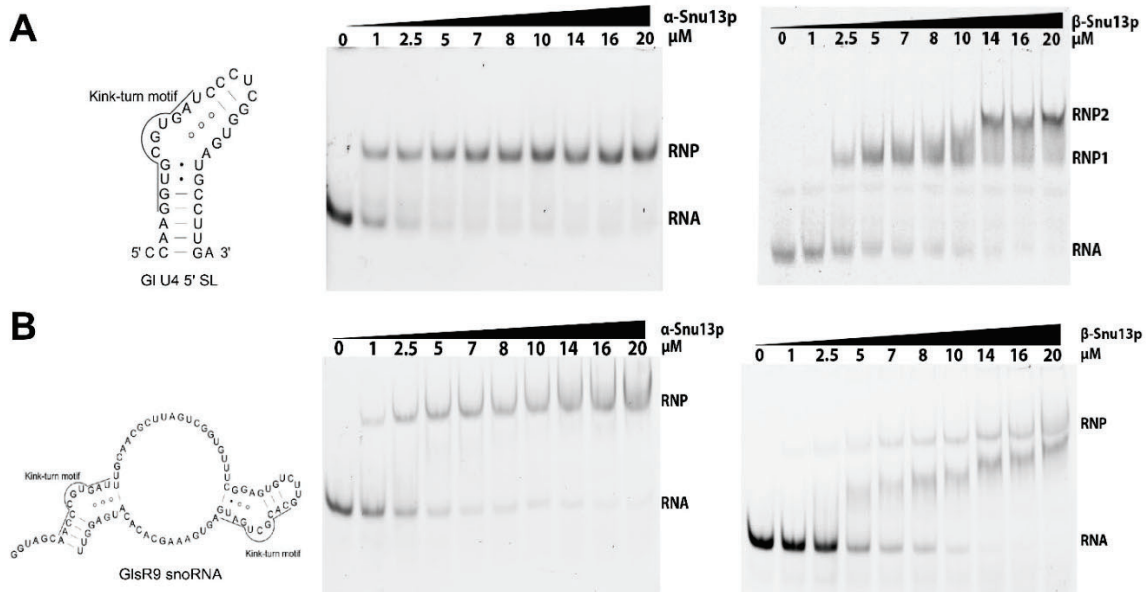
**Figure 3.1. Diplomonads possess two distinct Snu13p homologs.** (A) A maximum likelihood phylogenetic tree of L7Ae domain-containing proteins present in the translated proteomes of the metamonads *G. lamblia*, *G. muris*, *S. salmonicida*, and *M. exilis*. *S. cerevisiae* proteins are also included as well-characterized eukaryotic homologs, along with several archaeal L7Ae proteins. Numbers at nodes indicate bootstrap values. (B) Pairwise sequence alignment of  $\alpha$ -Snu13p and  $\beta$ -Snu13p homologs from *G. lamblia* generated using ClustalOmega, where \* represents identical residues, : are residues with very similar biochemical properties and . indicates somewhat similar biochemical properties. *G. lamblia* = *Giardia lamblia*, *G. muris* = *Giardia muris*, *S. salmonicida* = *Spironucleus salmonicida*, *M. exilis* = *Monocercomonoides exilis*, *M. jannaschii* = *Methanocaldococcus jannaschii*, *P. furiosus* = *Pyrococcus furiosus*, Asgard = Asgard group archaeon from marine sediment metagenome, *S. cerevisiae* = *Saccharomyces cerevisiae*.



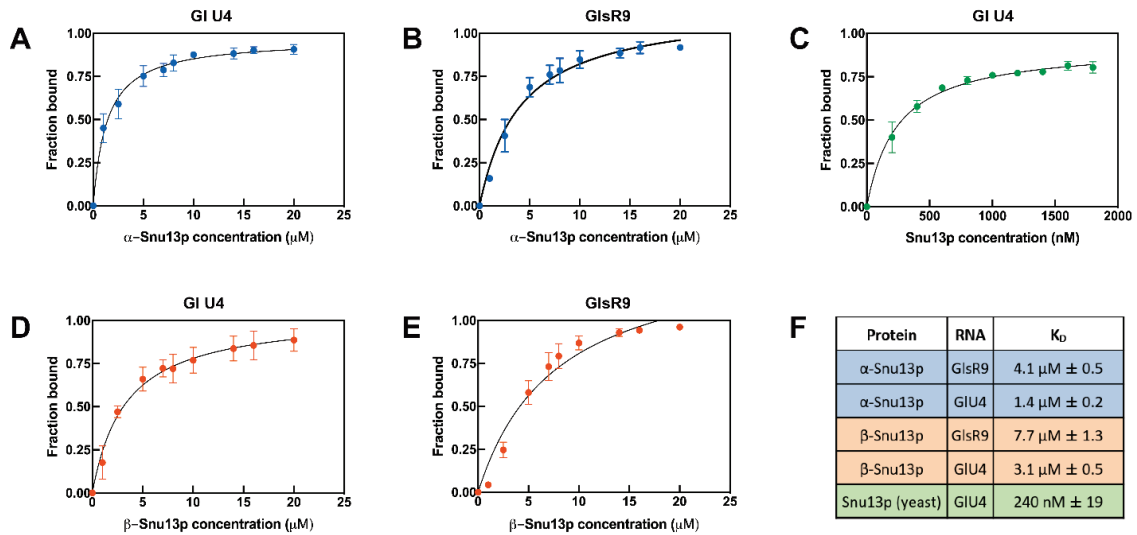
**Figure 3.2. Diplomonad Snu13p homologs form a clade within in a phylogenetic tree.** A maximum likelihood phylogenetic tree was generated using eukaryotic Snu13p homologs and archaeal L7Ae homologs based on species previously examined in Biswas et al. 2011, with the addition of newly identified diplomonad Snu13p homologs. Bootstrap values are indicated. Diplomonad homologs are boxed in blue and archaeal L7Ae homologs in red.

### 3.2.2 *G. lamblia* Snu13p homologs bind K-turns more weakly than other eukaryotic homologs

The sequence variation at key RNA binding amino acid residues in the *G. lamblia* Snu13p homologs prompted us to examine each protein's RNA binding affinities. Using Electrophoretic Mobility Shift Assays (EMSA), we assessed binding to the 5' stem loop of the *G. lamblia* U4 snRNA (G1 U4 SL) and a C/D snoRNA (GlsR9) (Figure 3.3A and B). Both these RNA classes contain K-turns and are bound by Snu13p homologs in other eukaryotes (Watkins, et al. 2000). We determined dissociation constants ( $K_D$ ) for  $\alpha$ -Snu13p binding of G1 U4 SL and GlsR9 to be  $1.4 \pm 0.2 \mu\text{M}$  and  $4.1 \pm 0.5 \mu\text{M}$  respectively versus  $3.1 \pm 0.5 \mu\text{M}$  and  $7.7 \pm 1.3 \mu\text{M}$  for  $\beta$ -Snu13p, showing marginally tighter binding by  $\alpha$ -Snu13p to both RNAs (Figure 3.4). Comparing EMSAs for the two homologs also revealed a second "supershifted" complex for  $\beta$ -Snu13p when binding G1 U4 SL at concentrations above  $10 \mu\text{M}$  that does not appear for  $\alpha$ -Snu13p (Figure 3.3A). This shift likely represents dimerization of the  $\beta$ -Snu13p protein as the G1 U4 SL RNA contains a single K-turn motif and is only 32 nucleotides in length making it unlikely that a second protein could bind to the RNA directly. Dimerization of only  $\beta$ -Snu13p suggests the two homologs differ in their ability to form unique protein-protein interactions.  $\beta$ -Snu13p also produced two bands per lane in EMSA when binding GlsR9. One of these bands likely represents binding of  $\beta$ -Snu13p to the K-turn formed by the internal C' and D' elements and not protein dimerization as it can be observed even at the lowest binding concentrations. This is further supported by the absence of a second band in lanes for  $\beta$ -Snu13p binding the *G. lamblia* snoRNA GlsR5, which cannot form a second K-turn (Figure A.2.3). The absence of such bands in the  $\alpha$ -Snu13p GlsR9 EMSAs suggest unique binding preferences do exist for the two homologs.



**Figure 3.3. EMSA analysis of *G. lamblia*  $\alpha$ -Snu13p and  $\beta$ -Snu13p binding to the *G. lamblia* U4 5' SL or GlsR9 C/D snoRNA.** Predicted secondary structures for the GI U4 (A) and GlsR9 (B) RNAs are depicted (left) with the K-turn structural region outlined on one strand. Binding assays representative of those used for  $K_D$  calculations for  $\alpha$ -Snu13p and  $\beta$ -Snu13p binding affinities for each RNA are shown. RNA indicates free RNA, RNP indicates a protein-RNA complex. Each lane contains 100 fmol of labeled GI U4 5' SL (A) or GlsR9 snoRNA (B) with a different concentration of protein (0 – 20  $\mu$ M).



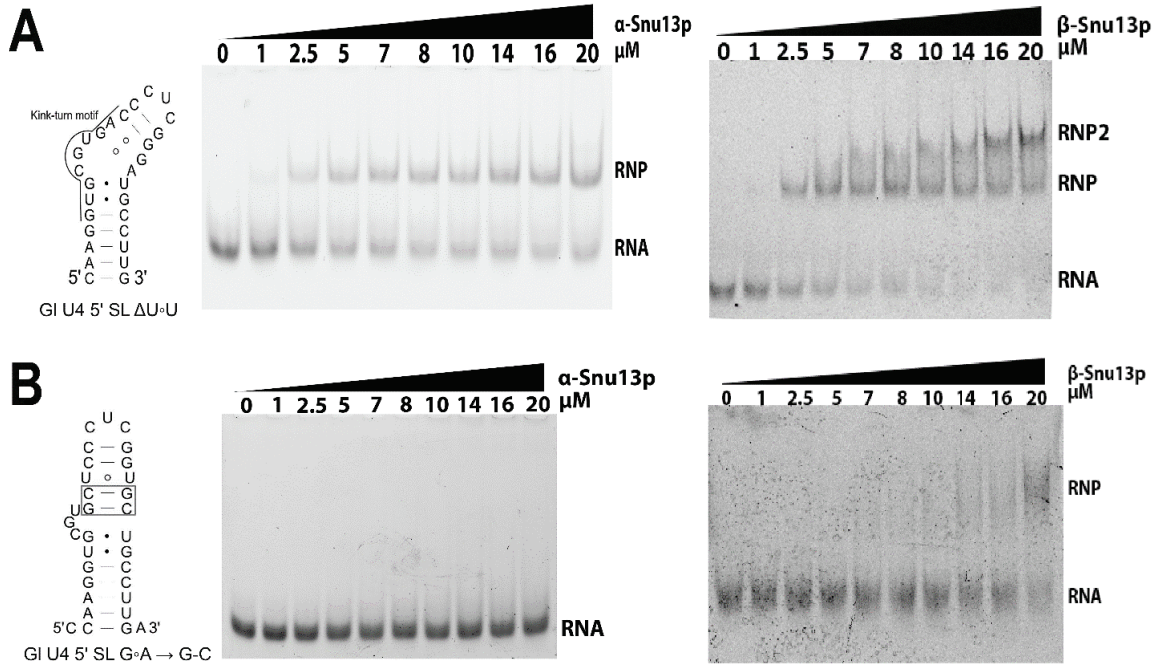
**Figure 3.4. *G. lamblia* Snu13p homologs bind with relatively low affinity to K-turn containing RNAs.** Binding curves generated in Prism using ImageJ analysis of EMSAs for  $\alpha$ -Snu13p (A and B) in blue,  $\beta$ -Snu13p (D and E) in red, binding to the GI U4 and GlsR9 RNAs respectively. *S. cerevisiae* Snu13p binding to GI U4 is shown in green in (C). (F) A table of the dissociation constants calculated from the curves using Prism (v9.1.0). Binding experiments were performed in triplicate.

The  $K_D$  values for the *G. lamblia* homologs are significantly higher than those determined for K-turn binding by any other eukaryotic Snu13p homolog (Chagot et al. 2019). To ensure neither our particular binding conditions nor the fluorescent RNA labels adversely affect the formation of protein-RNA interactions we analyzed binding of the well-studied *S. cerevisiae* Snu13p protein to our G1 U4 SL RNA. From this analysis we calculated a  $K_D$  of  $240 \pm 19$  nM for *S. cerevisiae* Snu13p, approximately five to ten-fold tighter binding than observed with the *G. lamblia* homologs, and consistent with previously determined dissociation constants for yeast Snu13p binding to K-turn containing RNAs (Figure 3.4C) (Chagot, et al. 2019).

We further probed K-turn binding by the *G. lamblia* Snu13p homologs by examining their ability to bind specific sequence variants of the G1 U4 SL RNA. Binding of both proteins is essentially completely abolished when performing binding assays with a G1 U4 SL RNA in which the tandem G $\circ$ A sheared pairs were replaced with canonical G-C pairs (U4 GC) to prevent K-turn formation (Figure 3.5B). Therefore, while RNA binding is relatively weak, both homologs do require K-turn formation for binding. The *G. lamblia* U4 K-turn also contains a non-canonical U $\circ$ U pair directly following the G $\circ$ A sheared pairs in the non-canonical stem. This feature is commonly found in C/D snoRNAs but is essentially absent from U4 K-turns throughout eukaryotes as an additional base-pair in this region can disrupt Prp31 recruitment to U4 (McPhee et al. 2014). We tested if this unusual non-canonical U $\circ$ U pair in G1 U4 is important for binding by the *Giardia* homologs (Figure 3.5A). EMSAs using a variant G1 U4 SL lacking the U $\circ$ U pair (G1 U4  $\Delta$ U $\circ$ U) showed no clear difference in binding compared to wild-type G1 U4 SL for either Snu13p homolog, indicating the U $\circ$ U pair is dispensable for binding, though it appears  $\alpha$ -Snu13p cannot



completely bind all the RNA at any tested concentration (Figure 3.5A). This is consistent with our finding that the U $\circ$ U pair is absent from the *G. muris* U4 snRNA (Chapter 2), which also contains the two Snu13p homologs.



**Figure 3.5. EMSA analysis of *G. lamblia*  $\alpha$ -Snu13p and  $\beta$ -Snu13p binding to variant GI U4 K-turn motifs.** Predicted secondary structures for the  $\Delta$ U $\circ$ U (**A**) and G $\circ$ A to G-C (**B**) variant GI U4 SL RNAs used for EMSAs are shown on the left. The tandem G $\circ$ A to G-C changed positions are boxed in (**B**). Representative EMSAs showing the binding of  $\alpha$ -Snu13p and  $\beta$ -Snu13p to each RNA are shown. RNA indicates free RNA, RNP indicates a protein-RNA complex. Each lane contains either the GI U4 5' SL  $\Delta$ U $\circ$ U (**A**) or GI U4 SL G $\circ$ A to G-C variant RNAs with a different concentration of protein (0 – 20  $\mu$ M).

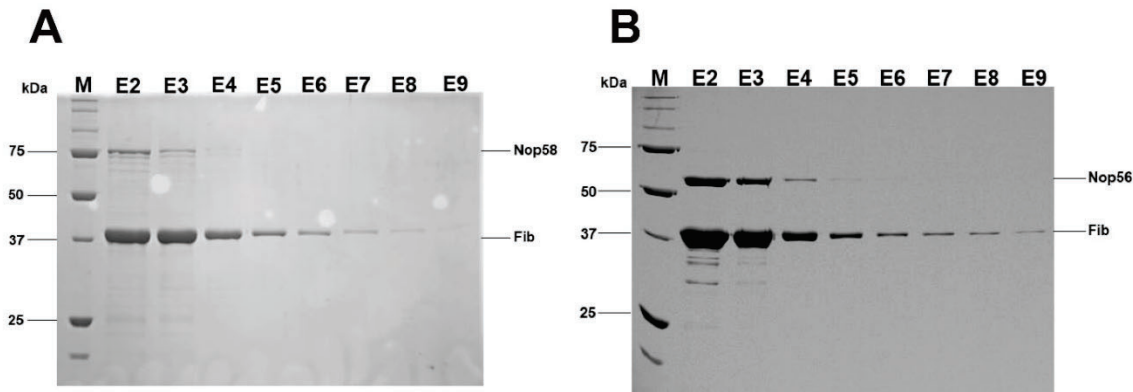
### 3.2.3 Protein-protein interactions and RNA binding properties of *G. lamblia* Nop56, Nop58 and fibrillarin

The comparatively weak binding of the *G. lamblia* Snu13p homologs to C/D snoRNAs led us to hypothesize that the three other core C/D snoRNP proteins (Nop56, Nop58, and fibrillarin) could play an enhanced role in RNA recognition and binding. We successfully purified recombinant *G. lamblia* fibrillarin from *E. coli*, but Nop56 and Nop58

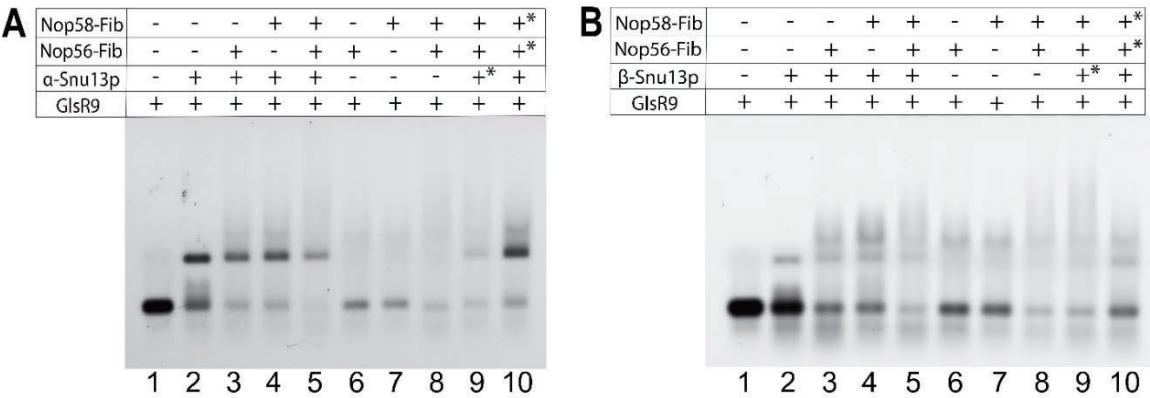
could not be stably expressed and purified. In archaea Nop5 and fibrillarin form a stable duplex *in vitro*; this led us to attempt co-expression of the two *G. lamblia* Nop proteins with fibrillarin to help increase Nop stability (Gagnon et al. 2012). We found that co-expression of either Nop56 or Nop58 with His-tagged fibrillarin in the presence of equimolar concentrations of glutamine and arginine (Golovanov et al. 2004) allowed for co-purification of the proteins as stable complexes (Figure 3.6). Analysis of RNA binding by these complexes using EMSA revealed that the Nop56-fibrillarin and Nop58-fibrillarin complexes can bind GlsR9 and GlsR5 C/D snoRNAs, while fibrillarin alone does not bind RNA at any examined protein concentration (Figure 3.7A and B, Figure A.2.4A and B). Binding of GlsR9 by each of the Nop-fibrillarin complexes appears tighter than observed for either Snul3p homolog, as a large proportion (>75%) of the RNA is in complex at a concentration of 1.6  $\mu$ M total (800 nM of each Nop-fibrillarin duplex) (Figure A.2.5). Nop-fibrillarin bands tend to be broader than those for Snul3p, potentially as a result of the complexes being more prone to falling apart in the gel matrix during electrophoresis. Similar binding results are obtained when using each of the Nop-fibrillarin complexes individually or as a combination of both complexes (data not shown), indicating similar contribution of each Nop containing complex to RNA binding. Because the Nop proteins require the presence of fibrillarin for stable purification we could not assess their binding in its absence. As a result, we cannot conclude with certainty if RNA binding is a property of the duplex or the Nop proteins alone. We did not determine specific  $K_D$  values for these protein-RNA interactions as our purification of the duplexes results in excess fibrillarin rather than stoichiometric amounts of the two proteins. As fibrillarin does not appear to bind GlsR9 on its own at these concentrations our measurements likely overestimate the



amount of duplexed protein present and available for binding. Therefore, the true binding affinity of the complex to snoRNAs is likely tighter than would be predicted by this data.



**Figure 3.6. *G. lamblia* Nop56p and Nop58p co-purify with His-tagged fibrillarin.** SDS PAGE of  $\text{Ni}^{2+}$  affinity purified fibrillarin found to be co-purifying with Nop58 (A) or Nop56 (B), isolated from *E. coli* cells that co-expressed Nop58 + fibrillarin (A) or Nop56 + fibrillarin (B). Lanes labeled M are protein molecular weight markers, lanes labeled E are successive elution fractions obtained during the affinity chromatography. Gels were visualized by staining with Coomassie R-250 Brilliant Blue.



**Figure 3.7. Analysis of Nop56-fibrillarin and Nop58-fibrillarin binding Glr9 snoRNA *in vitro*.** Agarose gel EMSA analysis of *G. lamblia* core C/D snoRNP proteins binding to the Glr9 C/D snoRNA. EMSA for  $\alpha$ -Snu13p (A) and  $\beta$ -Snu13p (B) using co-purified Nop56-fibrillarin and Nop58-fibrillarin complex. + for the Snu13p homologs indicates that the protein is present at a concentration of 3  $\mu\text{M}$  (1  $\mu\text{M}$  for +\*). For the Nop-fibrillarin duplexes + indicates the duplex is present at a concentration of 1000 nM (600 nM for +\*). All lanes contain 100 fmol of labeled Glr9 snoRNA.

To assess binding specificity for the Nop-fibrillarin complexes, we performed additional EMSAs using Nop56, Nop58 and fibrillarin with G1 U4 SL and a fragment of the *G. lamblia* RNase P RNA containing a K-turn. Both these RNAs are components of RNPs not predicted to contain Nop56, Nop58, or fibrillarin. We observed no interaction between the Nop-fibrillarin complexes and G1 U4 SL and only a very weak (potential) interaction with the RNase P fragment (Figure A.2.4C and D). Again, fibrillarin alone did not bind either RNA (Figure A.2.4A and B). The complexes formed between the Nop-fibrillarin duplexes and C/D snoRNAs therefore appears to be specific.

We next tested for the ability of the *G. lamblia* core C/D snoRNP proteins to assemble complete RNP complexes via agarose EMSA, to resolve larger complexes. Nop56-fibrillarin or Nop58-fibrillarin bind GlsR9 to form a single discrete RNP band as expected from the results described above (Figure 3.7A and B, lanes 6 and 7). When either of the Snup13 homologs is included, we observe the same Nop-fibrillarin band plus a second faster running band corresponding to a Snul3p-GlsR9 complex (Figure 3.7A and B, lanes 2, 3 and 4). Intriguingly, when both Nop-fibrillarin duplexes and either Snul3p homolog are present, both of the previously observed RNP bands become significantly fainter and a diffuse smear can be observed above the Nop-fibrillarin band (Figure 3.7A and B, lane 5). This could indicate that a complex containing all four core proteins is forming but is unstable and dissociates during electrophoresis. However, reactions containing both Nop56-fibrillarin and Nop58-fibrillarin together in the absence of either Snul3p homolog also form a more diffuse, slower running smear (Figure 3.7A and B, lane 8). These results indicate that the two Nop-fibrillarin complexes are capable of directly interacting with each other in the presence of RNA to form a larger complex *in vitro* even

in the absence of Snu13p (though does not exclude association with Snu13p when present), or that one of each Nop-fibrillarin complexes can separately bind to the regions around the C/D and C'/D' elements of the RNA without stable contacts forming between the complexes. Finally, we performed a similar test with the two Nop-fibrillarin duplexes in the presence of both Snu13p homologs to determine if a complete C/D snoRNP requires both homologs to be present in order to form. We did not observe a difference in these tests from reactions including just a single Snu13p homolog (Figure A.2.6 lane 8). Purification tests using the core C/D proteins in the absence of RNA were also performed but did not indicate the formation of a protein-only complex containing all four proteins (data not shown). These data together indicate that it is unlikely these proteins form a stable C/D snoRNP complexes *in vitro* in the absence of additional *trans*-acting assembly factors, but subsets of the proteins may form more transient interactions, at least under these conditions.

### ***3.2.4 Genes for several C/D snoRNP assembly factors appear to be absent from diplomonad genomes***

Research over the last 10 years has identified a collection of protein factors involved in the assembly of C/D snoRNPs in both yeast and humans. These include the R2TP complex (also known as the PAQosome), a co-chaperone of Hsp90 made up of the proteins Rvb1p, Rvb2p, Tah1p (RPAP3), and Pih1p (PIH1) as well as Rsa1p (NUFIP) and the zinc finger HIT domain proteins Hit1p (ZNHIT3) and Bcd1p (ZNHIT6) (See Chapter 1 for a detailed description) (Baldini et al. 2021). The inability of the core C/D snoRNP proteins from *G. lamblia* to stably assemble independent of such factors prompted us to search available metamonad genomes for homologs of C/D snoRNP assembly proteins. Probable homologs for each protein can be found in the oxymonad species *Monocercomonoides*

*exilis*, while the parabasalid *Trichomonas vaginalis* is missing only Pih1p. The three examined diplomonad species appear to be missing homologs for both Pih1p and Hit1p (Table 1 and Supplemental file 2). Several potential candidates for Tah1p/RPAP3 were detected in the diplomonad species, including *G. lamblia*, and are predicated to have the same domain architecture as human RPAP3 as determined by analysis with Phyre2, but a single strong putative candidate could not be identified (Supplemental file 2). The most recently described C/D RNP assembly factor, NOPCHAP1, was not also detected in *G. lamblia* or other metamonad genomes by our analysis or others (Abel, et al. 2021). Finally, all examined metamonad genomes encode a NUFIP domain containing protein, but none of these appear to be strong Rsa1p/NUFIP candidates as they result in very weak matches to homologs of these proteins in other species in both BLASTp and domain searches. In addition, the diplomonad homologs are significantly shorter in length than homologs found in higher eukaryotes or other metamonads, indicating that if these are functional Rsa1p/NUFIP homologs they are highly divergent.

Rsa1p/NUFIP binds directly to Snul3p in an RNA-independent manner in both yeast and humans (Rothé, et al. 2013; Quinternet et al. 2016). Despite overall divergence from other eukaryotic homologs the putative *G. lamblia* Rsa1p/NUFIP homolog conserves the amino acid residues involved in facilitating the Snul3p-Rsa1p/NUFIP interaction (Figure A.2.6A). We therefore experimentally tested whether the unusual putative Rsa1p/NUFIP homolog is able to associate with either Snul3p homolog from *G. lamblia*. Co-expression and co-purification experiments similar to those used to detect the Rsa1p-Snul3p interaction in yeast did not detect an interaction between the *G. lamblia* Rsa1p/NUFIP homolog with either *G. lamblia* Snul3p (data not shown). Further analysis

via EMSA found that a C-terminally truncated NUFIP homolog containing only the putative Snu13p interacting domain was unable to form a complex with either Snu13p homolog bound to a snoRNA (Figure A.2.6A and 6B, lanes 5-7). Together these results show that, in contrast to yeast, the predicted Rsa1p/NUFIP homolog is not able to interact directly with the *G. lamblia* Snu13p homologs in the presence or absence of a C/D snoRNA. Therefore, if the *G. lamblia* Rsa1p/NUFIP does function in C/D snoRNP assembly, the mechanism differs significantly from that of the yeast and human systems.

**Table 3.1. Predicted presence of C/D snoRNP assembly factors in metamonad species.**

Species	Rvb1	Rvb2	PIH1	Tah1/RPAP3	Hit1p	Nufip	Bcd1p
<i>M. exilis</i>	+	+	?	+	+	?	+
<i>T. vaginalis</i>	+	+	-	+	+	?	+
<i>G. lamblia</i>	+	+	-	?	-	?	+
<i>G. muris</i>	+	+	-	?	-	?	+
<i>S. salmonicida</i>	+	+	-	?	-	?	+

+ convincing homolog

? inconclusive homolog

- no homolog detected

### ***3.2.5 C/D snoRNAs but not the U4 snRNA co-precipitate with both G. lamblia Snu13p homologs***

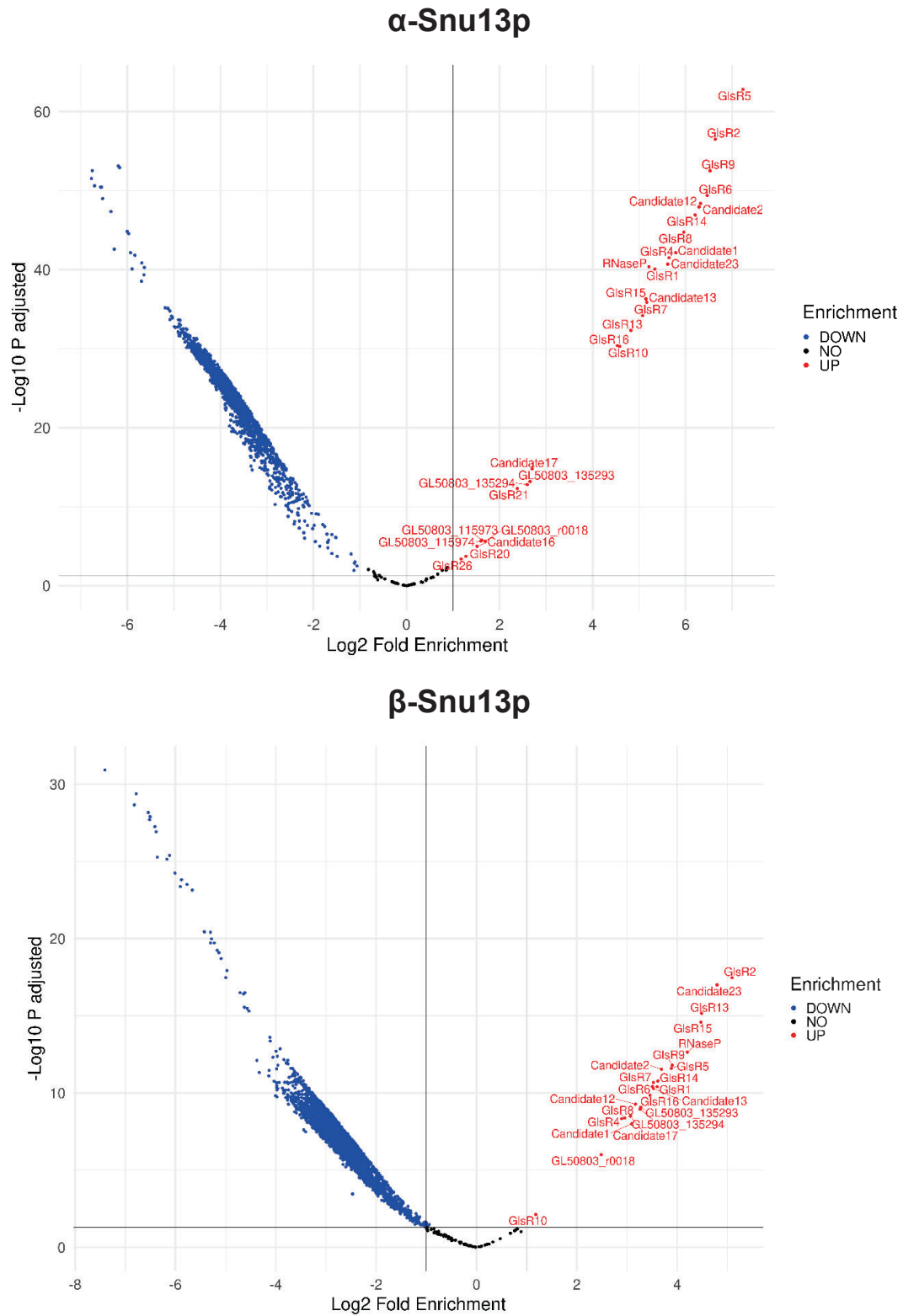
The specific but comparatively weak K-turn binding affinity of the *G. lamblia* Snu13p homologs *in vitro* led us to investigate what RNAs these proteins can stably associate with *in vivo*. RNA co-precipitations were performed using TAP tagged versions of both  $\alpha$ -Snu13p and  $\beta$ -Snu13p expressed in *G. lamblia* cells from recently developed expression vectors (Jerlström-Hultqvist et al. 2012). RNAs co-precipitated with each of the two homologs were used to construct Illumina sequencing libraries using TGIRT™-III, a thermostable reverse transcriptase capable of adapter template switching and processing through RNA structure. For each biological replicate a total RNA input control library was also constructed using total RNA extracted from soluble cell lysate taken from each sample

prior to co-precipitations experiments. Input control libraries were used to calculate enrichment values based on RNA abundance pre- and post- precipitation. Expressed protein control pull-downs using wild-type *G. lamblia* WB C6 cells showed essentially no nucleic acid is be purified from wild-type cells that are not expressing TAP tagged versions of the Snu13p homologs, as assessed by A260 readings and urea PAGE analysis (Figure A.2.7A). Comparison of input control libraries from the  $\alpha$ -Snu13p and  $\beta$ -Snu13p replicates show very strong correlation between RNA abundances indicating expression of the recombinant proteins did not significantly alter normal global RNA expression levels between strains (Pearson correlation analysis) (Figure A.2.7B).

As seen in Figure 3.8, 29 and 23 RNAs were significantly enriched in our  $\alpha$ -Snu13p and  $\beta$ -Snu13p libraries respectively, nearly all of which are C/D snoRNAs. All previously predicted *G. lamblia* C/D snoRNAs were enriched in purifications of both tagged homologs, confirming for the first time their association with the C/D snoRNP machinery. Universal C/D snoRNA enrichment for both homologs also discounts the possibility that individual C/D snoRNAs have strong preferential association with only one homolog. We additionally found that two previously predicted ncRNAs of unknown function, Candidate 12 and Candidate 17, were significantly enriched in our datasets. As described in Chapter 2, Candidate 17 is predicted to be the *G. lamblia* U3 snoRNA homolog (Gl-U3). Enrichment of Candidate 17 in both Snu13p homolog libraries strongly supports this classification, as U3 would be predicted to associate with the C/D snoRNP machinery. Closer inspection of the Candidate 12 sequence revealed putative C, D, C', and D' boxes (Figure A.2.8A). This RNA likely escaped previous classification as all four box elements diverge from the consensus sequences at one or more nucleotide position. We also

identified a potential modification target for Candidate 12 at position G664 of the LSU rRNA; a perfect 10 nucleotide base-pairing interaction between the rRNA and the guide region upstream of the D' box is predicted (Figure A.2.8B). Essentially all enriched RNAs were more highly enriched in the  $\alpha$ -Snu13p co-precipitations; however, differences in the purification conditions required for efficient isolation of the protein-RNA complexes could have impacted enrichment across experiments so we did not further analyze the (potential) differences between the two homologs.

A small number of RNAs were enriched in the  $\alpha$ -Snu13p but not the  $\beta$ -Snu13p libraries (Figure 3.8 and Supplemental file 2). These are mostly H/ACA snoRNAs, all of which are enriched just above the threshold values and have relatively low TPM/CPM abundances in both input control and experimental libraries compared to other enriched RNAs. This enrichment is likely due to indirect purification of H/ACA snoRNPs bound to pre-ribosomes which co-purify with Snu13p-associated complexes. Additionally, several predicted mRNAs were detected as enriched, but closer inspection of these RNA-seq reads found that they are likely erroneously annotated regions near rRNA operon ends rather than genuine unique mRNA coding regions (data not shown).



**Figure 3.8. C/D snoRNAs and RNase P are significantly enriched in co-precipitations with  $\alpha$ -Snu13p and  $\beta$ -Snu13p.**



**Figure 3.8. C/D snoRNAs and RNase P are significantly enriched in co-precipitations with  $\alpha$ -Snu13p and  $\beta$ -Snu13p.** Volcano plots for RNA-seq data from co-precipitation experiments analyzed in edgeR, plotting Log2 fold enrichment (Co-precipitated library/input RNA library) versus  $-\log_{10} P$  adjusted values (FDR). Each figure represents three replicates of paired co-precipitated RNA and input RNA control libraries. Vertical black lines indicate the Log2 Fold enrichment thresholds of 1 and -1. The horizontal black line marks the  $-\log_{10} P$  adjusted threshold set at  $P = 0.05$ . Significantly enriched RNAs are labeled and in red. Blue dots are significantly unenriched. Black dots indicate no significant change between input and co-precipitated libraries.

Surprisingly absent from the list of enriched RNAs was the U4 snRNA, which is bound by Snu13p homologs in all other examined eukaryotes and was observed to interact with both *G. lamblia* homologs *in vitro* (EMSAs described above). While still present, U4 abundance was significantly reduced in co-precipitation RNA-seq libraries for both Snu13p homologs compared to input control libraries and found at similar levels to other non-enriched, non-K-turn containing snRNAs and other ncRNA classes (Supplemental file 2). This result strongly suggests that neither Snu13p homolog is a component of the U4 snRNP in *G. lamblia*.

### **3.2.6 Ribosome processing complexes co-precipitate with *G. lamblia* Snu13p homologs**

We next used our tagged Snu13p homologs to perform protein co-precipitation experiments to further compare features of their associated complexes. Many *G. lamblia* proteins remain unannotated, therefore during our mass spectrometry analysis we attempted to annotate proteins lacking predicted identities. Co-precipitated proteins were assigned to one of seven categories based on evidence of association with a particular complex in other, more well studied, species (Table 3.2 and Supplemental file 2). Remaining proteins were recorded in an “other” category if they; 1) belong to a cellular complex that falls outside of the other six categories and showed limited representation in our data (e.g. one or two

proteins of a >20 protein complex were detected) or 2) were not successfully annotated as a homolog of a described protein.

In agreement with our RNA-seq analysis, both homologs clearly associate strongly with the core C/D snoRNP proteins, which are the strongest hits in the mass spectrometry data. More than 50 ribosomal proteins also consistently co-precipitated with each homolog, likely due to association of pre-ribosomes with precipitated snoRNPs (Table 3.2). We did not detect any of the bioinformatically predicated *G. lamblia* C/D snoRNP assembly factors in the pull-downs for either homolog. This result indicates that either the interactions are too transient to be detected by our methods or these proteins do not interact with the *G. lamblia* Snu13p homologs *in vivo*. This is particularly noteworthy for Rsa1p/NUFIP which is a direct binding partner of Snu13p in other species and is consistent with our lack of evidence for Snu13p-Rsa1p/NUFIP association *in vitro*. A relatively large number of “other” proteins were also co-precipitated in at least single replicates, especially with  $\alpha$ -Snu13p. More stringent purification conditions were tested (higher salt and detergent concentrations) to reduce potential spurious or non-specific interactions but these either significantly destabilized the core C/D snoRNP association or removed essentially all indirect interactions (data not shown). Also consistent with the absence of U4 from our RNA-seq data, no spliceosomal proteins were detected in any replicates for either Snu13p homolog.

**Table 3.2. Categorization of proteins detected in at least two replicates of protein co-precipitation experiments.**

Complex	$\alpha$ -Snu13p	$\beta$ -Snu13p	S4	Ucp1	Ucp2	Ucp3	Combined – Unique
Ribosome	66	50	63	65	63	56	68
C/D snoRNP	5	5	4	5	3	0	5
SSU Processome	34	18	8	8	3	0	35
Ribosome Biogenesis	17	3	5	4	3	0	18
Other	58	16	47	67	32	21	103
RNase P	2	0	0	0	0	0	2
Ucp	3	3	3	3	3	3	3
H/ACA snoRNP	2	0	2	2	2	1	2
<b>Total</b>	<b>187</b>	<b>95</b>	<b>132</b>	<b>154</b>	<b>109</b>	<b>81</b>	<b>236</b>

“Combined – Unique” column indicates the number of non-redundant proteins belonging to each complex (row) found in at least one of the six protein co-precipitation datasets

The detection of Gl U3 in our RNA-seq data prompted us to search our protein co-precipitation datasets for components of the small subunit (SSU) processome. Few SSU processome proteins are annotated in the *G. lamblia* genome database but searches using yeast SSU processome proteins revealed homologs for many of these are present in the *G. lamblia* genome. In total we detected 58 putative SSU processome proteins encoded in the *G. lamblia* genome including components of all the core sub-complexes (UtpA, UtpB, UtpC, Mpp10, and Utp7/Sof1/Utp14) and many additional factors, a combined 56 of which are present in our mass spectrometry data (Table 3.3 and Supplemental file 2). UtpA is the only core sub-complex that is lacking most constituents found in yeast, only containing three of seven proteins. When considering only the stable core processome complexes: 21/23 and 17/23 protein components present in the *G. lamblia* genome were detected in at least two of the mass spectrometry replicates for  $\alpha$ -Snu13p and  $\beta$ -Snu13p respectively. This indicates that we were able to consistently purify complete or near complete core SSU processomes rather than complex-intermediates or individual sub-complexes. As expected,

the additional transient factors were less consistently precipitated with only 13 and 1 additional protein(s) found in at least two replicates for  $\alpha$ -Snu13p and  $\beta$ -Snu13p. However, when considering individual replicates these values increase to 19 and 18 proteins, reflecting the more transient nature of the interaction between these proteins with the U3 snoRNP containing core processome. Combined data across pull-down experiments shows that 33 of the additional protein factors were detected, leaving only two SSU processome proteins that are encoded in the *G. lamblia* genome that were not detected by our co-precipitation experiments.

**Table 3.3. Protein composition of the SSU processome sub-complexes in *G. lamblia*.**

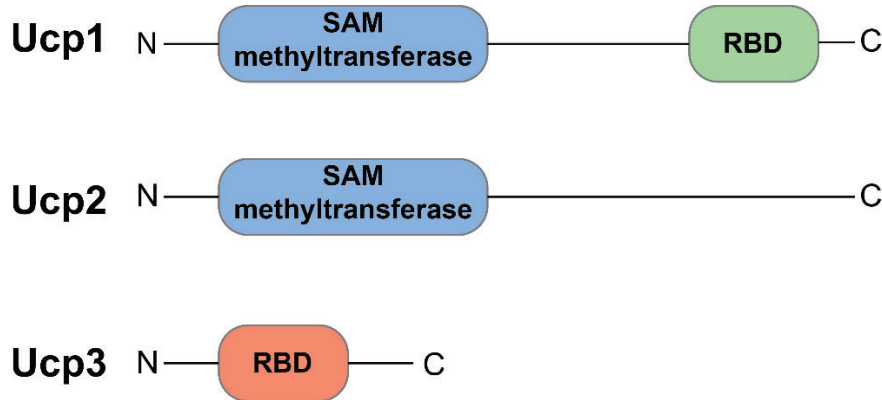
UtpA	UtpB	U3 snoRNP	Mpp10	Utp7/Sof1/Utp14	UtpC	Additional Factors		
Utp10	Utp13	Snu13p	Mpp10	Utp7	Rrp7	Utp11	Rcl1	Bud21
Utp17	Utp6	Fibrillarin	Imp3	Sof1	Utp22	Fcf2	Emg1	Lcp5
Utp15	Utp12	Nop56	Imp4	Utp14	Rrp36	Sas10	Utp23	Utp30
Utp4	Utp1	Nop58				Enp2	Mrd1	Rrt14
Utp5	Utp2	Rrp9				Rrp5	Nop9	Faf1
Utp8	Utp18					Krr1	Rrp8	Bud22
Utp9	Utp21					Bms1	Rrp12	Bfr2
						Kre33	Pno1	Efg1
						Nop14	Noc2	Fyv7
						Noc4	Cms1	Nop6
						Enp1	Utp20	Slx9
						Dbp4/8	Nob1	Sgd1
						Kri1	Esf2	
						Esf1	Dhr1	
						Utp3	Dhr2	
						Tsr1	Rok1	
						Nip7	Noc1	
						Utp24		

Present in pull downs
 Present in genome
 Absent from genome
 Ambiguous

## A

Ucp1	MPIIVKGAKLTDEFMASIKEVAGNVRELFSSRKFLSNVRETEKGVEIPCGYHKFFSAIQEK	60
Ucp2	MPHIIKGVKLTEEFTRTLKEIIGVDGFRSPYLANKARETDKGVEIPCGYGKLLAAVRQK	60
	***:*.***:* :*: .* : * : :.***:***** *:***:*	
Ucp1	HSNAELVSPDQIVLRQVFVERNKAFFARFKEQNADYLKSDKNALLVSSAPDGIIDSEKV	120
Ucp2	YPEADLLPLDQDEARKNCIDRNVSIFTKFKEYNADYLKTSKNVITVTSAPEGFMDPEKV	120
	:*:*: * : :*: :*:*** *****:*.***:*.***: * * *	
Ucp1	IAFLTKTFSKITPAKDHKFKITAKYQVLDAIREKVEADALKDGGSGASVKDRCAMMAIRFV	180
Ucp2	VAFLLTKLSKITPAQDHKFKITAKYQVLDAIRGRINAGALKDAGSSVQGRCATMGIRYV	180
	:*****:*****:*****:***** :*:*.***:*.***: * * *	
Ucp1	EGRAPQGRGTRVAVVVPVSDVEKIFNSLLADGYFERVPSIKFINLEERRKRAQERRERLDS	240
Ucp2	EGRPPQGRGTRVAVVVSSEIEKLFAALSADNLFERVPFIRFINPEKRRKHIQELRERVGS	240
	*** *****:***** *:***: * * * . *****:*** *:***: * * * : *	
Ucp1	AGDGSAAADSKKKPSEAAGRAKDKAGTDKADQSGSQSR--RGAQGDQSSSTKKRAADGKR	298
Ucp2	AGDGTLLGTNKKRRHQEAKGGVKKPGAKKATGIRVQLADDQKAAKPKKKQGGKQVQSQRER	300
	****: :.***: : * . * * :.*** * : * * :.*** : : *	
Ucp1	LPSFLKVANVPSTFDDIKGSLNENEHASILKAIAESHLQRRPRPDSSVISFFCTEENG	358
Ucp2	LPCLLTIAIGIPEALAFDDIKENLDKDEHADILKALAESRLRRPKQPNPSEVSFYCTVENG	360
	***:*.***:*.***:***** *:***:*****:*****:*****: * :***:*** *	
Ucp1	KILMEAFSNMDIDGSKLSVTLEEAR	383
Ucp2	KILRDAFGNMEINGAELRTTVSDVN	385
	*** :*.***:*.***: * :.***	

## B



**Figure 3.9. Ucp1-3 proteins detected in protein co-precipitations with *G. lamblia* Snul3p homologs and S4. (A)** Pairwise sequence alignment of newly identified and the highly-similar Ucp1 and Ucp2 proteins. \* represents identical amino acid residues, : are residues with very similar biochemical properties and . indicates somewhat similar biochemical properties. **(B)** Schematic of domains predicted by Phyre2 for the Ucp1-3 proteins detected in *G. lamblia* Snul3p co-precipitations. RNA binding domains (RBD) predicted from different PDB templates by Phyre2 are indicated in different colours (green or red). N and C terminal ends are indicated.

### ***3.2.7 Three abundant uncharacterized proteins co-precipitate with the *G. lamblia* Snu13p homologs***

In our Snu13p homolog protein co-precipitation data we consistently detected two uncharacterized proteins, given the interim names Ucp1 and Ucp2. BLASTp searches to the NCBI non-redundant protein database detected no homologs for either protein in any species outside of *Giardia*. Both proteins are present in all three mass spectrometry replicates for each Snu13p homolog and are found with peptide spectral match and peptide values near those of the core C/D snoRNP proteins. The proteins are nearly identical in size and alignments show significant sequence similarity, suggesting they may have arisen from a gene duplication event (Figure 3.9A). Using the SWISS-MODEL and Phyre2 web servers to predict potential protein domains found Ucp1 and Ucp2 both contain relatively high confidence matches to a variety of S-adenosyl-methionine (SAM)-dependent methyltransferase-like domains at their N-termini, with Ucp1 also predicted to contain a C-terminal RNA binding domain (RBD) (Figure 3.9B and Supplemental file 2). Previous mRNA sequencing analysis has shown that the Ucp1 and Ucp2 genes are also both very highly expressed (>98<sup>th</sup> percentile of expressed genes) (Franzén et al. 2013).

To further explore features of these proteins they were cloned into expression vectors and used for protein co-precipitation experiments. Both proteins co-precipitate nearly all the same ribosomal proteins as the Snu13p homologs along with a small number of proteins involved in ribosome biogenesis, however only Ucp1 consistently associated with all five C/D snoRNP protein components (Table 3.2 and Supplemental file 2). This indicates that the Ucp proteins may be ribosome-associated but likely do not interact directly with snoRNPs. Further support for this hypothesis was collected by co-

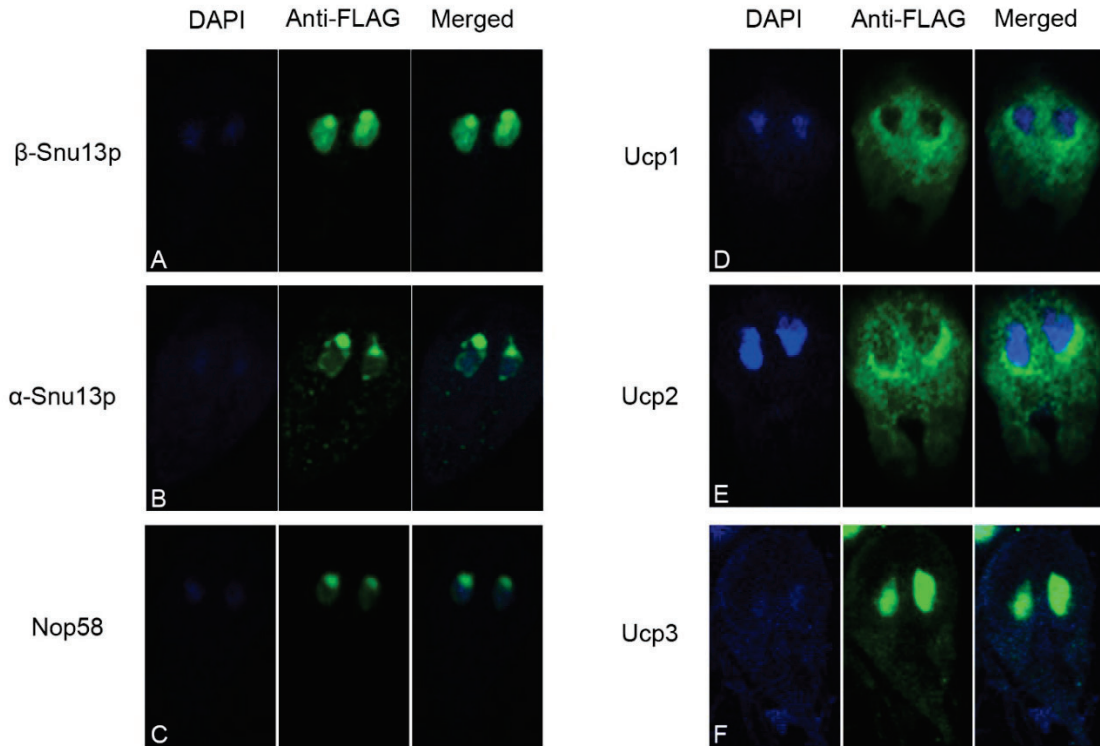
precipitations done using a TAP tagged version of the ribosomal protein S4, which resulted in a similar pattern of precipitated proteins to Ucp1 and Ucp2 (Table 3.2). A third uncharacterized protein, Ucp3, was found as a high abundance hit in the Ucp1 and Ucp2 pull-downs mass spectrometry data. Ucp3 is also present in all replicates for both Snu13p homologs and S4. Ucp3 is much smaller than Ucp1 or Ucp2, and domain predictions detected a single high confidence central domain that matches to various RNA binding domain structures (Figure 3.9B and Supplemental file 2). Co-precipitation of Ucp3 detected >50 ribosomal proteins, but essentially no ribosome biogenesis or rRNA modification complexes. Both Ucp1 and Ucp2 are seen in all three Ucp3 replicates, but C/D snoRNP proteins are not consistently observed across replicates, suggesting precipitation of Ucp3 with the Snu13p homologs was also indirect.

Searches of the *G. muris* genome found potential candidate homologs for all three Ucp proteins, but alignment of the Ucp candidates from the two *Giardia* species show significant sequence divergence (Figure A.2.9). Domain prediction analysis revealed identical organization for the *G. muris* and *G. lamblia* homologs (Supplemental file 2) which are of high confidence for *G. muris* Ucp2 and Ucp3, but prediction confidence scores for domains present in Ucp1 are significantly lower than in *G. lamblia*. The presence of the Ucp proteins in both *Giardia* species could indicate a conserved function for these proteins, however the significant sequence divergence observed between predicted orthologs, in particular Ucp1, may have also led to unique functions for the Ucp proteins in the two species.

### **3.2.8 Cellular localization of *G. lamblia* Snu13p homologs and Ucp proteins**

To determine if both Snu13p homologs localize to the *G. lamblia* nucleolus, as expected for C/D snoRNPs or if the two proteins differ in their cellular position, we analyzed protein localization by confocal immunofluorescence microscopy. An additional cell line expressing recombinant tagged Nop58 was also produced, which under normal conditions would be predicted to localize exclusively to the nucleolus with C/D snoRNPs. All three proteins are clearly concentrated at the apical side of both nuclei in individual cells, the previously described location of nucleolar formation in *G. lamblia* (Figure 3.10A-C and Supplemental file 3) (Jiménez-García, et al. 2008; Tian et al. 2010).  $\beta$ -Snu13p also shows a small degree of diffuse localization throughout the nucleus, but not in the cytoplasm. These findings agree with our RNA-seq and proteomics data indicating both homologs primarily function in ribosome maturation and that the addition of the tag does not impact nuclear import or protein localization.





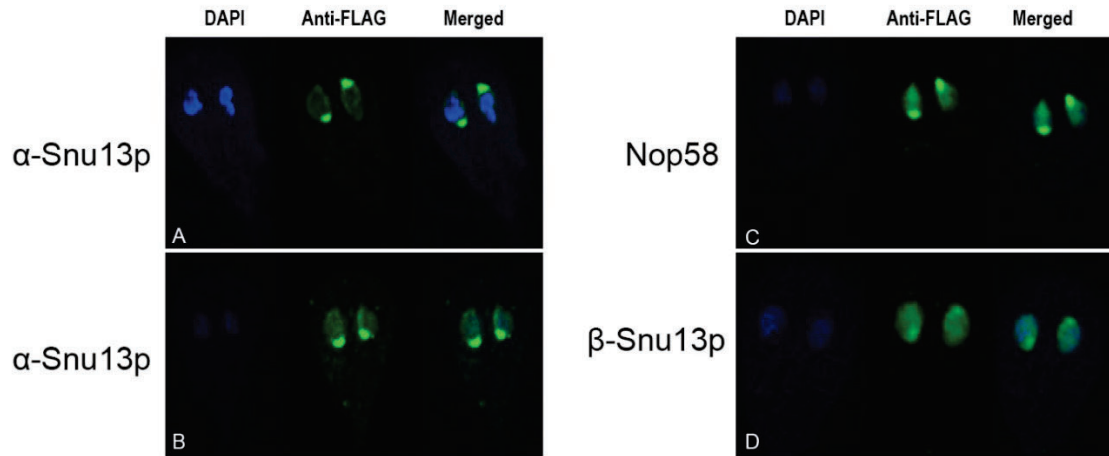
**Figure 3.10. Cellular localization of tagged *G. lamblia* proteins.** Representative confocal immunofluorescence microscopy images of *G. lamblia* expressing tagged  $\beta$ -Snu13p (A),  $\alpha$ -Snu13p (B), Nop58 (C), Ucp1 (D), Ucp2 (E), Ucp3 (F). DNA is stained with DAPI and appears blue. The protein of interest is visualized via an Alexa Fluor™ 488 fluorophore conjugated antibody and appears in green. Cells were visualized on an Olympus Fluoview FV1200 confocal microscope using the FV10-ASW 4.2 software at 60X magnification. Additional images of collections of cells for  $\alpha$ -Snu13p,  $\beta$ -Snu13p, Ucp1 and Ucp2 taken on a Cytation™ 5 Cell Imaging Multi-Mode Reader can be found in Supplementary File 2.

We also examined Ucp protein localization using this method. Ucp1 and Ucp2 are found at high concentrations at the outer basal side of the nucleus wrapping up each side, and in some cases surrounding the entire nucleus, but appear mostly absent from the nucleus itself (Figure 3.10D and E, Supplemental file 3). Both proteins are also detected in the cytoplasm becoming less concentrated further away from the nuclei. This pattern resembles localization of several protein classes previously found in the *G. lamblia* nuclear envelope (NE) and endoplasmic reticulum (ER) (Soltys et al. 1996; Elias et al. 2008; Touz

and Zamponi 2017). The similar unique localization pattern along with our proteomics data and their similar sequences suggest the two proteins are involved in the same complex. In contrast, Ucp3 is found exclusively in the nucleus with no apparent specific sub-nuclear localization (Figure 3.10F). Surprisingly, S4 was also found to localize throughout the nucleus rather than in the cytoplasm as expected for a ribosomal protein (Figure A.2.10A). Efforts to express an additional core ribosomal protein were unsuccessful and it is currently unclear why the majority of S4 would localize to the nucleus rather than the cytoplasm.

### ***3.2.9 G. lamblia nucleoli can form at the basal region of nuclei***

Expression of our tagged nucleolar proteins showed formation of discrete nucleoli at the apical side of both nuclei, as previous reported. However, during our analysis we found that in a smaller fraction of cases we could also observe instances where one (Figure 3.11A, C and D) or both nucleoli (Figure 3.11B) formed on the basal side of the nucleus. This phenomenon was observed for  $\alpha$ -Snu13p,  $\beta$ -Snu13p and Nop58, suggesting a general nucleolar phenomenon rather than a feature of any individual protein. Formation of asymmetric nuclei is intriguing as it supplies additional evidence for the existence of variation between the two nuclei, previously believed to be essentially identical. Basal nucleolar formation challenges the previously held notion that apical nucleoli formation is a universal feature of *G. lamblia* nuclei.

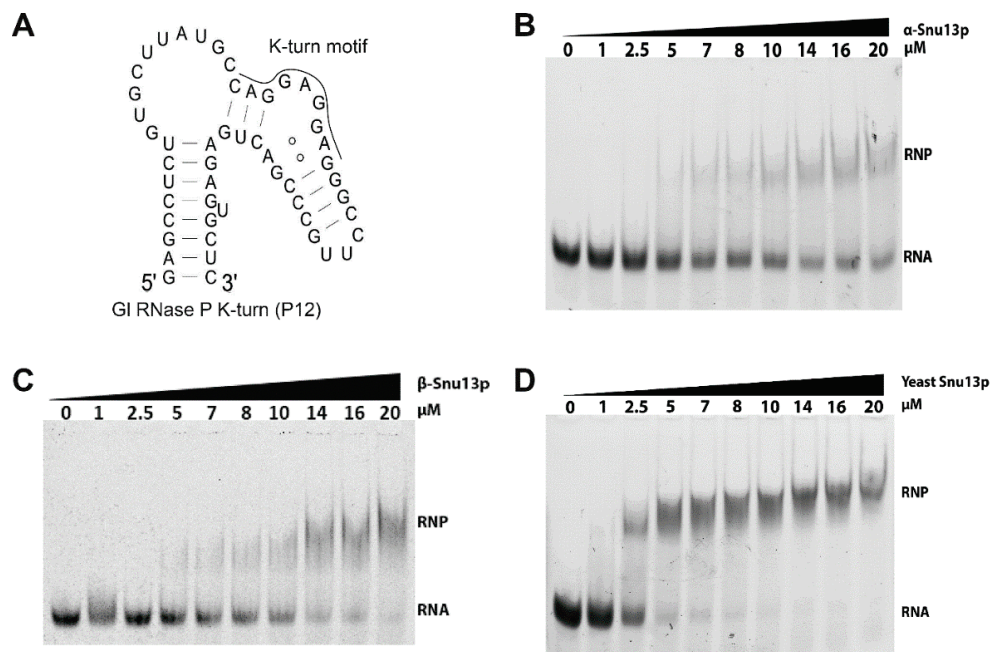


**Figure 3.11. *G. lamblia* nucleoli can localize to the basal side of one or both nuclei in individual cells.** Representative fluorescence confocal microscopy images of *G. lamblia* expressing tagged  $\alpha$ -Snu13p (**A and B**), Nop58 (**C**) and  $\beta$ -Snu13p (**D**) where one or both nucleoli are present on the basal side of the nucleus. DNA is stained with DAPI and appears blue. The protein of interest is visualized via an Alexa Fluor™ 488 fluorophore conjugated antibody and appears in green. Cells were visualized using an Olympus Fluoview FV1200 confocal microscope using the FV10-ASW 4.2 software at 60X magnification.

### 3.2.10 *G. lamblia* RNase P-GlsR15 RNA forms a hetero-diRNP

Our earlier observations (see Chapter 2) found that the RNase P RNA (RPR) from *Giardia* species is expressed and maintained as a single fused transcript with the C/D snoRNA GlsR15/GmsR15 which is predicted to target modification of tRNA<sup>Met</sup>. We hypothesized that this RNA likely forms a hetero-diRNP containing the protein components of both RNase P and C/D snoRNAs and is capable of both RNase P guided tRNA processing and snoRNA guided tRNA 2'-O-methylation. We used our proteomics, RNA-seq, and EMSA data to further probe this hypothesis. Analysis of our RNA-seq datasets showed that RPR is the only non-C/D snoRNA enriched in both the  $\alpha$ -Snu13p and  $\beta$ -Snu13p libraries and is among the most significantly enriched RNAs, indicating a strong, stable association with both homologs (Figure 3.8). Examination of the *G. lamblia* RPR sequence revealed a potential K-turn structure in a region of the RNA known as the P12

helix (Figure 3.12A). Archaeal RPR contains a K-turn in a homologous region that is bound by the protein L7Ae (Cho et al. 2010) and the human RPR P12 forms a K-turn bound in part by the protein Rpp38 (Wu, et al. 2018). Using EMSA we determined that both Snu13p homologs can weakly bind the P12 helix of *G. lamblia* RNase P RNA, but the interaction appears significantly less stable than the binding interaction with G1 U4 SL or GlsR9, especially for  $\alpha$ -Snu13p (Figure 3.12B and C). Consistent with our earlier binding analysis, yeast Snu13p bound the P12 K-turn at a much lower protein concentration than either *G. lamblia* homolog (Figure 3.12D), but significantly weaker than it bound G1 U4 SL and other K-turns. These results indicate that the P12 region of RPR is capable of K-turn formation but does not favorably interact with either Snu13p protein.



**Figure 3.12. EMSA analysis of Snu13p proteins binding the *G. lamblia* RNase P P12 K-turn.** (A) Predicted secondary structure for the P12 region of the *G. lamblia* RNase P RNA used for EMSAs with one strand of K-turn containing region indicated. Representative EMSAs for G1 RNase P P12 binding to *G. lamblia*  $\alpha$ -Snu13p (B),  $\beta$ -Snu13p (C), and *S. cerevisiae* Snu13p (D). RNA label indicates free RNA, RNP indicates a protein-RNA complex. Each lane contains 100 fmol of labeled G1 RNase P P12 K-turn RNA with a different concentration of protein (0 – 20  $\mu$ M).

Like other RNP complexes, RNase P proteins have not previously been annotated in the *G. lamblia* genome. Through BLASTp and domain searches we successfully identified putative homologs for the proteins Pop1, Pop4(Rpp29), Pop5, Rpr2(Rpp21), and Rpp1(Rpp30) in *G. lamblia*, but not the remaining four yeast proteins. Of the six tagged *G. lamblia* proteins used for co-precipitations in our study, only  $\alpha$ -Snu13p precipitated RNase P proteins. Two proteins, Pop1 and Rpp1, were found in all three  $\alpha$ -Snu13p replicates while Rpr2 and Pop5 were each found in a single replicate. The putative Pop4 homolog was not detected in any co-precipitation experiments. To assess if the absence of RNase P proteins from the  $\beta$ -Snu13p pull-downs was due to the shorter incubation time with the Strep-Tactin® resin compared with the 24-hour incubation used for  $\alpha$ -Snu13p we performed a single replicate of  $\beta$ -Snu13p with a 24-hour incubation. This trial contained a single RNase P protein, Pop1 (Table A.2.1 and Supplemental file 2). The 24-hour  $\beta$ -Snu13p replicate also contained dramatically more proteins overall than our other Snu13p pull-downs, in particular for proteins classified in the “other” category (145 “other” proteins compared to only 69 and 17 “other” proteins in the most abundant replicates of  $\alpha$ -Snu13p and  $\beta$ -Snu13p respectively). This indicates the 24-hour  $\beta$ -Snu13p replicate likely contains a larger number of non-specific protein interactions. We attempted to perform reciprocal co-precipitation experiments using the newly identified RNase P proteins to verify the RNP composition, but none were able to produce stable transfectants following insertion of the protein-coding regions into the *G. lamblia* expression vectors. These data support a preference for RNase P proteins to associate with  $\alpha$ -Snu13p but suggest that  $\beta$ -Snu13p likely also associates with the diRNP complex.

### 3.3 Discussion

#### 3.3.1 Insights into C/D snoRNP assembly in *G. lamblia*

The sequence variation observed in the diplomonad Snu13p homologs suggests that there is still considerable sequence evolution occurring in this lineage. This differs from the relatively high conservation of amino acid residues observed between other eukaryotic Snu13p proteins, eg. 78% identical residues between human and *Z. mays* homologs versus 39% between *G. lamblia* and *S. salmonicida*  $\alpha$ -Snu13p and 44% between human and *G. lamblia*  $\alpha$ -Snu13p (Figure A.2.11). Individual positions still often maintain residues with similar biochemical properties, eg. 67% for *G. lamblia* and *S. salmonicida*  $\alpha$ -Snu13p, and 74% for human and *G. lamblia*  $\alpha$ -Snu13p. This points towards more selective fine tuning of the proteins function rather than divergence in the absence of a selective pressure. Retention of the two homologs throughout the genomically reduced diplomonads indicates that they do each have a conserved cellular role. Previous studies found replacing individual loop 9 residues in the human Snu13p homolog (15.5K) with the corresponding archaeal L7Ae loop 9 residues increased  $K_D$  values anywhere from 3-fold to completely abolishing detectable binding (Gagnon, et al. 2010). 15.5K loop 9 and Q34 mutants in that study that match or closely resemble the residues found in the *G. lamblia* and other diplomonad homologs have  $K_D$  values that match well with our experimental observations for binding. It is therefore likely the variation at these two positions is a strong contributor to the weaker binding of the diplomonad homologs.

Previous experimental observations of archaeal Nop5 show it makes direct contacts with C/D sRNAs in archaea, in part through a highly conserved ALFR sequence which is important for efficient RNP formation (Ghalei et al. 2010). The ALFR sequence is

conserved in eukaryotic Nop56 and Nop58 and cross-linking experiments have found similar contacts for the eukaryotic homologs, but with asymmetric association of Nop58 with the C/D boxes and Nop56 with the C'/D' boxes (van Nues et al. 2011). Nop5-fibrillarin duplexes can also independently bind C'/D' regions of sRNAs in some archaeal species, though this interaction is much weaker than the L7Ae-RNA interaction (Tran et al. 2003). Our detection of direct and discriminative C/D snoRNA binding by the *G. lamblia* Nop56-fibrillarin and Nop58-fibrillarin complexes could be an evolutionary extension of these pre-existing contacts that have further evolved to compensate for the weaker binding of the *G. lamblia* Snu13p homologs.

Individual C/D snoRNP complexes in both archaea and eukaryotes contain at least two copies of L7Ae/Snu13p, one at the C/D boxes and one at the C'/D' boxes (Cahill, et al. 2002; Omer, et al. 2002; Szewczak, et al. 2002; Yu et al. 2018; Yang, et al. 2020) Our observations suggest that *G. lamblia* C/D snoRNPs do not require both Snu13p homologs be present in a single complex as co-precipitations detected high levels of the tagged Snu13p and other core C/D snoRNP proteins but much lower levels of the untagged Snu13p homolog. If both homologs were strictly required to form a single snoRNP complex, the untagged homolog would be expected to be present at similar levels as the two Nop proteins, as these proteins asymmetrically associate with C/D snoRNAs and are each present as single copies per RNA. Unfortunately, attempts to precipitate C/D snoRNPs using tagged recombinant Nop58 and fibrillarin both failed. fibrillarin did not stably express from transfected vectors and Nop58 is degraded following cell lysis under all tested purification conditions. Cumulatively our analysis show *G. lamblia* C/D snoRNPs can form



using either Snu13p homolog, and are significantly more reliant on the Nop-fibrillarin complexes for RNA recognition than other eukaryotes.

Both newly characterized snoRNAs, Candidate 12 and G1 U3, were previously predicted as ncRNAs of unknown function based on expression data and the presence of the ncRNA 3' processing motif (Chen, et al. 2007; Hudson, et al. 2012). In our Snu13p RNA co-precipitations we detected all currently predicted C/D snoRNAs including these two, but no additional putative C/D snoRNAs lacking the 3' end RNA processing motif. This likely means that all C/D snoRNAs in *G. lamblia* have now been identified and are processed by the same shared mechanism at their 3' ends.

### **3.3.2 C/D snoRNP assembly machinery in *G. lamblia***

Despite similarities between archaeal C/D sRNPs and the *G. lamblia* C/D snoRNPs our unsuccessful attempts to assemble a stable C/D snoRNP using only the core *G. lamblia* proteins suggests the complex still requires association with additional factors to properly assemble. Absence of detectable Rsa1p/NUFIP binding to either Snu13p in both our *in vivo* and *in vitro* experiments along with its divergent NUFIP domain sequence all cast doubt on a role for this protein in snoRNP assembly in *G. lamblia*. We attempted to further probe the C/D snoRNP assembly pathway using *in vivo* expression of our predicted Bcd1p homolog. Tagged Bcd1p could be stably expressed but attempts to purify it and associated proteins were unsuccessful under a variety of tested expression and purification conditions. Immunofluorescent localization of the protein found significant accumulation in the nucleus, but a strong signal was also observed diffusely throughout the cytoplasm, indicating the tagged protein may not properly localize (Figure A.2.10B).



The lack of any C/D snoRNP assembly factor homologs in our co-precipitation experiments and absence of several others from the *G. lamblia* genome entirely suggest that *G. lamblia* may utilize a distinct C/D snoRNP assembly pathway. In yeast, the assembly machinery is reliant on many protein-protein interactions between the assembly factors. The lack and/or size reduction of several of these factors in *G. lamblia* would likely require significant alteration to the overall structure of the assembly complexes relative to the machinery present in metazoans in order to remain functional (Massenet, et al. 2017; Yu, et al. 2018). Even highly conserved assembly factors like the *G. lamblia* Rvb1 and Rvb2 may not be involved in C/D assembly as they play a variety of different roles in most other eukaryotes and may be conserved in the *G. lamblia* genome for these other functions. *G. lamblia* may therefore utilize a distinct C/D snoRNP assembly pathway. However, we cannot rule out that their absence from our co-precipitation data is due to the transient nature of their interactions with the core C/D snoRNP proteins. It is possible *G. lamblia* possesses unique assembly factors, some of which could be present in our data; however, we did not detect any convincing evidence for such candidate proteins.

### ***3.3.3 G. lamblia Snu13p homologs do not interact with the spliceosomal U4 snRNA in vivo***

We initially hypothesized that retention of the two Snu13p homologs in *G. lamblia* was a consequence of subfunctionalization of the ancestral role of Snu13p in C/D snoRNPs and the U4 snRNP. The lack of U4 snRNA enrichment in our RNA co-precipitations and complete absence of spliceosomal components in protein co-precipitations with either *G. lamblia* Snu13p homolog do not support this hypothesis, instead strongly suggesting that neither homolog is part of the U4 snRNP in *G. lamblia*. Snu13p is the only known U4

snRNP protein in the *G. lamblia* genome, making it unlikely that the protein tag on either Snu13p was made inaccessible for binding to the purification resin by additional proteins, unless additional lineage specific U4 proteins have evolved (Hudson, et al. 2019). Low abundance of splicing complexes is also unlikely to explain the absence of splicing proteins in our data as we can detect other low abundance proteins like those of RNase P in our  $\alpha$ -Snu13p pull downs, indicating our experimental conditions and analysis were sufficiently sensitive.

In other species, the Nop domain containing protein Prp31 associates directly with Snu13p in the U4 complex and makes contacts with the U4 snRNA (Nguyen et al. 2016). Human Prp31 has also be found to significantly stabilize binding of the human 15.5K to a variant U4 RNA only weakly bound by 15.5K on its own (Schultz, et al. 2006). The absence of Prp31 from the *G. lamblia* genome means that unlike C/D snoRNPs there is no known secondary protein that could stabilize the relatively weak Snu13p-U4 interaction *in vivo*, potentially preventing complex formation from occurring. The absence of Prp31 from the *G. lamblia* and other diplomonad genomes may have also impacted the evolution of the Snu13p homologs in this lineage. Sequence and structural features of the U4 K-turn help dictate the selective recruitment of Nop56, Nop58, Prp31, or Rrp9 to the appropriate Snu13p-RNA complex (Schultz, et al. 2006; Liu et al. 2007). *Giardia* U4 snRNAs deviate from the sequence important for Prp31 recruitment to Snu13p bound U4 in humans. The loss of a need for a specific Prp31 recruitment signal in the U4 K-turn in the diplomonads allowed for both sequence variation in the RNA and a relaxation in RNA binding requirements for the Snu13p homologs, which do not need to be as discriminative, leading to their weaker K-turn binding affinity.

It is somewhat surprising that U4 in both *G. lamblia* and *G. muris* would retain the ability to form a K-turn structure if it is not bound by Snu13p considering the overall rapid evolution of their snRNAs (Chapter 2). A possible explanation which is consistent with our other data and model is that the association of Snu13p is transient but still important for correct folding of U4. L7Ae family proteins in some species can assist in K-turn formation, through an induced fit mechanism (Turner et al. 2005). Weak binding of *G. lamblia* U4 could mean Snu13p initially associates to facilitate formation of the K-turn, which in turn helps folding and association with the U6 snRNA but is not retained due to weak binding affinity and lack of Prp31 for stabilization. K-turns can also form in the absence of protein, especially in the presence of certain metal ions so we cannot rule out that Snu13p is also not required for folding. Regardless, we predict the *G. lamblia* U4 snRNA is likely still a constituent of the spliceosome based on the conservation of extensive base-pairing between the U4 and the U6 snRNAs (Hudson, et al. 2012). Even if a lack of U4 proteins impacts splicing efficiency this could be tolerated due to the small number of spliceosomal introns present in *G. lamblia*.

#### ***3.3.4 The G. lamblia SSU processome is similar in composition to other eukaryotes***

Our protein co-precipitation data is consistent with a previous bioinformatic survey of core components of SSU processome subcomplexes from across eukaryotes, including *G. lamblia* (Feng, et al. 2013), but importantly was successful in identifying additional proteins previously thought to be absent. The *G. lamblia* SSU processome contains many of the proteins found in yeast, the notable exception being the absence of several components of the UtpA subcomplex. Intriguingly, recent cryo-EM structures of the yeast processome found that many of the proteins of the UtpA and UtpB complexes are

evolutionarily related based on similarity in structure and domain architecture (Sun, et al. 2017). One possible explanation for the *G. lamblia* SSU processome tolerating these absences in UtpA but not other subcomplexes could then be that UtpB proteins are sufficiently similar to those of UtpA that they are capable of moonlighting in the UtpA complex to fill in for the absent proteins. It may also simply be that the missing UtpA proteins do not play as critical a role as other proteins in *G. lamblia*, allowing for their absence without significant negative impacts on ribosome biogenesis.

$\alpha$ -Snu13p protein co-precipitation data contained an Rrp9 homolog, a core component of the U3 snoRNP which was previously thought to be absent from the *G. lamblia* genome (Supplemental file 2) (Feng, et al. 2013). Nucleotides of the U3 B/C box determined to be required for Rrp9 recruitment to Snu13p in other eukaryotic species are also present in the diplomonad U3 snoRNAs. These B/C box nucleotides include a conserved G at the first position of the K-turn loop (L1) and a G-C that closes either the C or NC stem of the K-turn (Cléry, et al. 2007)(See also Chapter 2 for diplomonad U3 snoRNAs). A model for cooperative RNA binding by Snu13p and Rrp9 has been proposed in yeast. In the *G. lamblia* system, the contribution of Rrp9 to RNA binding could stabilize the interaction of Snu13p with the B/C box K-turn in a fashion similar to what we have proposed for Nop56 and Nop58 binding to the C/D boxes of modification guide snoRNAs. Rrp9 is present in all  $\alpha$ -Snu13p co-precipitation replicates but none of the  $\beta$ -Snu13p replicates. However, Rrp9 was detected in our 24-h incubation  $\beta$ -Snu13p co-precipitation data. Gl-U3 is enriched in both the  $\alpha$ -Snu13p and  $\beta$ -Snu13p libraries, and both homologs co-precipitate with the majority of other core SSU processome proteins. A potential model consistent with these data is the asymmetric association of  $\alpha$ -Snu13p with the U3 B/C K-

turn which recruits Rrp9, and association of  $\beta$ -Snu13p (or a second  $\alpha$ -Snu13p) with the U3 C'/D K-turn which recruits Nop58. This model can explain the enrichment of G1 U3 and SSU processome proteins in our co-precipitation experiments for both  $\alpha$ -Snu13p and  $\beta$ -Snu13p, and the less consistent co-precipitation of Rrp9 with  $\beta$ -Snu13p being due to a lack of direct association between the two proteins. Alternatively, lack of Rrp9 in  $\beta$ -Snu13p datasets could be a result of less stable association between the two proteins due to two sequences differences between Snu13p homologs. Both Snu13p homologs are likely components of the U3 snoRNP, but further work will be required to precisely determine the nature of their interaction with the two K-turn elements of the U3 snoRNA and recruitment of the newly identified Rrp9 homolog.

### ***3.3.5 Giardia specific proteins may be associated with ribosome function***

In our mass spectrometry datasets, Ucp1 is the only protein for which the largest classification category of co-precipitated proteins was “other”. Despite efforts to further annotate proteins in the “other” category most do not return BLAST hits other than to unannotated proteins in *Giardia* species or very weak generic domain matches. Ucp2 associates with fewer “other” proteins and most overlap with those detected with Ucp1, providing no obvious clues about the potential cellular activities involving these proteins. The work presented here provides a strong framework for understanding these proteins and hints at the potential for a novel protein complex made up of uniquely *Giardia* proteins but will require additional experimental investigation to better characterize.

While we can't make definitive conclusions about the roles of these proteins, we can speculate on plausible functions based on our current findings. For example, of the “other” proteins that could be annotated in these datasets a number are translation factors

(eg. EF1 and EF2) or nuclear export and quality control/ER associated proteins (eg. NEMF, ERP1, and Hsp70) (Supplemental file 2). Ucp1 and Ucp2 localize most strongly to the nuclear periphery, which matches the published localization of ER and NE associated proteins including members of the SNARE class of membrane associated proteins (eg. gQa4 and gQb5) and the ER protein chaperone BiP (Soltys, et al. 1996; Elias, et al. 2008; Touz and Zamponi 2017). This along with our prediction that Ucp1/Ucp2 associate with ribosomes suggests a possible role in ribosome localization or maturation. *G. lamblia* lack golgi and the ER takes over several roles in protein localization which could mean these cells require specialized proteins to help facilitate ribosome retention and interaction with this unique membrane. Alternatively, Ucp1/Ucp2 could themselves be specialized ribosomal proteins as the profiles of co-precipitated proteins for Ucp1, Ucp2 and S4 are strikingly similar (Table 3.2 and Supplemental file 2). Ucp1 could directly facilitate an interaction with the rRNA through its RBD and direct Ucp2 through protein-protein interactions. Recent publication of the cryo-EM structure of the *G. lamblia* ribosome did not feature the Ucp proteins, but all three are present in the mass spectrometry data of the ribosomes utilized for structure determination (Eiler et al. 2020), further supporting direct association with ribosomes.

Based on our protein co-precipitations Ucp3 also appears to associate with ribosomes, likely later during assembly as we detected no ribosome biogenesis factors in these experiments. Localization of Ucp3 throughout the nucleus matches with the unusual S4 localization we detected. Nuclear S4 localization could mean the tagged protein is predominantly excluded from mature ribosomes, however our tagged S4 is found in the densest fractions in glycerol gradient ultracentrifugation experiments and not in lighter

fractions, indicating it is incorporated into larger complexes (Figure A.2.12). Like Ucp3, S4 also co-precipitates relatively few ribosome biogenesis factors suggesting much of the protein is associated with mature ribosomes rather than assembly intermediates or being excluded. This may be the genuine localization for at least some *G. lamblia* ribosomes as they have not previously been localized, but this too will require further validation. As the Ucp3 consists predominately of a single RBD we also performed an initial investigation into the RNA binding properties of Ucp3. These tests were successfully used to co-precipitate Ucp3 associated RNAs. Urea PAGE analysis of Ucp3 associated RNAs shows an enrichment of RNAs between 200-600 nt in length, and retention of several large RNAs not observed in the Snu13p homolog co-precipitations (Figure A.2.13). Future RNA-seq analysis of these RNAs will help significantly in determining the specific role it plays in the cell and its association with ribosomes.

### ***3.3.6 G. lamblia nucleoli can be present at the apical or basal side of each nuclei***

Localization of nucleoli to the basal side of the nucleus in one or both nuclei of a *G. lamblia* cell could be the result of errors in the processes of nuclear division causing nuclei to be flipped in daughter cells. *G. lamblia* utilizes semi-open mitosis meaning the nuclear envelope does not completely disassemble during cell division (Sagolla et al. 2006). Additional work also suggests that nucleoli persist during this process and may even remain partially active (Lara-Martínez et al. 2016). Retention of intact nucleoli in a case where the nuclei was erroneously flipped during division would result in their basal localization as the nucleoli would not have the opportunity to reform at the usual apical side of the nucleus from diffuse nucleolar material following cytokinesis. Our consistent observation of these “flipped” nucleoli across cell lines expressing a variety of nucleolar proteins suggests

however, that this may represent a normal cellular alternative to apical nucleolar formation. These are the first reported cases of basal nucleolar formation in *G. lamblia* and identification of the mechanism through which basal nucleoli form and if this change has an impact on cellular processes will grant a better understanding of inter-nuclear diversity in *G. lamblia*, a topic which has garnered increased attention in recent years.

### **3.3.7 *G. lamblia* forms an RNase P-snoRNA diRNP in vivo**

Co-precipitation of the RPR with both *G. lamblia* Snu13p homologs and the presence of four core RNase P proteins in the  $\alpha$ -Snu13p mass spectrometry datasets add experimental support for our hypothesis (Chapter 2) that a composite RNase P-snoRNA diRNP forms on a single RNA from components of both RNPs in *Giardia*. The complement of RNase P proteins in *G. lamblia* is very similar to the archaeal RNase P complex, containing the same number of proteins (five) and only differing in that it is missing a Pop3 homolog (L7Ae in archaea) and contains a Pop1 homolog (absent in archaea) (Esakova and Krasilnikov 2010). This again adds to the list of *G. lamblia* features that either through ancestry or genomic reduction resemble archaeal features. Pop3/Rpp38 is a eukaryote specific L7Ae family protein which has been shown to bind the P12 K-turn of RPR in yeast and humans. Pop3/Rpp38 is absent from the *G. lamblia* genome which could support a hypothesis in which one of the Snu13p homologs fills this role. However, the very weak binding of the RPR P12 K-turn by both Snu13p homologs suggests that precipitation of the diRNP complex is more likely due to association of the homologs with the Glr15 C/D snoRNP domain. An additional protein could associate with Snu13p when bound to the RPR P12 and stabilize the interaction as we have proposed for Nop56/Nop58 and Rrp9 with C/D snoRNAs, but we currently have found no evidence for such an interaction. Of



note the *G. muris* RNase P does not contain the P12 K-turn but maintains the RNase P-snoRNA fused orientation and two Snu13p homologs, casting further doubt on a role for the Snu13p homologs in direct binding to this region of the RPR (Chapter 2). Finally, early tRNA maturation by RNase P occurs in the nucleolus in other eukaryotes, like yeast (Bertrand et al. 1998; Hopper et al. 2010). Localization of the Snu13p homologs to the *G. lamblia* nucleolus is consistent with a role in RNase P, assuming tRNA processing occurs similarly. This localization could also be a driving force in the emergence and evolution of the diRNP. Protein components of RNase P have been proposed to play a role in localization of the RNP (Esakova and Krasilnikov 2010). The absence of several of these proteins, including Pop3, could therefore be problematic for correct localization of RNase P. Presence of the snoRNP domain may therefore help facilitate proper localization of RNase P to the nucleolus for tRNA processing while also ensuring GlsR15 encounters its tRNA target rather than pre-rRNA.

Based on the enrichment of RPR and GlsR15 in the  $\beta$ -Snu13p RNA-seq data at levels similar to  $\alpha$ -Snu13p it seems highly likely this homolog also associates with the diRNP complex. The presence of Pop1 in the 24-hour  $\beta$ -Snu13p replicate could suggest that like we have proposed for the U3 snoRNP the Snu13p homologs associate asymmetrical with the K-turns of RPR and GlsR15. In this case, more consistent precipitation of RNase P proteins with  $\alpha$ -Snu13p would likely mean that this homolog more likely associates with the RPR K-turn while  $\beta$ -Snu13p exclusively associates with the GlsR15 domain. However, as described above our data is not entirely in line with this interpretation as it seems unlikely that either Snu13p homolog associates with the RPR K-turn. An alternative interpretation is that differences in RNase P protein co-precipitation is

due to differences in the purification conditions required for the two homologs. Again, in this case both homologs most likely associate with the GlsR15 K-turn exclusively. To help resolve this uncertainty, we attempted to perform reciprocal co-precipitations for proteomic analysis using four different RNase P proteins but were unable to generate stable transfectants. The consistent issues expressing proteins from this complex indicates overexpression of RNase P proteins or inclusion of tags on these proteins may be cytotoxic, perhaps due to interference with the tRNA processing pathway. Open questions remain regarding the exact composition of the complex and structure of the *G. lamblia* diRNP but the analyses presented here strongly support the formation of an RNase P-snoRNA diRNP in *G. lamblia*, making it the first RNase P fusion RNP ever reported.

### 3.4 Conclusions

Our proposed model requiring that additional protein factors be present to assist with K-turn binding in *G. lamblia* is unique among eukaryotes, but an extension of contacts already present early in the evolution of these RNPs. Absence of important C/D snoRNP assembly factors further highlight the presence of an unusual assembly pathway in *G. lamblia*. This model can explain both the tolerance for weak binding by Snu13p and the involvement in some RNP complexes but not others observed in other eukaryotes. Our findings support that the functions of the two Snu13p homologs are at least in part overlapping in *G. lamblia* but that they may differ in some regards including the manner of their association with the U3 and RNase P RNAs. Experimental evidence for the involvement of the Snu13p homologs in the RNase P-GlsR15 diRNP also show unique roles for Snu13p in *Giardia* that are not observed in other eukaryotes and reveal a unique tRNA processing complex. Additional distinct cellular roles for one or both Snu13p

homologs in *G. lamblia* may yet be uncovered and the sequence variation between Snu13p orthologs in diplomonads could reveal that their roles are variable even between species in this lineage. Finally, the identification of three abundant proteins (Ucp) that appear to associate with the ribosome but for which no homologs are found outside of *Giardia* could be important pieces in understanding the unusual *G. lamblia* protein transport systems and present yet another intriguing example of intriguing cellular innovation within protist lineages

### **3.5 Materials and Methods**

#### ***3.5.1 Culturing of Giardia lamblia***

*Giardia lamblia* cultures were grown in 22 mL glass screw cap tubes at 37°C in TYI-S-33 medium supplemented with bovine bile salts. Cultures tubes were placed at a 45° angle to increase cell adherence and grown to > 90% confluence. Every 3-4 days cells were subcultured into 50 mL glass screw cap tubes under the same conditions for experiments to increase yield. For cultures expressing recombinant proteins from either pAN or pAC vectors media was supplemented with 50 µg/mL of puromycin for selection.

#### ***3.5.2 Phylogenetic analysis of L7Ae domain proteins***

L7Ae domain containing proteins were predicted using the GenomeNet motif tool (<https://www.genome.jp/tools/motif/>) using translated proteomes downloaded from <http://www.giardiadb.org>. Sequences for proteins containing the requisite domains were downloaded and analyzed in MEGAX to produce a Maximum likelihood tree using an LG + G settings with 500 bootstraps. Sequences for Snu13p analysis were selected based on those included Biswas et al. 2011 and downloaded from GenBank. Phylogenetic analysis was performed using the same settings as above.

### ***3.5.3 Cloning of G. lamblia proteins into pAC and pAN expression vectors***

Protein coding genes were amplified from *G. lamblia* WB C6 (ATCC 30957) genomic DNA by PCR. PCRs were performed as 50 µL reactions in 1 x Phusion HF buffer, 0.2 mM each dNTPs, 20 pmol of forward and reverse oligonucleotide primers containing restriction digest sites (Table A.4.1), 100 ng *G. lamblia* genomic DNA and 1 U of Phusion® DNA polymerase (NEB). PCRs were typically run with an initial 98°C denaturing step of 3 min, followed by 35 cycles of denaturing at 98°C for 10 sec, annealing for 15 sec (at the melting temperature of the primer pairs), and extension at 72°C for 45 seconds, followed by a final 5 min extension at 72°C. Successful PCR reactions were cleaned up using the E.Z.N.A.® cycle pure kit (Omega Bio-tek) following the manufacturers instructions.

For cloning into pAN both the vector and gene insert PCRs were restriction digested with BamHI-HF and EcoRI (NEB) in a single reaction 50 µL reaction: 20 U BamHI-HF, 20 U EcoRI, 1 X NEB Buffer 4, and 1500 ng of pAN vector or 1000 ng of gene coding PCR. Restriction digests were incubated at 37°C for 1 h, then inactivated at 65°C. Vectors were treated with 10 U of Calf Intestinal Alkaline Phosphatase (NEB) for 1 h at 37°C then cleaned up with E.Z.N.A.® cycle pure kit. Inserts were ligated into pAN in 20 µL reactions with 3:1 insert:vector ratios in 1 X T4 ligase buffer (50 mM Tris-HCl, 10 mM MgCl<sub>2</sub>, 10 mM DTT, 1 mM ATP, pH 7.5) (NEB) and 400 cohesive end units T4 DNA ligase (NEB) at 16°C overnight. Ligation reactions were used directly to transform chemical competent *E. coli* DH5α cells and grown overnight at 37°C on 100 µg/mL ampicillin LB agar plates. Colonies were screened for inserts by PCR and vectors from positive colonies were sent for DNA sequencing (Psomagen Inc.).

#### **3.5.4 Transfection of *G. lamblia* cells with *pAC* and *pAN* vectors**

Cells collected from 200 mL of *G. lamblia* WB C6 (ATCC 30957) culture at >90% confluence were detached from tubes on ice for 30 minutes, then centrifuged for 10 mins at 1000 x g, 4°C. Old medium was removed and cells were resuspended in fresh TYI-S-33 supplemented medium to a concentration of approximately  $4.0 \times 10^7$  cells/mL. For each transfection 400 µL of cells were added to electroporation cuvettes (4 mm gap) on ice. 20 µL of the appropriate vector at a concentration of 2.5 µg/µL was added to cuvettes and incubated on ice for 5 minutes. Electroporations were performed using a Bio Rad Gene Pulser Xcell Electroporation system with the settings: 350 V, 1000 µF, 700 Ω. Cuvettes were placed back on ice for 10 minutes, then cells were transferred to culture tubes containing fresh TYI-S-33 supplemented medium containing no antibiotic and placed at 35°C to grow for 24 hours. The following day puromycin was added to the cultures to a final concentration of 50 µg/mL. After 3-4 days under puromycin selection medium was replaced with fresh medium containing puromycin and allowed to grow until adherent cells were observed at >80% confluence.

#### **3.5.5 Cloning of *G. lamblia* proteins for expression in *E. coli***

Coding sequences for *G. lamblia* proteins were amplified for *G. lamblia* genomic DNA with Phusion® DNA polymerase (NEB) as described above, see Table A.4.1 for specific oligonucleotide primer combinations. PCR products were cleaned up using the E.Z.N.A.® Cycle Pure kit (Omega Bio-tek) following the manufacturers instructions. Coding sequence products and either pET28a or pETDuet vectors were double digested with various combinations of restriction enzymes. Restriction digests were incubated for 1 h at 37°C, then heat inactivated for 20 mins at 65°C. Vector digests were treated with 10 U

of calf intestinal phosphatase for an additional 1 h at 37°C. All digests were cleaned up using E.Z.N.A.® cycle pure kit. Ligations were set up as described above for pAN vectors and incubated overnight at 16°C.

Ligation reactions were used to transform chemically competent *E. coli* DH5α cells then plated on LB + 100 µg/mL Amp agar and grown overnight at 37°C. The following day cells were screened for the correct insert by colony PCR. Colonies with inserts of the correct size were cultured overnight in liquid LB broth + 100 µg/mL Amp and plasmids were purified and sent for DNA sequencing (Psomagen) to confirm integrity of the coding sequences. Plasmids containing the correct coding sequences were used to transform either BL21 (DE3) or Rosetta II *E. coli* expression strains by heat shock transformation.

#### ***3.5.6 Overexpression and purification of recombinant G. lamblia proteins in E. coli***

*E. coli* BL21 (DE3) or Rosetta II strains containing vectors to express recombinant *G. lamblia* proteins were inoculated into liquid LB medium with appropriate antibiotics (50 µg/mL Kan for cells containing pET28a vectors, 100 µg/mL Amp for cell containing pETDuet vectors, 34 µg/mL Chloramphenicol for Rosetta II *E. coli* strain) and grown to an OD<sub>600</sub> of 0.4-0.6. Cultures were induced with Isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 1 mM and allowed to express protein for 4 hours at either 37°C or room temperature (Nop56 and Nop58 expressions) with gentle agitation. Cells were collected and resuspended in either Arg-Glu wash lysis buffer 20 (300mM NaCl, 50 mM NaH<sub>2</sub>PO<sub>4</sub>, 20 mM imidazole, 50 mM L-Arginine, 50 mM L-Glutamine) for cells expressing Nop56-fibrillarin and Nop58-fibrillarin, or wash/lysis buffer 20 (300 mM NaCl, 50 mM NaH<sub>2</sub>PO<sub>4</sub>, 20 mM imidazole) for all other protein expressions. Cells were lysed by French® press at an internal cell pressure of 14,000 psi. Lysate was cleared by

centrifugation at 10,000 rpm for 1 h, 4°C in an Eppendorf F-34-6-38 rotor. Cleared lysate was added to 2 mL of Ni<sup>2+</sup> NTA His-Bind® bed resin (Novagen) and incubated on ice with gentle agitation for 1 h. For  $\alpha$ -Snu13p and  $\beta$ -Snu13p purifications resin was transferred into 1mL of benzonase nuclease buffer (50mM NaCl, 1.5 mM MgCl<sub>2</sub>, 50 mM Tris Base, pH 8.0) and incubated at room temperature for 1 h with 125 U of benzonase (NEB), then washed twice with 5 column volumes of high salt benzonase nuclease buffer (500 mM NaCl, 1.5 mM MgCl<sub>2</sub>, 50 mM Tris Base, pH 8.0) to removed free oligonucleotides. Proteins were eluted from resin 6 times, each with a half column volume of Ni<sup>2+</sup> elution buffer (250 mM imidazole). Eluted protein was buffer exchanged into complex assembly buffer (150 mM NaCl, 20mM HEPES, 1.5mM MgCl<sub>2</sub>, 0.1 mM EDTA, 0.75 mM DTT, 10% glycerol, pH 7.0) and concentrated, then quantified by Bradford assay.

### ***3.5.7 In vitro transcription of Fluorescently labeled RNA***

Template DNA for *in vitro* transcriptions was generated by PCR amplification of the desired gene from either 100 ng of *G. lamblia* WB C6 DNA (for RNAs >50 nt) or by hybridization of overlapping forward and reverse primers (RNAs <50 nt) with forward primers containing T7 RNA polymerase promoter sequence (Table A.4.1). PCRs were performed as 50  $\mu$ L reactions in 1 x Phusion HF buffer, 0.2 mM each dNTPs, 40 pmol (or 200pmol for G1 U4 SL and associated variant RNAs) of both forward and reverse oligonucleotide primers and 1 U of Phusion® DNA polymerase (NEB). For <50 nt RNAs PCRs were run with an initial denaturing step at 98°C for 3 minutes, followed by 35 cycles of denaturing for 10 sec at 98°C, annealing of primers for 20 sec at 68°C, and extension of primers for 10 sec at 72°C. This was followed by a final elongation at 72°C for 5 min. For RNAs >50 nt conditions were the same as described above for gene amplification for

cloning into bacterial expression vectors. PCR reactions were purified by phenolic extraction using one reaction volume of phenol (pH 6.6), followed by two chloroform washes. DNA was then ethanol precipitated in 2.5 reaction volumes of 99% ethanol with 1/10<sup>th</sup> reaction volume of 3M sodium acetate (pH 5.2) in the presence of linear acrylamide carrier and incubated overnight at -20°C. Precipitated DNA was pelleted by centrifugation at 17, 000 x g for 10 min, then washed twice with 75% ethanol to remove salt, dried, and resuspended in ddH<sub>2</sub>O.

Fluorescently labeled RNAs were generated by *in vitro* transcription using purified DNA templates. For a 1X *in vitro* transcription reactions consisted of 20 µL reactions containing 1X T7 RNA polymerase buffer (40 mM Tris-HCl, 6 mM MgCl<sub>2</sub>, 1 mM DTT, 2 mM spermidine, pH 7.9), 100 to 200 ng DNA template, 1 X Fluorescein 12-UTP label mix (Roche), 40 U RNase inhibitor murine (NEB), and 100 U T7 RNA polymerase (NEB), in DEPC ddH<sub>2</sub>O. Reactions were incubated at 37°C for 2 h, then treated with 2 U of DNase I for 10 min at 37°C. Reactions were immediately phenol/chloroform extracted, then ethanol precipitated as described and resuspended as above but in DEPC treated ddH<sub>2</sub>O. RNAs were size selected on an 8 M Urea 15% PAGE gel. Fluorescent RNA was visualized on an Amersham™ Typhoon™ and bands of the correct size were gel extracted, crushed, and mixed with oligonucleotide elution buffer (0.5 M ammonium acetate, 10 mM magnesium acetate) and an equal volume of phenol (pH 6.6). Extracts were incubated overnight at room temperature with end-over-end rotation. Aqueous phases were removed and washed twice with chloroform then ethanol precipitated as described above and resuspended in complex assembly buffer. RNAs were quantified by running RNA samples along with serial dilutions of the Fluorescein-12-UTP labeling mix on 8 M Urea 15% PAGE gel then quantifying fluorescence intensity using ImageJ software and generating a standard curve.



### **3.5.8 Electrophoretic mobility shift assays (EMSA)**

Fluorescently labeled RNA was folded by heating to 65°C then cooling at 4°C/min to a final temperature of 4°C. Recombinant protein was diluted in complex assembly buffer to the desired concentrations in 10 µL reactions, also including 100 fmol fluorescently labeled RNA, and 2.5 µg *E. coli* tRNA. Reactions were incubated for 10 mins at 37°C then resolved on either a native PAGE gel containing 2% glycerol, run in 50 mM potassium phosphate pH 7.0 at 4°C, or on an agarose gel in 1 X TBE at room temperature. Fluorescent RNA was imaged using an Amersham™ Typhoon™. Where dissociation constants ( $K_D$ ) calculations were performed EMSA gels were run in triplicate and fluorescence intensity of unbound and protein-bound RNA was quantified using ImageJ software. Data output from ImageJ were used to generate curves and calculate dissociation constants by plotting protein concentration against proportion of RNA bound using Prism (v9.1.0).

### **3.5.9 Western blotting of tagged *G. lamblia* proteins**

*G. lamblia* cell pellets were thawed on ice for 20 minutes then resuspended in 2.5 mL of buffer A (50 mM Tris-HCl [pH 7.5], 100 mM NaCl, 5 mM MgCl<sub>2</sub>, 1 mM EDTA, 10% glycerol, 0.5% Nonidet P-40, 0.5 mM DTT, 1 mM Na<sub>3</sub>VO<sub>4</sub>, 2X Complete Mini Protease inhibitor tablet [Roche]) and lysed by Dounce homogenization followed by sonication (Output control 6, 50% duty cycle). Total cell lysate, Precision Plus Protein™ Dual Color Standard (Bio-Rad) and FLAG containing total lysate as a primary antibody control were resolved on separate lanes of a 15% SDS-PAGE gel at 250V. Gels were then equilibrated in western transfer buffer (25 mM Tris, 192 mM glycine, 20% methanol) for 15 minutes. Proteins were electroblotted by wet transfer to polyvinylidene difluoride (PVDF) membranes (activated in 100% methanol for 1 minute). Electroblotting was run

for 2 h 30 min at 20V, 4°C. Western blots were performed using Chemiluminescent Western Blotting Kit (Rockland Inc), according to manufacturers instructions. Briefly, 1 µL of mouse IgG antibody was added to membrane as a secondary antibody control. Each membrane was blocked for 1 h at room temperature in TBS + 0.1% Tween-20 + 1% bovine serum albumin (BSA). Primary mouse anti-FLAG antibody was diluted in TBS + 0.1% Tween-20 (TTBS) to 1:1000 and incubated with membrane for 1 h at room temperature with gentle agitation. Membranes were then washed 3 times with 1 X TTBS for 5 minutes each. Horseradish peroxidase (HRP) conjugated anti-mouse IgG antibody diluted to 1:20,000 in TTBS was then added to the membranes and incubated at room temperature for 1 h with gentle agitation. Membranes were again washed 3 times with 1 X TTBS for 5 minutes each. FemtoMax™ Supersensitive Chemiluminescent substrate was added to each membrane and FLAG-tagged proteins were visualized with an Amersham™ Imager 600.

#### ***3.5.10 Protein co-precipitations with *G. lamblia* Snu13p homologs***

*G. lamblia* cultures were placed on ice for 30 minutes to allow cells to detach from culture tubes. Cells were then transferred to 50 mL falcon tubes and pelleted by centrifuging at 1000 x g for 10 minutes at 4°C. Cell pellets from 500 mL of culture were pooled and stored at -80°C. Purification protocol for TAP tagged proteins was adapted from Jerlström-Hultqvist et al. (2012). *G. lamblia* cell pellets from TAP-tagged protein expression cultures were removed from -80°C and placed on ice to thaw, then resuspended in 5mL of buffer A or stringent buffer A (buffer A containing 1 M NaCl and 1% Nonidet P-40) and incubated on ice for 20 minutes. Resuspended cells were lysed by Dounce homogenization, followed by sonication (Output control 6, 50% duty cycle), then cleared of cell debris by centrifugation at 12,000 rpm in a Eppendorf F-34-6-38 rotor for 10 minutes at 4°C. 200 µL

of Strep-Tactin® Superflow™ beads (IBA) 50% slurry (100 µL bed resin) was washed twice by suspending in buffer A, then pelleting by centrifugation at 950 x g for 5 minutes at 4°C and removing supernatant. Soluble cell lysate was added to the washed Strep-Tactin® resin and incubated for 1 h ( $\beta$ -Snu13p(1h), Ucp1, Ucp2, Ucp3, S4) or overnight ( $\alpha$ -Snu13p,  $\beta$ -Snu13p(24h)) at 4°C with end-over-end rotation. Resin was pelleted by centrifugation (950 x g, 4°C, 5 minutes), and supernatant was collected as unbound flow through. Resin was resuspended in 400 µL of buffer Strep (buffer A with 0.1% instead of 0.5% Nonidet P-40) and transferred to Pierce Spin Cap Spin Columns (Pierce). Columns were centrifuged at 900 x g for 1 min at room temperature, and flow through collected as wash 1. An additional 400 µL of buffer Strep was added to the resin and incubated for 2 minutes at 4°C with end-over-end rotation. Columns were centrifuged as before. This process was repeated for a total of 6 washes. Bound protein was eluted twice from the resin by addition of 500 µL of desthiobiotin elution buffer (50 mM Tris-HCl [pH7.5], 150 mM NaCl, 5 mM MgCl<sub>2</sub>, 1 mM EDTA, 10% glycerol, 2 mM D-desthiobiotin [IBA]) to the column followed by a 10-minute incubation at 4°C with end-over-end rotation. Columns were centrifuged at 900 x g for 1 minute at room temperature to collect the first elution, then repeated for a second elution.

Samples were concentrated for mass spectrometry via acetone precipitation by adding 4 volumes of ice cold 80% acetone containing 15 mM NaCl (Crowell et al. 2013), then incubated at -20°C overnight. The following day precipitated proteins were pelleted by centrifugation at 17,000 x g for 10 minutes, then washed twice with ice cold 80% acetone. Protein pellets were dried under vacuum to remove trace acetone, then resuspended in ddH<sub>2</sub>O. Elutions containing protein were resolved on 15% SDS-PAGE gels then stained with Coomassie Brilliant Blue R-250. Regions containing visible protein bands

were extracted as a single piece and sent to the Alberta Proteomics and Mass Spectrometry Facility for mass spectrometry analysis.

### ***3.5.11 Analysis and annotation of G. lamblia proteins***

Unannotated proteins detected in the co-precipitation experiments with *G. lamblia* Snul3p homologs were searched against the NCBI non-redundant protein database using BLASTp for potential homologs. Proteins that remained unannotated were analyzed using Phyre2 (Kelley et al. 2015) and classified based on identified domain architecture compared to homologous proteins in yeast or humans wherever possible. To identify SSU processome proteins in the *G. lamblia* genome we downloaded sequences for all known small subunit processome from the Saccharomyces Genome Database (<https://www.yeastgenome.org/>) and used them in BLASTp searches to the *G. lamblia* proteome in Giardiadb (<https://giardiadb.org/giardiadb/>). Reciprocal searches were then performed using BLASTp against the NCBI non-redundant protein database to support annotation. BLASTp searches of the Ucp 1-3 proteins to the nr protein database in NCBI and giardiadb were performed to look for potential protein homologs. Additional domain predictions were performed using Phyre2, SWISS-MODEL, I-TASSER, and GenomeNet Motif tools (<https://www.genome.jp/tools/motif>) using default settings. *G. muris* homologs were identified using BLAST searches to the *G. muris* genome in giardiadb using the *G. lamblia* homologs.

To identify metamonad C/D snoRNP assembly factors searches were performed with BLAST in giardiadb using *S. cerevisiae* C/D snoRNP assembly factor protein sequences as the query. Hits from giardiadb were used to perform reciprocal BLAST searches to the NCBI non-redundant (nr) protein database. A second approach was also

employed by searching metamonad annotated proteomes (downloaded from giardiadb) for domains using the GenomeNet Motif tool (<https://www.genome.jp/tools/motif/>) to the Pfam database on default settings (E value < 1.0). Results of this search were examined for conserved domain architecture found in either the *H. sapiens* or *S. cerevisiae* C/D snoRNP assembly factor proteins. Proteins containing the correct relevant domains were used in reciprocal BLASTp searches to the NCBI nr protein database to look for related homologs in species outside of yeast and humans.

### ***3.5.12 Immunofluorescence microscopy***

*G. lamblia* cell cultures expressing recombinant FLAG-tagged proteins were grown as described above. Cell cultures at >90% confluence were detached from culture tubes by placing on ice for 10 minutes. 7 mL of each culture was centrifuged for 10 minutes at 1000 x g, 4°C to pellet cells. Medium was aspirated off and cell pellets were resuspended in 1 mL of TYI-S-33 supplemented medium. Resuspended cells were transferred to 6 well plates containing a sterile glass microscope coverslip in each well. Cells were incubated at 37°C in a humidified chamber for 1 h to allow adherence to the cover slips. Medium was removed and cells were fixed with 2 mL of 4% paraformaldehyde in 1X phosphate buffered saline (PBS) for 20 minutes at room temperature. The fixing agent was removed, and wells were washed twice with 1 mL 1 X PBS. A single combined blocking and permeabilization step was performed by adding 2 mL IFA blocking/permeabilization buffer (3% BSA, 0.1% Triton X-100, in 1 X PBS) to each well and incubating for 1 h at 37°C in a humidified chamber. Primary mouse IgG antibody (Rockland Inc) was diluted to 1:100 in 1 X PBS + 0.1% Tween-20 (PBST) + 3% BSA, 1 mL was added to each well and incubated for 1 h at 37°C in a humidified chamber. Each well was washed 3 times with 2 mL PBST for 5 min.

Alexa Fluor™ 488 fluorophore conjugated secondary rabbit anti-mouse IgG antibody (Invitrogen) was diluted to 1:400 in PBST and added to each well, then incubated at room temperature in a humidified chamber in the dark for 1 h. The secondary antibody solution was removed, and wells were washed 3 times with 2 mL PBST for 5 minutes each. 300 nM DAPI in PBST was added to each well and incubated at room temperature for 10 minutes then washed 3 times with 2 mL PBST for 5 min each. Cover slips were removed from 6 well plates and mounted onto glass slides using Prolong™ Gold Antifade Mountant (ThermoFisher Scientific) then left to cure for 24 h before visualizing. Cells fixed to 6 well plates were visualized using a Cytation™ 5 Cell Imaging Multi-Mode Reader at 20X magnification. Immunofluorescence slides were visualized on an Olympus Fluoview FV1200 confocal microscope using the FV10-ASW 4.2 software at 60X magnification.

### ***3.5.13 RNA co-precipitations with *G. lamblia* Snu13p homologs***

The same protocol used to co-purify tagged proteins with cellular proteins for mass spectrometry was followed to co-purify associated RNAs up until the wash steps. The only exception being that following the clearing of the cell lysates, 100  $\mu$ L of the cleared lysate was collected for RNA extraction to make RNA input control libraries. For washes, resin with bound protein was resuspended in 400  $\mu$ L of buffer strep and transferred to a Pierce Spin Cap Spin Column (Pierce). Columns were centrifuged at 900 x g for 1 min and flowthrough collected as the first wash. 400  $\mu$ L of buffer strep was added to each column and incubated for 2 mins at 4°C with end-over-end rotation. Columns were centrifuged again at 900 x g for 1 min and flowthrough collected. This process was repeated for a total of 6 washes for  $\beta$ -Snu13p and 8 washes for  $\alpha$ -Snu13p (optimized for the least number of washes required to ensure no RNA is detected in final wash). For the respective final wash,

resin and buffer strep were transferred to 1.5 mL microcentrifuge tubes incubated for 2 mins at 4°C with end-over-end rotation then centrifuged for 1 min at 900 x g. Supernatant was removed and 1 mL of TRIzol® reagent was added to the resin and corresponding 100 µL of soluble cell lysate then mixed by pipette and incubated for 5 minutes at room temperature. 200 µL of chloroform was added to each extraction tube and shaken manually for 15 seconds then incubated at room temperature for 3 minutes. Extractions were then centrifuged for 15 minutes at 12,000 x g. The aqueous phase containing RNA for each extraction was collected and precipitated by adding 1 µL of linear acrylamide carrier, 1/10th volume of 3 M sodium acetate (pH 5.2) and 2.5 volumes of ice cold 99% ethanol, then incubating at -20°C overnight. The following day precipitations were centrifuged at 17,000 x g for 10 minutes to pellet RNA. RNA pellets were washed twice with ice cold 75% ethanol then dried under vacuum and resuspended in 20uL of DEPC treated ddH<sub>2</sub>O.

### ***3.5.14 TGIRT™-III library preparation for RNA-seq***

Resuspended RNA was fragmented to an average size of 70-100 nucleotides using the NEBNext® Magnesium RNA Fragmentation Module (NEB) according to the manufacturer's instructions, incubated at 94°C for 7 mins. Fragmentation reactions were cleaned up using the RNA Clean and Concentrator™ Kit (Zymo Research) following a modified version of the manufacture's instructions where 8 volumes of ethanol were used to increase retention of small RNAs. Fragmented RNAs were treated with T4 Polynucleotide kinase (NEB) (1 x T4 PNK buffer (70 mM Tris-HCl, 10 mM MgCl<sub>2</sub>, 5 mM DTT, pH 7.6) 1 U T4 Polynucleotide kinase) to ensure all RNAs possess 3'-OH ends. Reactions were cleaned up again using RNA Clean and Concentrator™ Kit (Zymo

Research) as described above. Treated RNAs were analyzed on an Agilent Bioanalyzer 2100 with an RNA pico 6000 chip to assess size distribution of fragments.

RNA-seq libraries were prepared using the TGIRT™-III Kit (InGex). 10X Primer Mix containing R2 RNA and R2R DNA (Table A.4.1) was denatured for 2 minutes at 82°C then allowed to anneal by cooling to 25°C at a constant rate of 0.1°C/s. First strand cDNA was synthesized using template switching TGIRT™-III reverse transcriptase with control input RNA and co-precipitated RNA samples as template (1X Reaction buffer (450 mM NaCl, 5 mM MgCl<sub>2</sub>, 20 mM Tris-HCl, pH 7.5), 10 mM DTT, 100 nM of annealed R2 RNA/R2R DNA, 500 nM TGIRT™-III) and incubated at room temperature for 30 minutes. dNTPs were added to each reaction to a final concentration of 1.25 mM each, then incubated at 60°C for 15 minutes. Reactions were inactivated by addition of 5M NaOH and incubation at 95°C for 3 mins then cooled to room temperature and neutralized with 5M HCl. RT reactions were cleaned up twice sequentially with the MinElute® Reaction Cleanup Kit (Qiagen). R1R DNA adapters were adenylated using a 5' DNA Adenylation Kit (NEB). 5' Adenylated R1R DNA adapters were then ligated to the first strand DNA using 5' App DNA/RNA ligase (1x NEBuffer 1 (10 mM Bis-Tris-Propane-HCl, 10 mM MgCl<sub>2</sub>, 1 mM DTT), 5 mM MnCl<sub>2</sub>, 40 pmol 5' App DNA/RNA ligase (NEB), 1 µM 5' adenylated R1R DNA), incubating for 2 hours at 65°C, then inactivating at 90°C for 3 minutes. Ligation reactions were cleaned up using MinElute® Reaction Cleanup Kit (Qiagen). PCR amplification of cDNA for final library construction was done using Phusion DNA polymerase (NEB) (1X HF Buffer, 200 µM each dNTPs, 1 U (2000U/mL) Phusion polymerase, 200 nM Illumina Multiplex primer, 200 nM Illumina Barcode Primer (Table A.4.1)) and thermocycler parameters: 98°C 5s, (98°C 5s, 60°C 10s, 72°C 15s) x 14



cycles. Final PCR amplified libraries were cleaned up using AMPure XP beads (Beckman Coulter) at a 1.4X ratio, according to the manufacturer's instructions.

Size distribution and ligated adapter contamination was analyzed on an Agilent Bioanalyzer 2100 with a DNA HS chip. Libraries were quantified using the NEBNext® Library Quant Kit for Illumina® according to the manufacturer's instructions. Quantified libraries were pooled in a single tube to a final concentration of 7 nM each for sequencing. Libraries were sent to GENEWIZ® for sequencing on an Illumina HiSeq system with paired end 150 bp reads.

### **3.5.15 Analysis of RNA sequencing libraries**

RNA-seq library sequence quality was assessed with FastQC, then adapter sequences and short reads were removed from sequence read files in fastq format using the Cutadapt software (v2.10) (Martin 2011) (with settings -a AAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC, -A GATCGTCGGACTGTAGAACTCTGAACGTGTAG, -j 6, -m 20). Low quality nucleotides were removed using Trimmomatic v0.39 (with settings TRAILING:30). In order to map reads an index was created for the RNA-seq aligner STAR (v2.7.6a) (Dobin et al. 2012) using *Giardia lamblia* genome version 49 and the *Giardia lamblia* annotation gff file (both acquired from giardiadb.org) manually supplemented with non-coding RNAs using the parameters --genomeSAindexNbases 10, --sjdbGTFtagExonParentTranscript Parent, --sjdbOverhang100 and all other parameters default. Reads were then aligned to the *Giardia lamblia* genome version 49 using STAR and the following parameters: --alignIntronMax 65, --outSAMtype BAM Unsorted, --outReadsUnmapped Fastx, --

outSAMprimaryFLAG AllBestScore, --outFilterScoreMinOverLread 0.5, --outFilterMatchNminOverLread 0.5, and all other parameters default.

Aligned reads for each library were assigned to genomic features using the featureCounts feature of the R package Rsubread (v2.2.6) (Liao, et al. 2019). Additional modifications to the gff annotation file including removal of small predicted open reading frames overlapping other features and addition of rRNA fragment annotations were made prior to genomic feature assignment. The following parameters were used for featureCounts: isGTFAnnotationFile = TRUE, requireBothEndsMapped = TRUE, minOverlap = 10, countMultiMappingReads = TRUE, fraction = TRUE, all other parameters set to default. Raw counts from featureCounts were used to calculate normalized counts per million (CPM) and transcript per million (TPM) as previously described (Boivin et al. 2018).

Enrichment values for co-precipitated RNAs were calculated by comparison of precipitated RNA reads and control input RNA reads using the glmFIT function from the edgeR (v3.30.3) package in R. Only genes with at least 5 counts per million (CPM) in each replicate were further considered for analysis, using a significance threshold of FDR  $P < 0.05$  and  $>1 \log_2$  fold change up or down in read count. FDR  $P$  values were calculated in R using the  $P$  value obtained from edgeR analysis. To check for potential novel ncRNAs the small number of unassigned reads for RNA-seq libraries were extracted from the processed FASTQ files using custom Python3 scripts and filterbyname.sh from the BBmap suite (v38.89) (<https://sourceforge.net/projects/bbmap/>). Unassigned reads were mapped to the genome with STAR as described above, then reads were used to generate features using Cufflinks (Trapnell et al. 2012).

### ***3.5.16 Glycerol gradient ultracentrifugation of tagged S4 cell lysate***

*G. lamblia* cell lysates were prepared from cell pellets from 450 mL of culture as described for protein and RNA co-precipitations above. Glycerol gradients were prepared from 10-30% glycerol in buffer A lacking protease inhibitor. 2.5 mL of soluble cell lysate was loaded onto a 35 mL gradient and centrifuged for 20 h at 26, 000 rpm, 4°C in an SW28 rotor. Following centrifugation, the gradients were collected as 2 mL fractions starting with the densest fractions using an ÄKTA prime plus system. Large complexes that ran through the entire gradient to form pellets on the bottom of the centrifugation tubes were then resuspended in buffer A + 10% glycerol, then acetone precipitated as described for protein mass spectrometry preparation to concentrate. Gradient fractions and resuspended pellets were analyzed via western blot as described above to determine protein locations within the gradient or protein pellet.

**Chapter 4: RNA-Seq employing a novel rRNA depletion strategy reveals a rich repertoire of snoRNAs in *Euglena gracilis* including box C/D and  $\Psi$ -guide RNAs targeting the modification of rRNA extremities.**

Reprinted with permission from:

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

**Authors:** Ashley N. Moore<sup>†</sup>, David C. McWatters<sup>†</sup>, Andrew J. Hudson, and Anthony G. Russell

<sup>†</sup> These authors contributed equally to this manuscript.

**Publication:** RNA Biology

**Publisher:** Taylor & Francis

**Date:** Oct 3, 2018

Rights managed by Taylor & Francis

**Contributions:**

DCM performed the bulk of bioinformatic analysis and wrote the final draft of the manuscript. ANM performed experiments, initial bioinformatic analysis, and wrote a first draft of the manuscript. AJH assisted with  $\Psi$ -guide snoRNA analysis. All authors read and approved the final manuscript.

**Changes incorporated for the thesis:**

Additional analysis of U1 snRNA and tRNA repeats were performed following publication and further findings from the RNA-seq library were added. Also, minor changes were made to figures 1 and 2 and the introduction for clarity within the context of the thesis. Some additional more recent references were added where appropriate.

## 4.1 Introduction

As described in Chapter 1, non-coding RNAs (ncRNAs) have essential roles in an array of gene expression mechanisms in all organisms (Huttenhofer et al. 2005; Mattick and Makunin 2006; Matera et al. 2007; Goodrich and Kugel 2009; Mercer et al. 2009; Faust et al. 2012). Comprehensive strategies to identify ncRNAs have been developed utilizing deep sequencing technologies in the form of RNA-seq. During this procedure it is advantageous to fractionate and enrich the ncRNA population of interest from total cellular RNA prior to cDNA library creation to remove very abundant cellular RNAs that would otherwise dominate the sequence reads. This allows for the more efficient and cost-effective identification of less abundant and novel non-coding RNA species. Commercial kits have been developed to remove rRNA during library preparation; however, they are only available (or work efficiently) for a limited number of model organisms. They also increase the number of sample handling steps which can further increase the likelihood of generating unnatural RNA degradation products. Such kits are not yet available for most protists, a collection of primarily single-celled eukaryotic organisms that includes *Euglena gracilis*, the organism investigated in this study.

*E. gracilis* is a particularly interesting organism in which to characterize ncRNAs because of the many unusual features of its cellular biology and gene expression strategies that suggest it may contain a large collection of ncRNAs (Ebenezer et al. 2017; McWatters and Russell 2017). mRNA transcriptome studies have been performed (O'Neill et al. 2015; Yoshida, et al. 2016; Ebenezer, et al. 2019) that indicate that *E. gracilis* has extensive protein-coding potential and it has been suggested that expression of nuclear protein-coding genes is extensively controlled at the post-transcriptional level. This organism contains a

large subunit (LSU) rRNA that is naturally fragmented into 14 discrete pieces (compared to 2 in most other eukaryotes) by post-transcriptional processing events (Schnare, et al. 1990; Schnare and Gray 1990). *E. gracilis* rRNA is also the most extensively modified of any examined organism to date, containing 211 2'-*O*-methylations (Nm) and 116 pseudouridine (Ψ) modifications (Schnare and Gray 2011). The LSU rRNA is more extensively modified than the non-fragmented small subunit (SSU) rRNA, which instead has a similar amount of modification as its human counterpart. These extra modifications are predicted to help stabilize the highly fragmented LSU during ribosome assembly (Schnare and Gray 2011).

This extensive degree of modification and pre-rRNA processing makes snoRNAs, which are responsible for guiding 2'-*O*- methylation and Ψ formation along with targeting pre-rRNA cleavage, of particular interest in *E. gracilis*. Since *E. gracilis* has so many rRNA modifications and processing sites it is predicted to also contain a large collection of targeting snoRNAs. Previously we had identified snoRNAs that guide 47% of the experimentally mapped rRNA 2'-*O*-methylation sites but only 11% of the Ψ sites. Two of the *E. gracilis* rRNA nucleotide modifications are located very close to the 3' ends of two different rRNA species. We had not yet uncovered any snoRNA species capable of targeting these sites and thus the mechanism of modification was still unclear.

The small number of *E. gracilis* Ψ-guide RNAs identified previously structurally differ from the H/ACA box snoRNAs first characterized in other eukaryotes. The prototypical structure consists of two extended stems, either or both of which are interrupted by single-stranded regions that base-pair to rRNA to form pseudouridylation guide pockets (see Figure 1.2B, Chapter 1). A single-stranded linker region containing the

H box sequence (ANANNA) separates the two stems, and an 'ACA' consensus sequence box element follows the second stem and is usually located 3 nt from the 3' end of the RNA (Watkins and Bohnsack 2012). In contrast, the small number of *Euglena*  $\Psi$ -guide RNAs identified previously possess only a single-stem structure (no H box) and possess an AGA rather than an ACA (box) sequence motif at their 3' ends (Russell, et al. 2004; Moore and Russell 2012). AGA box  $\Psi$ -guide RNAs have also been identified in trypanosome species (Liang et al. 2005; Eliaz et al. 2015). Whether or not this is a structurally common form for these RNAs in Euglenozoa, the evolutionarily-diverse phylum containing the euglenids (including *Euglena* species), kinetoplastids (including trypanosomes) and many other classes of protist organisms, requires a more comprehensive characterization of  $\Psi$ -guide RNAs in *E. gracilis* and its relatives (Russell, et al. 2004, 2006; Moore and Russell 2012).

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

## **4.2 Results and Discussion**

### ***4.2.1 Preventing amplification of unwanted RNAs during RNA-Seq library construction***

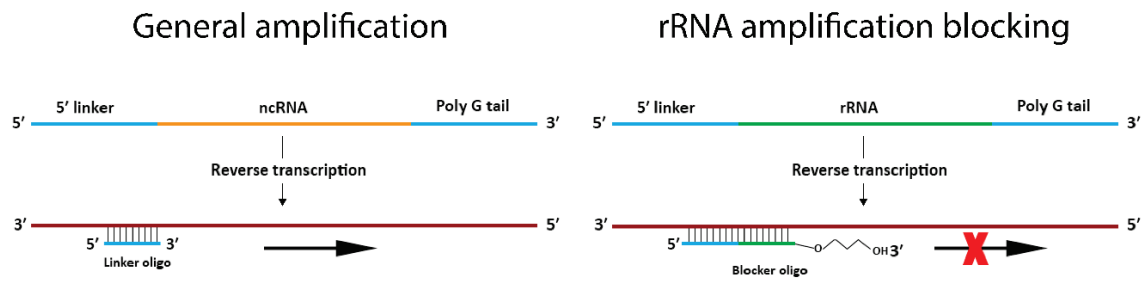
In *E. gracilis* the large subunit rRNA is naturally fragmented into 14 discrete pieces most of which fall within the size range of many other small ncRNA species. This complicates the generation of small RNA libraries in *E. gracilis* and other Euglenozoa since high levels of rRNA dominate the sequence reads from even a carefully size-selected library. To address this issue we have developed a strategy adapted from a technique

previously described for eliminating unwanted DNA sequences from environmental samples (Vestheim and Jarman 2008). This strategy utilizes blocking oligonucleotides which contain a hydrocarbon chain modification at their 3' end during the PCR amplification step of cDNA library construction to prevent amplification of targeted sequences.

A set of blocker oligonucleotides was designed to anneal to the end of the added 5' linker sequence + 5' end of each of the 14 LSU rRNA fragments (Figure 4.1a, Figure A.3.1a). To be effective, the blocker oligonucleotides must anneal efficiently to the unwanted target cDNAs while not significantly affecting the amplification of all other cDNAs. We found that 13 nt of complementarity with 5' linker sequence + 13 nt of complementarity with the 5' end of an LSU rRNA fragment worked effectively. We also found that having a 10-fold excess of each blocker oligonucleotide relative to adaptor-specific general amplification oligonucleotide greatly diminished the amplification of specific LSU sequences (Figure 4.1b and Figure A.3.1b). Without employing blocker oligonucleotides, LSU fragments that are prevalent in a size-selected (< 400 nt) *E. gracilis* RNA fraction are efficiently amplified. This can be detected by including specific LSU reverse primers with the adapter-specific amplification primers during the PCR step of the procedure (Figure 4.1b, lanes 1 and 4). When an LSU fragment-specific blocker oligonucleotide is also included during the PCR step, the targeted LSU sequence is either greatly reduced or undetectable following PCR amplification (Figure 4.1b, lanes 2, 3, 5 and 6, and Figure A.3.1b). Importantly, this was also observed when multiple blocker oligonucleotides were used in the same PCR amplification (data not shown). Using this approach, the relative number of rRNA sequence reads in RNA-Seq data is greatly reduced.



a)



b)

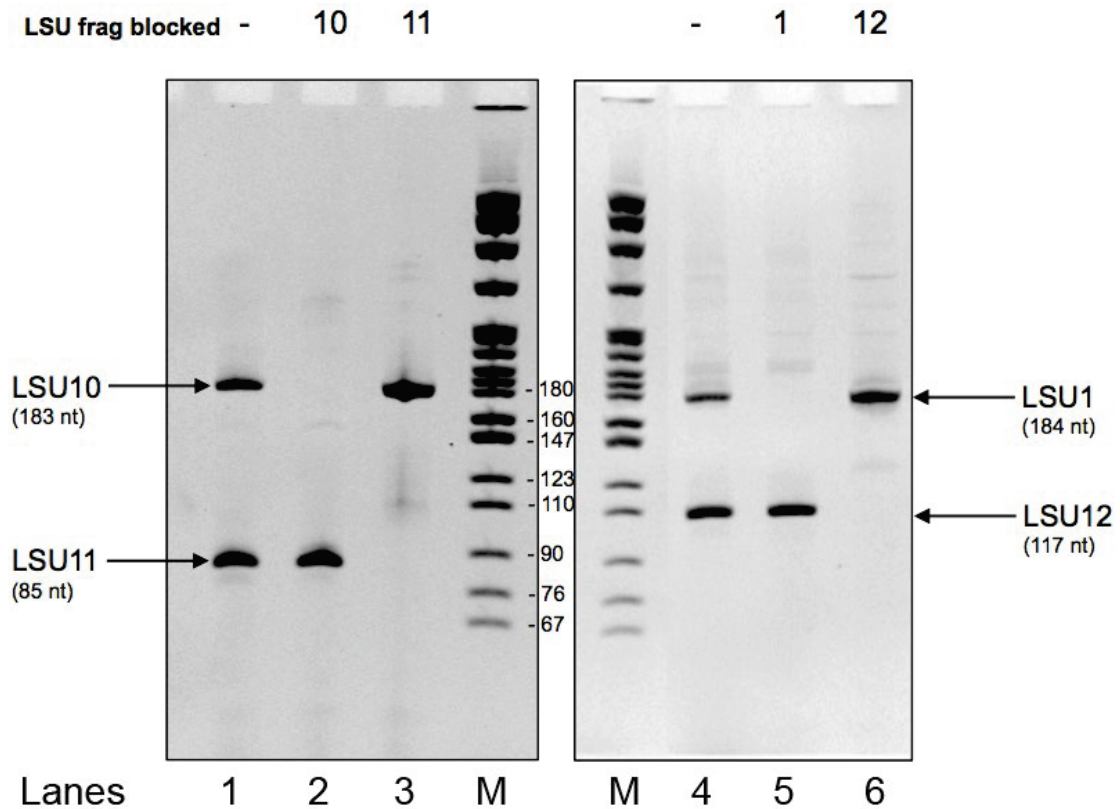
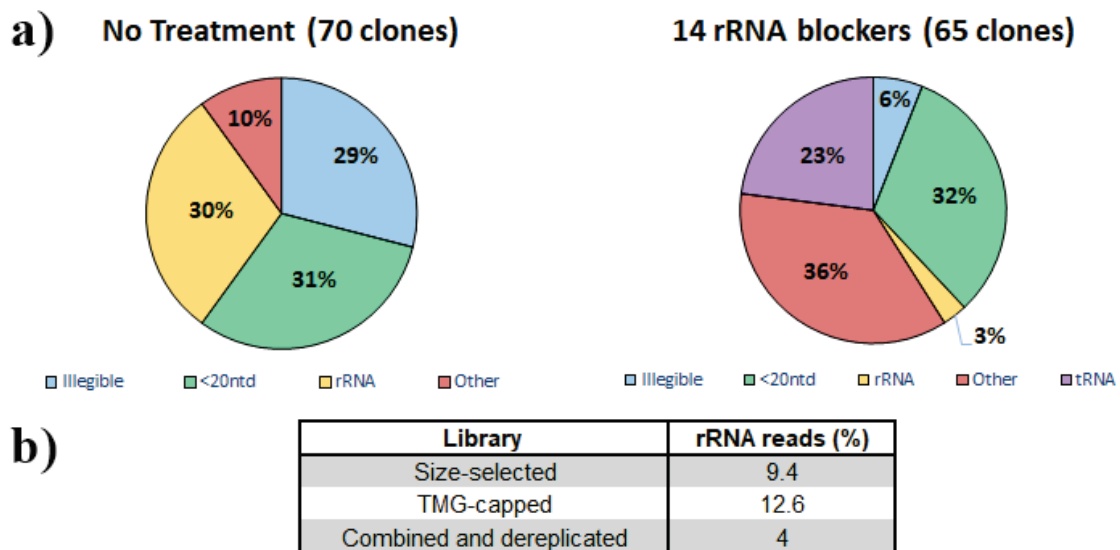


Figure 4.1. Primer blocking strategy to prevent rRNA amplification during small RNA library preparation.

**Figure 4.1. Primer blocking strategy to prevent rRNA amplification during small RNA library preparation.** **a)** Cartoon schematic of the rRNA oligo blocker strategy. In the general amplification all selected ncRNAs (orange) are Poly G tailed (blue) and a 5' linker (blue) is ligated to the RNA. The RNA is reverse transcribed, and a short linker oligo is annealed to the first strand cDNA (dark red) followed by PCR amplification. In cases where the ncRNA is an rRNA fragment (green), following reverse transcription, a blocker oligo containing a C3 hydrocarbon spacer, which is present in excess, base-pairs across the 5' linker - rRNA boundary specifically recognizing rRNA fragment sequence and preventing the universal linker oligo from annealing. The C3 hydrocarbon spacer then inhibits polymerase extension of the oligo on rRNA cDNAs. **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.

The efficiency of blocking was further assessed, prior to library deep sequencing, by shotgun cloning cDNA library products and sequencing 178 clones by Sanger sequencing of PCR products obtained via bacterial colony PCR. When no blocker oligonucleotides were used, 30% of the total unfiltered reads were rRNA (Figure 4.2a). After removing reads that were poor quality (29%) and also those less than 20 nucleotides in length, the remaining sequence reads we termed 'informative sequences'. Of these informative sequences, 75% were rRNA fragments. Initially, blocker oligos had only been designed to the 10 smallest LSU species, whose mature sizes fall in the range of small ncRNAs. Following amplification including these blockers, rRNA still dominated the informative reads; however, most of the rRNA sequences were now fragments of the 4 largest LSU species (i.e. the blocking of smaller LSU species was successful). Additional blocker oligonucleotides were then designed such that all 14 LSU species were targeted for depletion during library construction to also reduce amplification of these LSU rRNA degradation products. This was very effective, resulting in only 3% of all reads matching rRNA sequence and a significant enrichment of informative 'other' sequences (36%); that

is, sequences that consist of ncRNAs other than rRNA and tRNA were now evident (Figure 4.2a). This indicates that adding blocker oligonucleotides targeting all 14 *E. gracilis* LSU fragments at the PCR amplification step of library synthesis is very effective in reducing the number of rRNA sequences in the final library, especially when considering that in studies in other organisms, RNA-Seq data from total RNA samples with no rRNA depletion step typically contain >90% rRNA reads (He et al. 2010; Chen and Duan 2011; O'Neil et al. 2013; Peano et al. 2013).



**Figure 4.2. Sequencing results of a *Euglena gracilis* small RNA library before and after use of primer blocking to prevent amplification of rRNA fragments. a)** PCR-amplified library products were cloned and sequenced using Sanger sequencing to assess the efficacy of the primer blocking strategy. Sequence reads that were legible and longer than 20 nt were considered informative sequences. **b)** Proportion of reads attributed to rRNA for the size-selected, TMG-capped, and combined and dereplicated small RNA libraries generated using RNA-Seq.

Two different LSU rRNA-depleted *E. gracilis* small RNA libraries were created, a size-selected (< 400 nt) library and a trimethyl guanosine (TMG) cap pull-down non-size-selected library, and both were individually sequenced using paired-end 250 bp sequencing on an Illumina MiSeq platform (Genome Québec). In total, following quality control filtering, there were 3,080,604 high-quality reads when combining the reads from the size-selected and TMG-cap pull-down libraries. In the size-selected library reads, 9.4% of reads were rRNA sequence, and 12.6% of the TMG-capped library reads were rRNA (Figure 4.2b). Following dereplication of the combined library reads, there were 727,447 unique reads and searching for previously annotated *E. gracilis* RNAs revealed that approximately 4% of these reads were rRNA, 19% were snRNAs, 1.4% were known tRNAs (only one nuclear-encoded but a complete set of chloroplast tRNAs had been previously characterized in this organism) and <1% were previously characterized snoRNA sequences. Cumulatively, this indicated that our rRNA depletion strategy was very successful and worked similarly when employed on the two independently processed library fractions (size-selected and TMG cap pull-down). It is also the first indication of the apparent diversity of ncRNAs in this organism.

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

#### 4.2.2 Identification of new snoRNAs

Previously identified modification-guide snoRNAs in *E. gracilis* displayed a relatively uniform size distribution, between 50 and 90 nt, so we focused on examining sequence reads in that size range. First we identified candidates by scanning for conserved sequence and structural features (Russell, et al. 2004, 2006; Moore and Russell 2012), and then requiring that candidates be able to base-pair to corresponding rRNA target modification sites (Schnare and Gray 2011). This approach identified 82 new box C/D snoRNAs, 31 box AGA RNAs (Figures A.3.2, A.3.3, and A.3.4) and numerous isoforms of both types – we define isoforms as those sequence-related RNAs predicted to target the same modification site. Cumulatively, including all biochemically, genomically (PCR-mediated), and now RNA-Seq identified RNAs, we have characterized snoRNAs that guide modifications of approximately 88% of the 2'-O-methylated sites and 45% of pseudouridylated sites in *E. gracilis* rRNA, 227 unique snoRNA species in total.

In order to examine snoRNA representation in our data set we used single end reads from both the size-selected and TMG-capped libraries to calculate reads per million (RPM) for each newly identified snoRNA and all previously identified snoRNAs. We found that RPM for the newly identified snoRNAs were consistent with the range of RPM values found for previously identified snoRNAs in both our libraries (Table A.3.3). Additionally, when comparing reads found in the two libraries we observed that RNAs from the size-selected library consisted of a variety of both mature and precursor forms while the TMG-capped reads were generally more uniform in size and had a higher proportion of reads representing mature RNAs. All but 4 methylation guide and 2 pseudouridylation guide RNAs were detected in the size-selected library while only 78% and 51% of each type

respectively were found in the TMG-capped library. When considering the very large observed relative enrichment of the U3 snoRNA and U2 snRNA in the TMG-capped library (Table A.3.3), two ncRNAs anticipated to have hypermethylated caps, the lack of significant enrichment (or even complete absence) in this library of most *E. gracilis* modification guide snoRNAs may suggest that a large fraction of these RNAs do not possess hypermethylated caps.

All of the 82 new Box C/D snoRNAs identified by RNA-Seq appear to be single-guide RNAs and predominantly utilize the region upstream of the D' box to target modification, similar properties to the previously characterized *E. gracilis* box C/D RNAs (Russell, et al. 2004, 2006; Moore and Russell 2012). Double-guide RNAs are exceedingly rare in *E. gracilis* and of the 182 different box C/D RNA species now identified, only 2 appear to be double-guides and in both cases, each species utilizes its two guide regions to target nearby 2'-O-Me sites. This is noticeably different from what is observed in other eukaryotes and even more so, in archaeal organisms where double-guide box C/D RNAs constitute a much higher fraction of the total modification-guide RNA repertoire. The *Euglena* snoRNAs identified so far are also more uniform and smaller in size than those in other characterized eukaryotes and show closer resemblance to their archaeal counterparts.

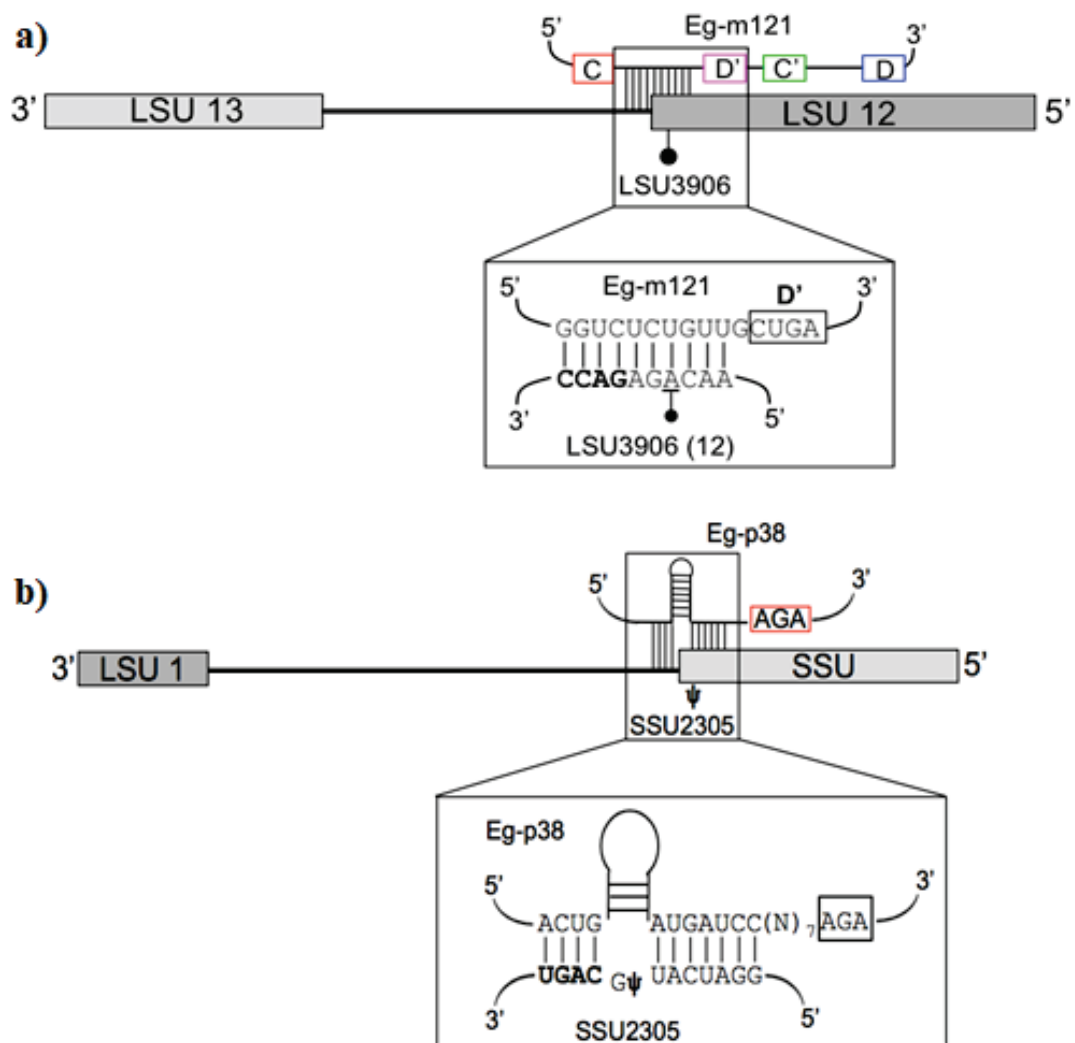
#### ***4.2.3 Importance of rRNA spacer regions and timing of snoRNA-guided rRNA modification***

Two of the new snoRNAs identified by RNA-Seq have the required base-pairing potential to target the modification sites found at the 3' extremities of rRNA species. The first, Eg-m121 guides 2'-O-methylation at position LSU3906, near the 3' end of LSU species 12. Base-pairing interactions between the snoRNA and rRNA are typical of the

length most commonly observed between box C/D RNAs and rRNA in *E. gracilis* (10 bp) and most interestingly, this base-pairing interaction extends into the (intergenic) spacer region that separates LSU species 12 and 13 on the primary rRNA transcript (Figure 4.3a). The second, Eg-p38 RNA guides  $\Psi$  formation at position SSU2305, the penultimate nucleotide at the 3' end of the SSU rRNA and the base-pairing interaction between the snoRNA and rRNA extends into ITS 1, the spacer between SSU rRNA and LSU species 1 (Figure 4.3b). In fact, the entire interaction between the 5' half of the snoRNA bi-partite base-pairing interaction to form the pseudouridine pocket with the rRNA target site occurs using only the ITS region. To our knowledge, these are the first examples of modification guide snoRNAs predicted to employ base-pairing interactions to mature rRNA-spacer sequence boundaries.

The way in which rRNA modifications coordinate with other maturation steps such as pre-rRNA cleavage is not well understood. Co-transcriptional modification has been observed in yeast prior to pre-rRNA cleavage (Kos and Tollervey 2010), but it is unclear if this a common feature among different eukaryotes. This is particularly interesting in the case of *E. gracilis* as the fragmentation of the LSU, along with the high degree of rRNA modification, results in significant enrichment of rRNA modifications near cleavage sites compared to other eukaryotes. For the two snoRNAs described above, the interaction between snoRNA and rRNA must occur before pre-rRNA cleavage that removes these particular spacer regions during the biogenesis pathway that generates the mature 3' ends. It is commonly suggested that snoRNA-rRNA base-pairing interactions occur very early in the ribosome biogenesis pathway (Watkins and Bohnsack 2012). In *E. gracilis*, the LSU is highly modified and naturally fragmented into 14 pieces indicating a high degree of

complexity in pre-rRNA processing and ribosome assembly. It is interesting to consider that these rRNA modifications and associated modification complexes may also play some role in removal of these spacer sequences to generate mature LSU rRNA fragments as appears to be the case for maturation of the six LSU fragments from *Trypanosoma brucei* (Chikne, et al. 2019).



**Figure 4.3. Identified *E. gracilis* snoRNAs whose guide regions base-pair with pre-rRNA intergenic sequence.**



**Figure 4.3. Identified *E. gracilis* snoRNAs whose guide regions base-pair with pre-rRNA intergenic sequence.** Two snoRNAs were identified that each guide a modification site found at 3' extremities of rRNA subunit species with their guide regions base-pairing to intergenic sequence. **a)** Eg-m121 snoRNA guides a 2'-*O*-methylation at position A3906 which is located 3 nucleotides from the mature 3' end of LSU fragment 12. The snoRNA guide region pairs with 4 nucleotides of the spacer region (inset, bolded nucleotides) between LSU fragments 12 and 13. **b)** Eg-p38 snoRNA guides a  $\Psi$  modification at position U2305 which is located 2 nucleotides from the mature 3' end of the SSU rRNA. The entire 5' portion of the pseudouridine pocket (relative to the snoRNA) is formed by base-pairing to the internal transcribed spacer 1 region (bolded nucleotides), the pre-rRNA region located between the 3' end of the SSU and the 5' end of the 5.8S rRNA (LSU 1).

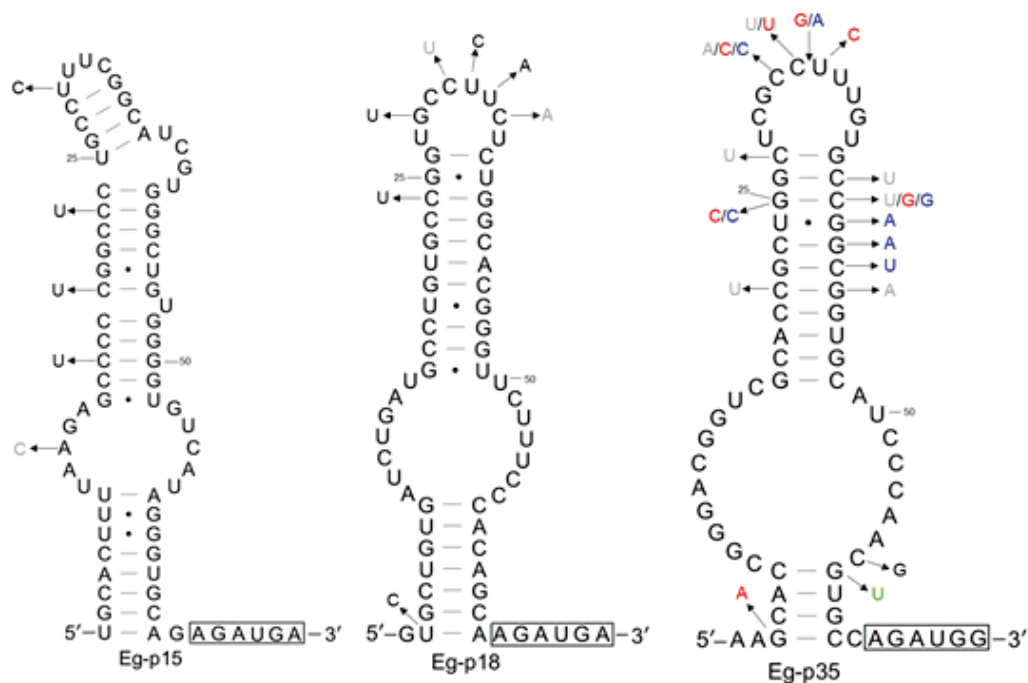
#### 4.2.4 Identification and characterization of novel $\Psi$ -guide snoRNAs

In previous studies, only 12 *E. gracilis*  $\Psi$ -guide RNA species had been identified, all of which contained a conserved 'AGA' sequence 3 nt from their 3' ends (i.e. AGA box RNAs) (Russell, et al. 2004; Moore and Russell 2012). The small number identified was due to the inefficient immunoprecipitation of these RNA-containing complexes when using antibodies targeted at the protein Cbf5p, the more challenging nature of identifying encoding regions for these structurally more complex RNAs compared to identifying box C/D RNAs within PCR-amplified *Euglena* genomic snoRNA cluster regions, and the fact that genomic amplification primers were primarily based on identified box C/D RNA sequences that presumably favor amplification of clusters encoding primarily box C/D RNAs. Bioinformatic analysis of the small ncRNA library has now significantly increased those identified to 45  $\Psi$ -guide snoRNA species with predicted rRNA target sites. Even though we allowed for much greater size variation when searching the library, the size range of these RNAs is 60 – 72 nt, with an average size of 66 nt, a remarkably uniform size distribution compared to  $\Psi$ -guide RNAs characterized in other eukaryotes. It has previously been observed that the interaction between the target rRNA and the snoRNA is

responsible for positioning the target substrate uridine (and  $\Psi$  pocket) 13 – 16 nt from either an ACA or H box sequence for the typical two-stem  $\Psi$ -guide RNA structure characterized in other eukaryotes (Ganot, et al. 1997; Ni, et al. 1997). The *Euglena*  $\Psi$ -guide RNAs share this property (distance from their AGA box) with the exceptions of Eg-p7, Eg-p30 and Eg-p37 (17-18 nt, Table A.3.4). It is currently uncertain whether structural differences in *Euglena*  $\Psi$ -guide snoRNP structure may allow for such variation while still efficiently targeting a modification site since we cannot definitively rule out the possibility that other isoforms of these RNAs may exist with more optimal distances that weren't amplified or detected in our library.

The increased number of characterized  $\Psi$ -guide snoRNAs now allows for a more thorough examination of the common structural features of these RNAs (see Table A.3.4 and A.3.5). When considering all *E. gracilis* extended 'single-stem'  $\Psi$ -guide snoRNAs discovered to date, the basal stems vary from 4 – 9 bp in length (6 bp median) and the majority have predicted canonical base-pairing interactions in this region. There also appears to be a preference for higher G-C content in the basal stem (4 G-C bp median), with only a single RNA possessing a basal stem containing <50% G-C content. There are only 7 instances where this stem appears to be interrupted by bulged nucleotides. The average size of the more variable apical stem is 12 bp (range of 7 – 16 bp) and the majority of the identified snoRNAs have at least one mismatch or bulged nucleotide in this region. When the nucleotide changes observed in the various identified isoforms of a  $\Psi$ -guide RNA species are mapped onto its predicted RNA secondary structure, sequence variation is common in the apical stem and predicted to affect secondary structure (Figures 4.4 and A.3.6). Much less frequently do sequence changes occur that alter the structure of the basal

stem. This indicates that sequence and structural variability in the apical stem is likely accommodated in *Euglena*  $\Psi$ -guide snoRNP complexes. This is consistent with what has been observed in yeast, where basal stems are essential for snoRNA accumulation while the apical stems do not contribute as significantly to snoRNA stability (Balakin et al. 1996; Bortolin et al. 1999). Both stems are however essential for the pseudouridylation reaction (Bortolin, et al. 1999). Nucleotide substitutions between isoforms in *Euglena* are also very commonly found in the apical loop region (Figure 4.4) seemingly indicating a lack of strict structural/sequence motifs for snoRNA stability and possibly functionality, in those regions.



**Figure 4.4. Examples of predicted secondary structures of *E. gracilis* AGAUGN box snoRNA species and isoforms.** The boxed nucleotides highlight the AGAUGN box consensus sequence element and arrows indicate sequence variation between characterized isoforms. Nucleotide changes most frequently occur in the stem above the internal single-stranded loop regions that form the pseudouridylation pocket and/or in the apical loop region. Each different color represents the nucleotide changes present in a single isoform species. Black = Eg-p#.1; Grey = Eg-p#.2; Red = Eg-p#.3; Blue = Eg-p#.4; Green = Eg-p#.5. See Figure S6 for additional examples.

In the collection of *Euglena*  $\Psi$ -guide snoRNA sequences, 50% contain a uridine immediately downstream of the AGA box ('AGAUNN') and 32% of the RNAs have the 3' end sequence 'AGAUGN' (Figure A.3.4). We therefore define a new consensus motif for these RNAs and refer to them as AGAUGN box  $\Psi$ -guide RNAs. It is also noteworthy that this sequence resembles the internal portion of the consensus C and C' box sequences (UGAUGA) of the methylation guide box C/D snoRNAs.

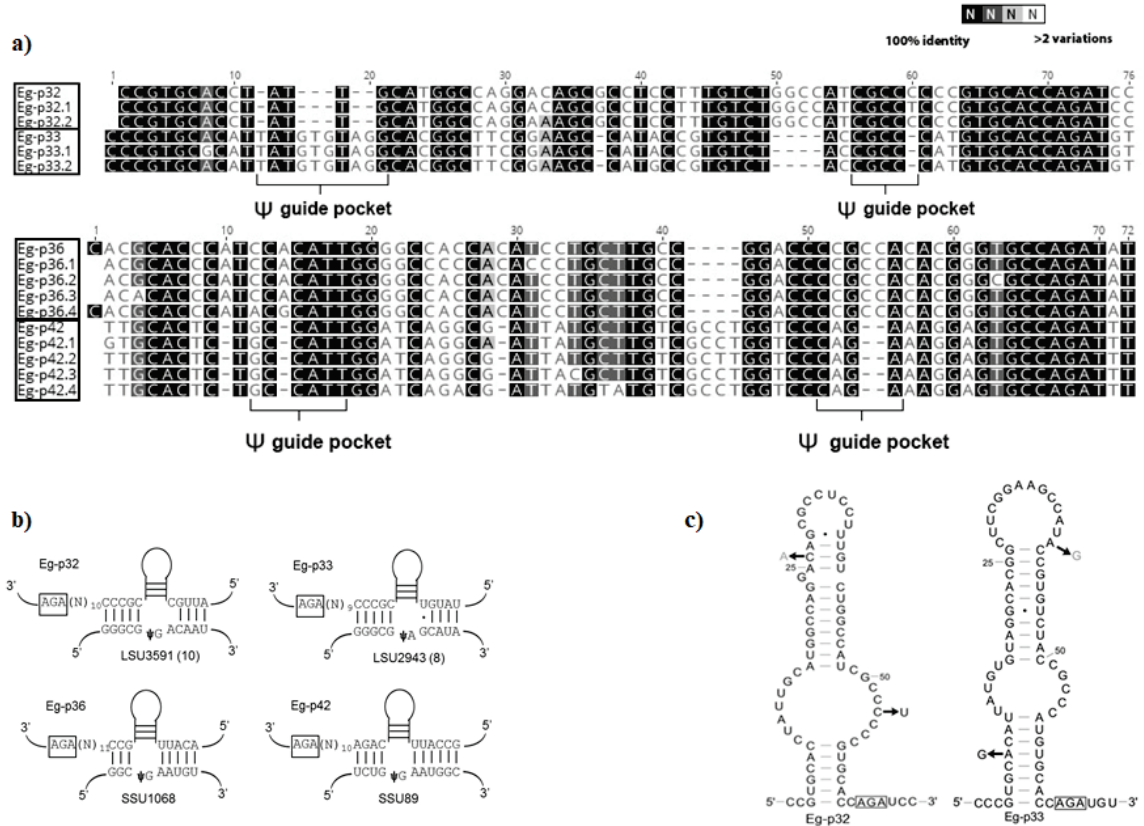
Additional programs such as snoSeeker (Yang et al. 2006), snoGPS (Schattner, et al. 2004), and Psiscan (Myslyuk et al. 2008) have been developed to search for snoRNAs in sequence data libraries. However, they are either based on conserved yeast or mammalian snoRNA features, many of which are not present in *E. gracilis* snoRNAs, or depend on a computational pipeline based on trypanosome sequence analysis and comparison. Some of these programs were tested for their ability to successfully find already characterized *Euglena* snoRNA sequences in the library sequence data, but were largely unsuccessful. Therefore, we found that manually inspecting the initial hits provided through pattern search parameters (see methods) generated the best overall results. This strategy is based on the features of the previously identified snoRNAs in *E. gracilis* and hence it is possible that the remaining "missing" snoRNAs may be significantly structurally different from those already characterized. Perhaps a set of *Euglena*  $\Psi$ -guide RNAs exist which contain multiple extended stem-loop structures. However, as no double-stem H/ACA-like snoRNAs were identified using the software listed above which are optimized to search for such structures, any *Euglena* RNAs of this variety would have to be significantly structurally divergent from other eukaryotic (yeast and human) H/ACA box snoRNAs on which the programs were trained.

Single stem  $\Psi$ -guide snoRNAs possessing conserved AGA sequences have also been identified in trypanosomes, a group of organisms within the Euglenozoans. Analysis of *Leishmania major* (Eliaz, et al. 2015) and *Trypanosoma brucei* (Liang, et al. 2005)  $\Psi$ -guide snoRNAs reveals significant structural similarity to the RNAs identified in our study. Some AGA box  $\Psi$ -guide RNAs in trypanosomes exceeding 100 nucleotides in length have recently been identified but these RNAs appear to still maintain the extended single stem secondary structure (Eliaz, et al. 2015). The genome-wide search for snoRNAs in these trypanosomes found rRNA modification sites with no apparent corresponding snoRNA guide. It has been proposed that these modifications might be carried out by protein-only enzymes (Liang, et al. 2005; Eliaz, et al. 2015). This may also be the case in *E. gracilis* as there still remains a collection of rRNA modifications for which no snoRNA guide has been identified. The unusual rRNA processing pathway could require the cooperative action of both stand-alone enzymes and snoRNA-guided modification complexes.

#### **4.2.5 Evolution of $\Psi$ -guide snoRNA species in *Euglena***

During the analysis of the newly identified AGAUGN box snoRNAs we found a few sequence isoform groups in which the isoform members displayed more sequence variation than is typical for the isoforms of a single  $\Psi$ -guide species. The first pair, Eg-p32 and Eg-p33 were initially clustered in our bioinformatics analysis as a single snoRNA species; however, closer inspection revealed that these two RNAs contain significant changes in their  $\Psi$ -guide pocket regions resulting in the targeting of two different  $\Psi$  modification sites (Figure 4.5a and 4.5b). Similarly another pair, Eg-p36 and Eg-p42, possess a high degree of sequence similarity to each other but contain significant changes in their guide regions (Figure 4.5a and 4.5b). The predicted structural differences that are

evident when comparing pairs such as Eg-p32 and Eg-p33 (Figure 4.5c) further illustrates the previously described properties of this class of *Euglena* ncRNA. This includes the lengthening of the basal stem in Eg-p33 without creating bulges or mismatches, apparent tolerance of a few different mismatches/bulges in the apical stem of isoforms of Eg-p32 and significant sequence and length variation in the apical loops of the pairs. In both cases, these  $\Psi$ -guide snoRNA pairs are evolutionarily-related and the encoding genes have likely evolved by a mechanism similar to that previously characterized for box C/D snoRNA and rRNA target site evolution (Moore and Russell 2012). *Euglena* snoRNA genes are often found within tandemly repeated clusters containing multiple unique snoRNA isoforms of one or both classes (Russell, et al. 2004; Moore and Russell 2012). snoRNA gene duplication followed by sequence divergence within modification guide regions allows the targeting and emergence of new modification sites (Moore and Russell 2012). This mechanism of snoRNA evolution has also been observed in plants (Barneche et al. 2001; Brown et al. 2001), nematodes (Zemann et al. 2006), and trypanosomes (Eliaz, et al. 2015). When comparing  $\Psi$ -guide RNA homologs between different trypanosome species, Eliaz *et al.* observed both snoRNA gene duplication and sequence divergence in the pseudouridylation pockets of the RNAs that allows the targeting of different rRNA nucleotides, similar to what we observe with the *Euglena*  $\Psi$ -guide snoRNA paralogs (see above).



**Figure 4.5. Evolution of *E. gracilis* Ψ-guide snoRNA species.** **a)** Alignment of library cDNA sequences of isoforms of Ψ-guide snoRNAs. Two different sequence isoform clusters were found to have extensive sequence similarity (isoforms of Eg-p32 are similar to Eg-p33, Eg-p36 to Eg-p42) yet display sequence divergence in the regions that form the pseudouridylation pocket thus targeting different rRNA modification sites. Regions containing the nucleotides corresponding to the pseudouridylation guide pocket for each of the RNAs are labeled. Letters on a black background indicate identical nucleotides present at that position in 100% of isoforms, dark gray background indicate 1 isoform differs, light gray indicates 2 isoforms differ, and white background >2 isoforms differ at that position. **b)** Illustration of the predicted base-pairing interactions between the snoRNAs (top) and target rRNA pseudouridylated sites (bottom). Experimentally confirmed pseudouridine sites are indicated as “Ψ”. Within the AGAUGN box elements the highly conserved AGA sequence is highlighted and the number of nucleotides (N) to the base-paired region is indicated. LSU = large subunit rRNA, SSU = small subunit rRNA and the *E. gracilis* LSU “fragment” species where the modification site resides is indicated in parentheses. Full-length snoRNA sequences are shown in Figure S4. **c)** Predicted secondary structures of the Eg-p32 and Eg-p33 pair with nucleotide changes mapped. Black nucleotides = Eg-p#.1; Grey nucleotides = Eg-p#.2.



Multiple isoforms are present for many of the newly identified *E. gracilis* snoRNA species of either class. Up to 9 isoforms were found for a single  $\Psi$ -guide RNA species (Figure A.3.4). This further highlights the high frequency of snoRNA gene duplication and abundance of this class of RNA in *Euglena*. The apparent frequency of snoRNA duplication events is likely the mechanism that allows for the extensive RNA modification in this organism and may be an adaptation to allow for (or even cause) the complex ribosomal biogenesis pathway; both extensive fragmentation and modification.

#### ***4.2.6 Orphan snoRNAs in Euglena gracilis***

In addition to snoRNAs predicted to target rRNA, numerous potential orphan “snoRNAs” were identified (Figure A.3.5). These RNAs are similar in size to rRNA targeting snoRNAs, have canonical sequence box elements and in the case of AGAUGN box RNAs, contain the conserved secondary structural features. However, no mapped modified nucleotide is present in the rRNA in regions that show any limited base-pairing potential to the appropriate regions of these RNAs. Of the newly identified AGAUGN box RNAs found in our library, 8 (20%) were orphans. For box C/D class snoRNAs, 17 (17%) of the newly identified RNAs were considered orphans. In total, orphan “snoRNAs” represent 11% of the combined total of all snoRNA species identified so far in *Euglena*. These orphan RNAs also do not appear to be involved in modifying snRNAs based on examining any base-pairing potential to the mapped *Euglena* snRNA modification sites or other snRNA regions not yet experimentally examined for modifications. Previous analysis of trypanosome snoRNAs also failed to identify any guides for snRNA modifications, with the exception of the spliced-leader RNA (Eliaz, et al. 2015). The apparent scarcity of



snRNA modification-guide RNAs in Euglenozoa may indicate that these modifications are performed by stand-alone protein enzymes or guided by structurally novel RNAs.

While we do not know the function of these orphan *Euglena* RNAs, the discovery of pseudouridylation of mRNA species in yeast and humans (Carlile, et al. 2014; Lovejoy et al. 2014; Schwartz, et al. 2014) raises the possibility that these orphans may be used to target modification sites in mRNA or other cellular RNA species. Alternatively, they could be involved in *Euglena*'s unique rRNA processing/assembly pathway, similar to what has been observed in some trypanosomes, or be processed into other types of ncRNA (Eliaz, et al. 2015; Chikne, et al. 2019). Our method for identification of box C/D snoRNAs relied primarily on the presence of the consensus sequence box elements at expected positions relative to RNA 5' and 3' ends and searches for significant base-pairing interactions to modified sites in *E. gracilis* rRNA. Comprehensive identification of all potential box C/D RNAs in the library sequences is particularly challenging for orphans (without rRNA base-pairing) because of the short length of C and D box elements, the even more sequence degenerate C' and D' boxes, and the absence of any obvious extended secondary structural conservation in this class of *Euglena* snoRNA. Consequently, we are likely underestimating the abundance of “snoRNAs” that target modification or perform other functions on non-rRNA species. The conserved ‘AGAUGN’ box and secondary structural features makes it somewhat easier to identify orphan RNAs within this other “snoRNA” class. Recent publication of a draft *Euglena* genome estimates only 1% of the genome is coding, but is predicted to be only 20% assembled and did not analyze ncRNAs (Ebenezer, et al. 2019). This further suggests there is likely a large number of uncharacterized ncRNAs remaining

in the *E. gracilis* genome and highlights the need for distinct ncRNA analysis to detect these unusual RNAs.

#### **4.2.7 tRNA identification**

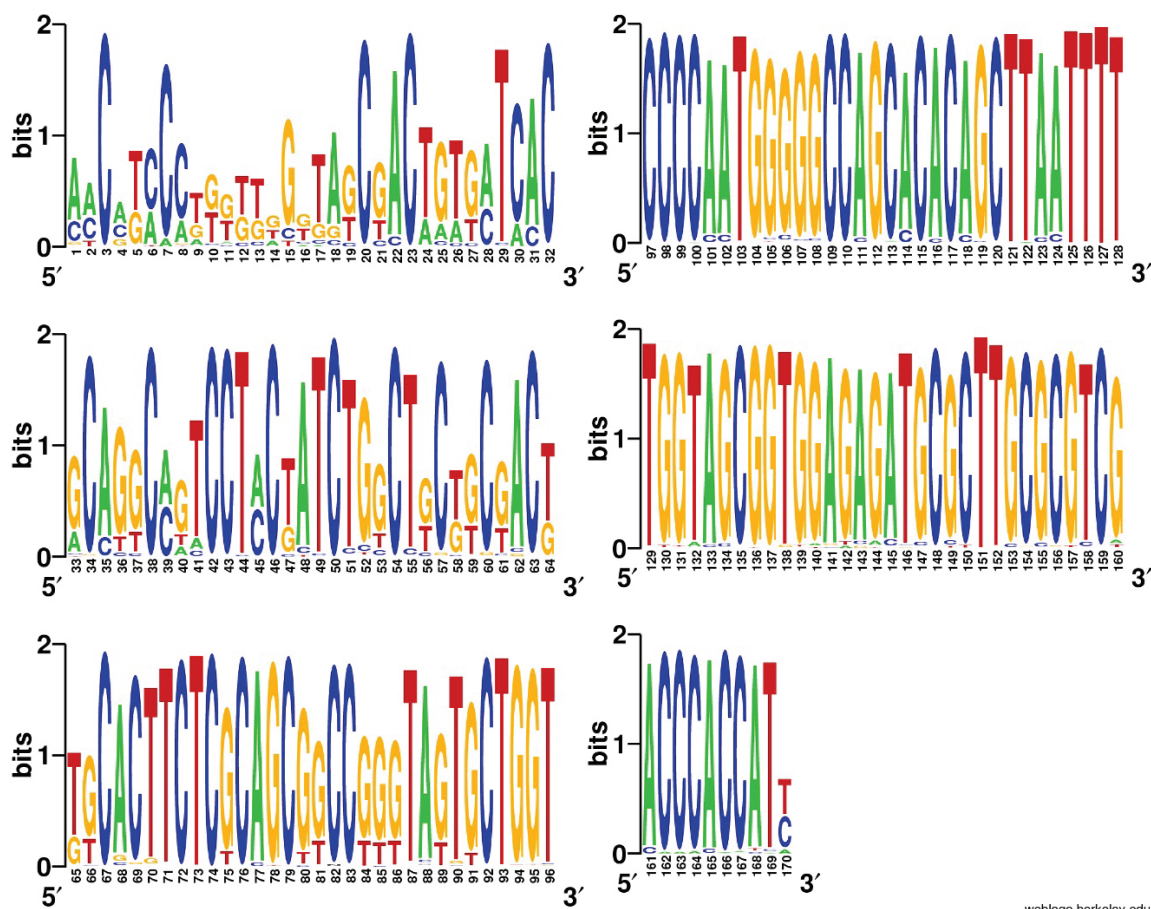
As a result of sparse genomic sequence information very few *E. gracilis* nuclear tRNAs have been identified to date. Using the tRNA identification program ARAGORN (Laslett and Bjorn 2004) to detect library sequences with requisite conserved primary and secondary structural potential, we have identified 14 tRNAs in our library which do not map onto the complete *E. gracilis* chloroplast genome sequence (Hallick et al. 1993). Analysis of the *E. gracilis* mitochondrial genome failed to identify any tRNA coding genes, suggesting that these 14 tRNAs are encoded in the nuclear genome. We cannot rule out that some of these tRNAs may function in the mitochondria (Dobáková et al. 2015). These tRNAs include single tRNA isoacceptors for Met (CAU), His (GUG), Ser (AGA), Cys (GCA), Leu (CAG), and multiple isoacceptors for Glu (CUC and UUC), Gln (UUG and CUG), Pro (UGG and CGG), and Ala (AGC, CGC, and UGC) (Figure A.3.7). While some of these identified sequences possess mature CCA 3' ends, a large collection of the tRNA reads are precursors which contain additional sequence at their 5' and 3' ends. An abundance of nucleoside modifications in *E. gracilis* tRNAs likely explains the higher representation of precursors and the incomplete representation of all tRNA species in the library as many tRNA modifications block reverse transcriptases from extending beyond the modified site, requiring additional enzymatic treatments of the RNA prior to RT in order to obtain full-length cDNAs (Wilusz 2015). The limitation of sequence size for Illumina sequencing may also explain the absence of tRNA precursors containing introns from being identified in the library. Correct identification of the new *Euglena* tRNAs was

further confirmed through alignment to the tRNA isoacceptors in *Trypanosoma* and *Leshmania* species. This showed close sequence similarity of tRNAs between all these species (Tessier et al. 1991). We found performing PCR based assays like those used to examine snoRNA repeats shows at least some tRNAs are also found in genomic repeats (data not shown). This further highlights the degree of repetition found for nearly all ncRNA classes in *E. gracilis*.

#### **4.2.8 U1 snRNA sequences with variable 5' ends**

During analysis of our small RNA library, we found a collection of 836 unique U1 snRNA sequences that each differ from the previously published *E. gracilis* U1 snRNA sequence (Breckenridge, et al. 1999). Examination of these variant sequences revealed the sequence changes were heavily biased towards the 5' end of the RNA, the region responsible for the binding of the 5' splice site of spliceosomal introns (Figure 4.6). As described in Chapter 1, in addition to conventional spliceosomal introns, *E. gracilis* contains non-conventional mRNA introns that do not conform to the typical spliceosomal 5' and 3' splice site sequences. The splicing components and mechanism involved in removal of *Euglena* non-conventional introns is currently unknown (McWatters and Russell 2017). It is intriguing to consider the possibility that *E. gracilis* could employ a large collection of variant U1s to recognize the diverse splice boundaries of these non-conventional introns. However, only a single gene locus had been detected that contains the previously characterized U1 snRNA (Breckenridge, et al. 1999) and our attempts to PCR amplify genomic repeats of U1 snRNAs (assuming they would exist in such an arrangement) by designing oligonucleotides targeting several of the variants did not generate detectable PCR amplicons (data not shown).

These variant U1 RNAs are often represented as single reads in the RNA-Seq libraries. It is possible these are artifacts of the library construction method, but it is unclear how they would have been generated as it seems to be a phenomenon restricted to the U1 sequences (the other snRNAs in the library do not display this) and often times only a few nucleotides differ from the conventional sequence in any given U1 RNA sequence read. Variation could also be the result of misincorporation (mismatching) of nucleotides during reverse transcription caused by U1 modified nucleotide or differentially edited sites, but this would indicate an incredibly high density of modifications in U1 that had not previously been detected (Breckenridge, et al. 1999) and these sites are also heavily biased towards the 5' third of the U1 sequence (Figure 4.6). This unusual finding clearly warrants further investigation as it could functionally affect the splicing pathways and additional deeper sequencing of *E. gracilis* ncRNAs and co-precipitation experiments targeting U1 complexed proteins will be possible first steps towards further exploration.



weblogo.berkeley.edu

**Figure 4.6. Sequence logo of variant *E. gracilis* U1 sequences.** The logo was prepared from 836 variant U1 snRNA sequences identified when combining the two *E. gracilis* small RNA libraries. Nucleotides are numbered based on positioning relative to the primary U1 isoform previously identified (Breckenridge, et al. 1999). Individual nucleotide height is proportional to frequency of appearance at that position with a bit score of 2 indicating 100% conservation. Generated using the WebLogo webserver.

## 4.3 Materials and Methods

### 4.3.1 Library construction

*E. gracilis* total RNA (~112 µg) was resolved on a 15% denaturing polyacrylamide gel and RNA fragments less than 400 nt in size were excised and isolated (Sambrook and Russell 2001). A poly-G tail was added to the 3' ends of the size-selected RNA (Rederstorff and Huttenhofer 2011). The tailing reaction contained size-selected or TMG cap-enriched

*Euglena* RNA, 1X Poly(A) Polymerase (PAP) buffer (USB), 0.5 mM GTP, 60 U of yeast PAP (USB) and 20 U of RNase Inhibitor (NEB) incubated at 37°C for 60 min. The reaction was extracted once with phenol:chloroform (1:1), then twice with chloroform and the aqueous phase was ethanol precipitated with added acrylamide carrier. The RNA was then treated with 10 U of Tobacco Acid Pyrophosphatase (TAP) (epicentre®) in a 10 µL reaction containing 1X TAP buffer (epicentre®) and 20 U of RNase Inhibitor (NEB) at 37°C for 60 min and the RNA was extracted and precipitated (as above).

An RNA oligonucleotide linker was ligated to the 5' termini of the TAP-treated RNA (Rederstorff and Huttenhofer 2011). The RNA was first mixed with 200 pmol of linker and incubated at 65°C for 5 min. The ligation reaction containing 10 U of T4 RNA ligase (NEB), 1 mM ATP, 1X T4 RNA ligase buffer (NEB), and 20 U of RNase Inhibitor (NEB) was then performed at 4°C overnight (16 hrs), after which another 10 U of T4 RNA ligase was added and the reaction further incubated at 37°C for 30 min. The RNA was then extracted and precipitated.

An antisense primer containing an adaptor sequence and poly C stretch was designed to anneal to the 3' poly-G tail. This primer (100 pmole) was incubated with 10 µL of prepared RNA from the previous step and dNTPs (500 µM) at 65°C for 5 min and then immediately chilled on ice. Superscript II RT (Invitrogen) was used to synthesize cDNA at 47°C for 60 min following the manufacturer's protocol. The cDNA was then used as template for PCR amplification with Phusion Taq Polymerase (Thermo Scientific) using oligonucleotides designed to anneal to the 3' poly-G tail and the 5' linker sequence with or without the addition of blocking primers (also see below and Table A.3.1 for oligonucleotide sequences; Table A.3.6 for PCR conditions). When assessing relative

levels of rRNA (and not employing the blocking primers), PCR products were purified by gel-extraction and cloned into the pJET1.2/blunt vector following the manufacturer's protocol. Transformed *E. coli* cells were then used for colony PCR screening, using primers that anneal upstream and downstream of the cloning site. Automated DNA sequencing of these PCR product clones was performed by MacroGen Corp USA.

#### ***4.3.2 Preventing amplification of large subunit rRNA fragments***

Blocking primer sets were designed with a C3 spacer (3 hydrocarbon) modification at their 3' end and each of these modified primers anneals both to the 3' end of the added 5' linker sequence and to the 5' end of a specifically-targeted individual LSU rRNA fragment (see Table A.3.2 for blocker oligonucleotide sequences). At the PCR amplification step of library preparation, in addition to the general amplification primers, each blocking primer was also added to the reaction to a concentration of 5 pmole/ $\mu$ L to prevent amplification of the unwanted rRNA species. The final resulting PCR-generated cDNA library was purified using the E.Z.N.A.® Cycle-Pure Kit (Omega) and sent to Genome Québec for high-throughput sequencing using the Illumina MiSeq 250 platform.

#### ***4.3.3 Bioinformatic analysis***

The Illumina MiSeq sequence reads, for both size-selected and TMG-enriched libraries, were first sorted based on the presence of the 5' linker sequence using the FASTQ Barcode Splitter tool from the FASTX-Toolkit (Hannon Lab website). The 5' and 3' adaptor sequences were then removed (allowing 2 mismatches) using the Trim Ends tool in Geneious v8.0.4 software and cutadapt software package. Typically, the sequence quality was very poor following the 3' poly G tract and therefore the 3' ends were trimmed downstream of a poly G tract  $\geq 12$  nt long. The two most highly abundant sequences were

also removed from the collection. The UCLUST algorithm, a component of the USearch (Edgar 2010) software package, was used to cluster related sequences together based on pair-wise alignments, using an identity threshold of 0.8. To remove previously characterized *Euglena* RNAs from the newly formed sequence clusters, databases of *E. gracilis* snoRNAs, rRNA, snRNAs and tRNAs were created. First, the UBLAST algorithm (Edgar 2010) was used to find matches between the database and the RNA-Seq library sequences using an E-value of  $1e-9$ . Then to ensure removal of as many sequences as possible, searches using the USearch global algorithm were performed with an id value of 1. Matches to these databases were subsequently removed prior to library analysis.

Two approaches were used to identify new snoRNAs. First, trimmed sequences between 50 – 80 nt in length were extracted (using Geneious) and then scanned for *E. gracilis* snoRNA features using the pattern matching program ‘Scan for Matches’ (Dsouza et al. 1997). A consensus pattern was created based on all previously identified *Euglena* snoRNAs including size, sequence box elements, and secondary structure potential. To identify additional box C/D snoRNAs, trimmed sequences between 55 – 90 nt were analyzed using the Snotscan webserver (Lowe and Eddy 1999) with *E. gracilis* rRNA sequence, including internal transcribed spacer sequences, as potential modification targets.

Positive hits from both approaches were then further manually inspected as previously described (Moore and Russell 2012). Sequences that strictly maintained conserved features of snoRNAs but did not display significant base-pairing potential to any mapped modified rRNA site were sorted into the orphan ‘snoRNA’ category. Reads per million (RPM) for snoRNA species were calculated using quality filtered single end reads from size-selected and TMG-capped libraries. Individual RNAs were quantified using USearch algorithm searches with an id value of 0.95, then normalized for library size.



For tRNA identification, the USearch algorithm was used to BLAST characterized *T. brucei* tRNAs against our trimmed and dereplicated library. Potential candidates from the library were then further analyzed using the ARAGORN webserver (Laslett and Bjorn 2004) to look for conserved sequence and structural elements indicative of tRNAs.

## Chapter 5: Conclusions and Future Directions

In this thesis study, I examined ncRNA and RNP repertoires in the diplomonads *G. lamblia* and *G. muris*, and the Euglenozoan *E. gracilis*, leading to a number of novel insights regarding ncRNA function and evolution. These insights include: i) the presence of an RNase P-snoRNA diRNP in both *Giardia* species capable of targeting a conserved modification in elongator tRNA<sup>Met</sup>; ii) structurally reduced telomerase and U3 RNAs in *Giardia*; iii) two unique diplomonad Snu13p homologs involved in various complexes that may require secondary proteins for stable K-turn binding *in vivo* and iv) a large collection of *E. gracilis* snoRNAs that shed light on Ψ-guide snoRNA structure, evolution, and the timing of pre-rRNA processing.

In Chapters 2 and 3, I describe a novel RNase P-snoRNA diRNP from the species *G. lamblia* and *G. muris*, that contains at least one of the Snu13p homologs. The snoRNA domain of the diRNP is capable of targeting elongator tRNA<sup>Met</sup> for 2'-O-methylation. The snoRNA-tRNA base-pairing is maintained in both species, supporting this proposed function; however, the methylation status of the target tRNA<sup>Met</sup> nucleotide is yet to be experimentally examined and reliance of this modification on the Glr15/GmsR15 domain will also be required. This could be determined through primer extension-based methylation mapping assays and anti-sense/morpholino hybridization to the snoRNA guide region to prevent guide-target base-pairing. Beyond this, obtaining structures of the diRNP in the presence and absence of the presumed substrates, using techniques like cryo-EM, will be important in understanding the mechanism by which it can perform both functions. This will be particularly interesting for the diRNP in complex with pre-tRNA<sup>Met</sup> which would be a substrate for both 5' end processing and 2'-O-methylation by RNase P and

GlsR15/GmsR15 respectively. The ability to further investigate the RNase P-snoRNA diRNP is complicated by the difficulty in initial attempts to generate stable transfectants containing tagged RNase P proteins in *G. lamblia* (data not shown) and may require alternative strategies such as the production of antibodies against RNase P proteins that target the native complex to facilitate this type of analysis.

In Chapter 3, two Snu13p homologs were described in diplomonads and shown to be components of C/D snoRNPs in *G. lamblia* by using protein and RNA co-precipitation experiments. Analyzing the role of the *G. lamblia* Snu13p homologs found that the presence of additional proteins including Nop56, Nop58, and Rrp9 may be required for stable K-turn binding and RNP assembly *in vivo*. We were not able to definitively determine if the two homologs play any distinct roles in the cell as we had initially hypothesized, but my data suggests they could associate differently with the U3 snoRNP and RNase P-snoRNA diRNP. In the case of U3, *in vitro* binding assays of each homolog to the two unique K-turns present in U3 (C'/D and B/C) will be a logical next step is assessing these potential differences. For the diRNP, co-precipitations targeting RNase P specific proteins should reveal whether one or both Snu13p homologs associate with the complex. Additionally, cross-linking based RNA-seq techniques like CLIP-seq could be applied to specifically localize the regions of RNA bound by the two homologs (Hafner et al. 2021). Cross-linking based protein analysis could also be utilized to enhance our protein-protein interaction analysis. This would address the potential association between the U3 specific protein Rrp9 or RNase P protein components and the Snu13p homologs and could be used to probe the association of the Ucp1-3 proteins with each other, core ribosome components, and endoplasmic reticulum and nuclear envelope proteins under more

stringent conditions. Validation of these interactions could then be done using yeast two-hybrid (Y2H) assays to assess direct protein-protein interactions.

Developing a system to obtain gene knockouts, such as CRISPR-Cas9, has been slow in *G. lamblia* as the use of genetic tools like this are complicated by the presence of its two nuclei. Recently, methodologies were developed to allow CRISPR interference (CRISPRi) based knockdowns in *G. lamblia* which are promising for the study of gene function (McInally et al. 2019; Jex et al. 2020). Use of this CRISPRi tool to knockdown the two Snul3p homologs individually and assess the effects on cell viability and growth, along with the impact on various Snul3p containing complexes and pre-rRNA processing would also help unravel differences in their cellular roles.

In Chapter 2, homologs for four snRNAs, U1, U2, U4, and U6 were detected in *G. muris*. U1 has diverged significantly from the U1 of *G. lamblia* showing the relatively rapid evolution within diplomonads of normally well-conserved eukaryotic RNAs. Our inability to detect a strong candidate for U5 in either examined *Giardia* species, despite having characterized what appears to be a large proportion of all ncRNAs in these species, could mean that a functional homolog of U5 is truly not present in *Giardia*. Additional support for this hypothesis comes from the absence of Prp31 and Prp6 homologs, proteins that facilitate the interaction of U4/U6 with U5 snRNPs in humans and yeast (Makarova et al. 2002; Nguyen, et al. 2016). While a handful of U5 proteins are present in *Giardia* including a Prp8 homolog, these could associate with the splicing machinery through other contacts independent of the U5 snRNA. This would be novel due to the importance of U5 in anchoring the 5' exon during splicing but could conceivably be achieved through protein contacts. Purification of a (tagged) conserved U5 protein or the structurally non-canonical

candidate U5 snRNA will be required to draw more definitive conclusions. The apparent absence of both Snu13p homologs from the U4 snRNP (Chapter 3) adds to the list of unusual features of the U4/U6•U5 tri-snRNP complex and may require additional validation.

The *G. lamblia* 3' ncRNA processing motif is conserved in all *G. muris* ncRNAs described so far and the motif consensus sequences are nearly identical in the two species (Chapter 2). This indicates a strong selective pressure to maintain this processing pathway and suggests it could be conserved throughout all *Giardia* species. Despite its importance, we currently have little understanding of the mechanism by which the motif is processed, or the machinery involved. It is unclear as to what could have driven the evolution of a ncRNA processing pathway shared between so many distinct ncRNA classes. Intriguingly, a recent structural analysis of the human RNase MRP complex determined a consensus target sequence of 5'-\*RCRC-3' (\* denotes cleavage site) and previous biochemical analysis of RNase MRP substrates in yeast found 5'-\*CUC-3' to be the most common cleavage site (Esakova et al. 2011; Lan, et al. 2020). RNase MRP is primarily known for targeting pre-rRNA containing the sequence 5'-\*ACACAA-3', but is also found to target other RNA species for cleavage. These sites resemble the *Giardia* 3' processing motif (most commonly 5'-UCCUUU\*ACUCAA-3'), and RNase MRP cleaves its targets at the same relative position described for cleavage of the 3' motif (Hudson 2014). It could be that the RNase MRP complex evolved this specialize role in *Giardia* species. The presence of the processing motif downstream of the 28S rRNA operon in *G. muris* may also indicate RNase MRP could further contribute to pre-rRNA processing through cleavage of the motif as

well. Future analysis should explore the possible involvement of the *Giardia* RNase MRP complex in 3' motif processing.

We described 113 new snoRNAs from *E. gracilis* (Chapter 4) allowing us to characterize conserved features of their unique  $\Psi$ -guide RNAs and their evolution. While we have identified many modification-guide snoRNAs, the mechanism by which the 16 ITS sequences are processed out of the *E. gracilis* pre-rRNA transcript is still unclear. The description of at least 15 snoRNAs in trypanosomes involved in pre-rRNA processing suggests there may be a large collection of snoRNAs also involved in removal of these novel ITS regions in *Euglena* (Chikne, et al. 2019). This could include previously identified snoRNAs, or the orphan snoRNAs described in Chapter 4. With the reduction in cost and increased efficiency of RNA-seq technologies, larger scale small RNA libraries produced using a thermostable RT like TGIRT could help identify additional ncRNAs previously not readily amplified due to the stability of their secondary structure, including the remaining snoRNAs. Further development of a procedure for snoRNA knockdowns in *E. gracilis* like those used in *T. brucei* will also be paramount in unraveling the role of snoRNAs in particular processing events.

Finally, the appearance of 836 variant U1 snRNAs in our small RNA library could represent a unique solution to splicing of the *Euglena* non-conventional introns. However, these RNA sequences are typically represented by very few RNA-seq reads and I was unsuccessful in detecting genomic repeats containing the variant U1 sequences. I have produced polyclonal antibodies against the predicated *E. gracilis* U1A protein and optimized RNA co-precipitation conditions. If the variant U1 snRNAs are genuinely present, expressed, and involved in splicing they are predicted to be associated with U1A

based on their sequence and structural features. U1A co-precipitations may give insights into non-conventional splicing regardless of variant U1 function if the spliceosome is involved. Additional experiments including cross-linking, ligation, and sequencing of hybrids (CLASH) or psoralen analysis of RNA interactions and structures (PARIS) may also be valuable in assessing the RNAs associated with the non-conventional introns (Kudla et al. 2011).

The totality of my findings highlight the value in studying diverse sets of species. This type of inquiry can lead to a deeper understanding of cellular complexes, even those well-studied in model species, and uncover novel innovations that have evolved under diverse conditions and novel selective pressures. Comparative analysis of ncRNPs and other complexes in a greater variety and number of species will undoubtedly greatly expand our knowledge of the type of variation that is possible in nature and help uncover the most core and ancestral features of these molecular machines.

## References

- Abel Y, Paiva ACF, Bizarro J, Chagot M-E, Santo Paulo E, Robert M-C, Quinternet M, Vandermoere F, Sousa PMF, Fort P, et al.** 2021. NOPCHAP1 is a PAQosome cofactor that helps loading NOP58 on RUVBL1/2 during box C/D snoRNP biogenesis. *Nucleic Acids Res* **49**:1094-1113.
- Adam RD.** 2001. Biology of *Giardia lamblia*. *Clin Microbiol Rev* **14**:447-475.
- Adam RD, Dahlstrom EW, Martens CA, Bruno DP, Barbian KD, Ricklefs SM, Hernandez MM, Narla NP, Patel RB, Porcella SF, et al.** 2013. Genome sequencing of *Giardia lamblia* genotypes A2 and B isolates (DH and GS) and comparative analysis with the genomes of genotypes A1 and E (WB and Pig). *Genome Biol Evol* **5**:2498-2511.
- Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, Agatha S, Berney C, Brown MW, Burki F, et al.** 2019. Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *J Eukaryot Microbiol* **66**:4-119.
- Akopian D, Shen K, Zhang X, Shan S.** 2013. Signal Recognition Particle: An Essential Protein-Targeting Machine. *Annu Rev Biochem* **82**:693-721.
- Altman S.** 2011. Ribonuclease P. *Philos Trans R Soc Lond B Biol Sci* **366**:2936-2941.
- Andersson JO, Sjögren ÅM, Horner DS, Murphy CA, Dyal PL, Svärd SG, Logsdon JM, Ragan MA, Hirt RP, Roger AJ.** 2007. A genomic survey of the fish parasite *Spironucleus salmonicida* indicates genomic plasticity among diplomonads and significant lateral gene transfer in eukaryote genome evolution. *BMC Genomics* **8**:51.
- Ankarklev J, Franzén O, Peirasmaki D, Jerlström-Hultqvist J, Lebbad M, Andersson J, Andersson B, Svärd SG.** 2015. Comparative genomic analyses of freshly isolated *Giardia intestinalis* assemblage A isolates. *BMC Genomics* **16**:697.
- Archibald JM.** 2007. Nucleomorph genomes: structure, function, origin and evolution. *BioEssays* **29**:392-402.
- Åsman AKM, Curtis BA, Archibald JM.** 2019. Nucleomorph Small RNAs in Cryptophyte and Chlorarachniophyte Algae. *Genome Biol Evol* **11**:1117-1134.
- Atzorn V, Fragapane P, 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.
- Aulds J, Wierzbicki S, McNairn A, Schmitt ME.** 2012. Global Identification of New Substrates for the Yeast Endoribonuclease, RNase Mitochondrial RNA Processing (MRP) *J Biol Chem* **287**:37089-37097.
- Avery OT, Macleod CM, McCarty M.** 1944. Studies on the chemical nature of the substance inducing transformation of *Pneumococcal* types : induction of transformation by



a desoxyribonucleic acid fraction isolated from *Pneumococcus* type III. *J Exp Med* **79**:137-158.

**Ayadi L, Galvanin A, Pichot F, Marchand V, Motorin Y.** 2019. RNA ribose methylation (2'-O-methylation): Occurrence, biosynthesis and biological functions. *Biochim Biophys Acta Gene Regul Mech* **1862**:253-269.

**Baker DL, Youssef OA, Chastkofsky MI, Dy DA, Terns RM, Terns MP.** 2005. RNA-guided RNA modification: functional organization of the archaeal H/ACA RNP. *Genes Dev* **19**:1238-1248.

**Balakin AG, Smith L, Fournier MJ.** 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.

**Baldini L, Charpentier B, Labialle S.** 2021. Emerging Data on the Diversity of Molecular Mechanisms Involving C/D snoRNAs. *Non-Coding RNA* **7**:30.

**Ban N, Nissen P, Hansen J, Moore PB, Steitz TA.** 2000. The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. *Science* **289**:905-920.

**Barandun J, Chaker-Margot M, Hunziker M, Molloy KR, Chait BT, Klinge S.** 2017. The complete structure of the small-subunit processome. *Nat Struct Mol Biol* **24**:944-953.

**Barneche F, Gaspin C, Guyot R, 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.

**Baserga S, Steitz J.** 1993. The Diverse World of Small Ribonucleoproteins. *Cold Spring Harbor M Arch* **24**:359-381.

**Baudin-Baillieu A, Fabret C, Liang X-h, Piekna-Przybylska D, Fournier MJ, 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.

**Beltrame M, Tollervey D.** 1995. Base pairing between U3 and the pre-ribosomal RNA is required for 18S rRNA synthesis. *EMBO J* **14**:4350-4356.

**Bertrand E, Houser-Scott F, Kendall A, Singer RH, Engelke DR.** 1998. Nucleolar localization of early tRNA processing. *Genes Dev* **12**:2463-2468.

**Best AA, Morrison HG, McArthur AG, Sogin ML, Olsen GJ.** 2004. Evolution of eukaryotic transcription: insights from the genome of *Giardia lamblia*. *Genome Res* **14**:1537-1547.

**Biswas S, Buhrman G, Gagnon K, Mattos C, Brown BA, Maxwell ES.** 2011. Comparative Analysis of the 15.5kD Box C/D snoRNP Core Protein in the Primitive

Eukaryote *Giardia lamblia* Reveals Unique Structural and Functional Features. *Biochemistry* **50**:2907-2918.

**Bizarro J, Charron C, Boulon S, Westman B, Pradet-Balade B, Vandermoere F, Chagot M-E, Hallais M, Ahmad Y, Leonhardt H, et al.** 2014. Proteomic and 3D structure analyses highlight the C/D box snoRNP assembly mechanism and its control. *J Cell Biol* **207**:463-480.

**Bizarro J, Dodré M, Huttin A, Charpentier B, Schlotter F, Branlant C, Verheggen C, Massenet S, Bertrand E.** 2015. NUFIP and the HSP90/R2TP chaperone bind the SMN complex and facilitate assembly of U4-specific proteins. *Nucleic Acids Res* **43**:8973-8989.

**Boivin V, Deschamps-Francoeur G, Couture S, Nottingham RM, Bouchard-Bourelle P, Lambowitz AM, Scott MS, Abou-Elela S.** 2018. Simultaneous sequencing of coding and noncoding RNA reveals a human transcriptome dominated by a small number of highly expressed noncoding genes. *RNA* **24**:950-965.

**Borovjagin AV, Gerbi SA.** 1999. U3 small nucleolar RNA is essential for cleavage at sites 1, 2 and 3 in pre-rRNA and determines which rRNA processing pathway is taken in *Xenopus* oocytes. *J Mol Biol* **286**:1347-1363.

**Bortolin ML, Ganot P, 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.

**Boulon Sv, Marmier-Gourrier N, Pradet-Balade Brr, Wurth L, Verheggen Cl, Jádý BtE, Rothé B, Pescia C, Robert M-Cc, Kiss Ts, et al.** 2008. The Hsp90 chaperone controls the biogenesis of L7Ae RNPs through conserved machinery. *J Cell Biol* **180**:579-595.

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

**Breckenridge DG, Watanabe Y-i, Greenwood SJ, Gray MW, Schnare MN.** 1999. U1 small nuclear RNA and spliceosomal introns in *Euglena gracilis*. *Proc Natl Acad Sci U S A* **96**:852-856.

**Bringmann P, Appel B, Rinke J, Reuter R, Theissen H, Lührmann R.** 1984. Evidence for the existence of snRNAs U4 and U6 in a single ribonucleoprotein complex and for their association by intermolecular base pairing. *EMBO J* **3**:1357-1363.

**Brow DA, Guthrie C.** 1988. Spliceosomal RNA U6 is remarkably conserved from yeast to mammals. *Nature* **334**:213-218.

**Brown JW, Clark GP, Leader DJ, Simpson CG, Lowe T.** 2001. Multiple snoRNA gene clusters from Arabidopsis. *RNA* **7**:1817-1832.

- Burki F, Roger AJ, Brown MW, Simpson AGB.** 2020. The New Tree of Eukaryotes. *Trends Ecol Evol* **35**:43-55.
- Cacciò SM, Lalle M, Svärd SG.** 2018. Host specificity in the *Giardia duodenalis* species complex. *Infect Genet Evol* **66**:335-345.
- Cahill NM, Friend K, Speckmann W, Li Z-H, Terns RM, Terns MP, Steitz JA.** 2002. Site-specific cross-linking analyses reveal an asymmetric protein distribution for a box C/D snoRNP. *EMBO J* **21**:3816-3828.
- Carlile TM, Rojas-Duran MF, Zinshteyn B, Shin H, Bartoli KM, Gilbert WV.** 2014. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature* **515**:143-146.
- Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, Zhao Q, Wortman JR, Bidwell SL, Alsmark UCM, Besteiro S, et al.** 2007. Draft Genome Sequence of the Sexually Transmitted Pathogen *Trichomonas vaginalis*. *Science* **315**:207-212.
- Caton EA, Kelly EK, Kamalampeta R, Kothe U.** 2017. Efficient RNA pseudouridylation by eukaryotic H/ACA ribonucleoproteins requires high affinity binding and correct positioning of guide RNA. *Nucleic Acids Res* **46**:905-916.
- Chagot M-E, Quinternet M, Rothé B, Charpentier B, Coutant J, Manival X, Lebars I.** 2019. The yeast C/D box snoRNA U14 adopts a “weak” K-turn like conformation recognized by the Snu13 core protein in solution. *Biochimie* **164**:70-82.
- Charette JM, Gray MW.** 2009. U3 snoRNA genes are multi-copy and frequently linked to U5 snRNA genes in *Euglena gracilis*. *BMC Genomics* **10**:528.
- Chen J-L, Greider CW.** 2003. Template boundary definition in mammalian telomerase. *Genes Dev* **17**:2747-2752.
- Chen X, Collins LJ, Biggs PJ, Penny D.** 2009. High Throughput Genome-Wide Survey of Small RNAs from the Parasitic Protists *Giardia intestinalis* and *Trichomonas vaginalis*. *Genome Biol Evol* **1**:165-175.
- Chen X, Rozhdestvensky TS, Collins LJ, Schmitz J, Penny D.** 2007. Combined experimental and computational approach to identify non-protein-coding RNAs in the deep-branching eukaryote *Giardia intestinalis*. *Nucleic Acids Res* **35**:4619-4628.
- Chen XS, Penny D, Collins LJ.** 2011. Characterization of RNase MRP RNA and novel snoRNAs from *Giardia intestinalis* and *Trichomonas vaginalis*. *BMC Genomics* **12**:550.
- Chen Z, Duan X.** 2011. Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. *Methods Mol Biol* **733**:93-103.
- Chikne V, Shanmugha Rajan K, Shalev-Benami M, Decker K, Cohen-Chalamish S, Madmoni H, Biswas VK, Kumar Gupta S, Doniger T, Unger R, et al.** 2019. Small

nucleolar RNAs controlling rRNA processing in *Trypanosoma brucei*. *Nucleic Acids Res* **47**:2609-2629.

**Cho I-M, Lai LB, Susanti D, Mukhopadhyay B, Gopalan V.** 2010. Ribosomal protein L7Ae is a subunit of archaeal RNase P. *Proc Natl Acad Sci U S A* **107**:14573-14578.

**Chu S, Archer RH, Zengel JM, Lindahl L.** 1994. The RNA of RNase MRP is required for normal processing of ribosomal RNA. *Proc Natl Acad Sci U S A* **91**:659-663.

**Clerget G, Bourguignon-Igel V, Marmier-Gourrier N, Rolland N, Wacheul L, Manival X, Charron C, Kufel J, Méreau A, Senty-Ségault V, et al.** 2020. Synergistic defects in pre-rRNA processing from mutations in the U3-specific protein Rrp9 and U3 snoRNA. *Nucleic Acids Res* **48**:3848-3868.

**Cléry A, Senty-Ségault V, Leclerc F, Raué HA, Branlant C.** 2007. Analysis of Sequence and Structural Features That Identify the B/C Motif of U3 Small Nucleolar RNA as the Recognition Site for the Snu13p-Rrp9p Protein Pair. *Mol Cell Biol* **27**:1191-1206.

**Cloutier P, Poitras C, Durand M, Hekmat O, Fiola-Masson É, Bouchard A, Faubert D, Chabot B, Coulombe B.** 2017. R2TP/Prefoldin-like component RUVBL1/RUVBL2 directly interacts with ZNHIT2 to regulate assembly of U5 small nuclear ribonucleoprotein. *Nat Commun* **8**:15615.

**Collins K.** 2006. The biogenesis and regulation of telomerase holoenzymes. *Nat Rev Mol Cell Biol* **7**:484-494.

**Cordingley JS, Turner MJ.** 1980. 6.5 S RNA; preliminary characterisation of unusual small RNAs in *Trypanosoma brucei*. *Mol Biochem Parasitol* **1**:91-96.

**Crick F.** 1970. Central Dogma of Molecular Biology. *Nature* **227**:561-563.

**Crick FH.** 1958. On protein synthesis. *Symp Soc Exp Biol* **12**:138-163.

**Crooks GE, Hon G, Chandonia JM, Brenner SE.** 2004. WebLogo: a sequence logo generator. *Genome Res* **14**:1188-1190.

**Crowell AMJ, Wall MJ, Doucette AA.** 2013. Maximizing recovery of water-soluble proteins through acetone precipitation. *Anal Chim Acta* **796**:48-54.

**d'Orval BC, Bortolin M-L, Gaspin C, Bachellerie J-P.** 2001. Box C/D RNA guides for the ribose methylation of archaeal tRNAs. The tRNATrp intron guides the formation of two ribose-methylated nucleosides in the mature tRNATrp. *Nucleic Acids Res* **29**:4518-4529.

**Dann SM, Le CHY, Hanson EM, Ross MC, Eckmann L.** 2018. *Giardia* Infection of the Small Intestine Induces Chronic Colitis in Genetically Susceptible Hosts. *J immunol* **201**:548-559.

- Darzacq X, Jády BE, Verheggen C, Kiss AM, Bertrand E, 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.
- del Campo J, Sieracki ME, Molestina R, Keeling P, Massana R, Ruiz-Trillo I.** 2014. The others: our biased perspective of eukaryotic genomes. *Trends Ecol Evol* **29**:252-259.
- Dennis PP, Tripp V, Lui L, Lowe T, Randau L.** 2015. C/D box sRNA-guided 2'- O -methylation patterns of archaeal rRNA molecules. *BMC Genomics* **16**:632.
- Deryusheva S, Gall JG.** 2013. Novel small Cajal-body-specific RNAs identified in *Drosophila*: probing guide RNA function. *RNA* **19**:1802-1814.
- Dimova DK, Dyson NJ.** 2005. The E2F transcriptional network: old acquaintances with new faces. *Oncogene* **24**:2810-2826.
- Dobáková E, Flegontov P, Skalický T, Lukeš J.** 2015. Unexpectedly Streamlined Mitochondrial Genome of the Euglenozoan *Euglena gracilis*. *Genome Biol Evol* **7**:3358-3367.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR.** 2012. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:15-21.
- Dragon F, Gallagher JE, Compagnone-Post PA, Mitchell BM, Porwancher KA, Wehner KA, Wormsley S, Settlege RE, Shabanowitz J, Osheim Y, et al.** 2002. A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. *Nature* **417**:967-970.
- Dsouza M, Larsen N, Overbeek R.** 1997. Searching for patterns in genomic data. *Trends Genet* **13**:497-498.
- Duan J, Li L, Lu J, Wang W, Ye K.** 2009. Structural Mechanism of Substrate RNA Recruitment in H/ACA RNA-Guided Pseudouridine Synthase. *Mol Cell* **34**:427-439.
- Dunbar DA, Baserga SJ.** 1998. The U14 snoRNA is required for 2'-O-methylation of the pre-18S rRNA in *Xenopus* oocytes. *RNA* **4**:195-204.
- Dupuis-Sandoval F, Poirier M, Scott MS.** 2015. The emerging landscape of small nucleolar RNAs in cell biology. *WIREs RNA* **6**:381-397.
- Dvinge H, Guenthoer J, Porter PL, Bradley RK.** 2019. RNA components of the spliceosome regulate tissue- and cancer-specific alternative splicing. *Genome Res* **29**:1591-1604.
- Ebenezer TE, Carrington M, Lebert M, Kelly S, Field MC.** 2017. *Euglena gracilis* Genome and Transcriptome: Organelles, Nuclear Genome Assembly Strategies and Initial Features. In: Schwartzback SD, Shigeoka S, editors. *Euglena: Biochemistry, Cell and*

Molecular Biology. Advances in Experimental Medicine and Biology: Springer. p. 159-182.

**Ebenezer TE, Zoltner M, Burrell A, Nenarokova A, Novák Vanclová AMG, Prasad B, Soukal P, Santana-Molina C, O'Neill E, Nankissoor NN, et al.** 2019. Transcriptome, proteome and draft genome of *Euglena gracilis*. *BMC Biology* **17**:1-23.

**Edgar RC.** 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**:2460-2461.

**Eiler DR, Wimberly BT, Bilodeau DY, Rissland OS, Kieft JS.** 2020. The *Giardia lamblia* ribosome structure reveals divergence in translation and quality control pathways. *bioRxiv*:2020.2009.2030.321331.

**Einarsson E, Ma'ayeh S, Svärd SG.** 2016. An up-date on *Giardia* and giardiasis. *Curr Opin Microbiol* **34**:47-52.

**Elias EV, Quiroga R, Gottig N, Nakanishi H, Nash TE, Neiman A, Lujan HD.** 2008. Characterization of SNAREs determines the absence of a typical Golgi apparatus in the ancient eukaryote *Giardia lamblia*. *J Biol Chem* **283**:35996-36010.

**Eliasz D, Doniger T, Tkacz ID, Gupta SK, Kolev NG, Unger R, Ullu E, Tschudi C, Michaeli S.** 2015. Genome-wide analysis of small nucleolar RNAs of *Leishmania major* reveals a rich repertoire of RNAs involved in modification and processing of rRNA. *RNA Biol*:0.

**Elliott BA, Ho H-T, Ranganathan SV, Vangaveti S, Ilkayeva O, Abou Assi H, Choi AK, Agris PF, Holley CL.** 2019. Modification of messenger RNA by 2'-O-methylation regulates gene expression in vivo. *Nat Commun* **10**:3401.

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

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

**Esakova O, Perederina A, Quan C, Berezin I, Krasilnikov AS.** 2011. Substrate recognition by ribonucleoprotein ribonuclease MRP. *RNA* **17**:356-364.

**Esguerra J, Warringer J, Blomberg A.** 2008. Functional importance of individual rRNA 2'-O-ribose methylations revealed by high-resolution phenotyping. *RNA* **14**:649-656.

**Falaleeva M, Pages A, Matuszek Z, Hidmi S, Agranat-Tamir L, Korotkov K, Nevo Y, Eyras E, Sperling R, Stamm S.** 2016. Dual function of C/D box small nucleolar RNAs in rRNA modification and alternative pre-mRNA splicing. *Proc Natl Acad Sci U S A* **113**:E1625-E1634.



- Falaleeva M, Welden JR, Duncan MJ, Stamm S.** 2017. C/D-box snoRNAs form methylating and non-methylating ribonucleoprotein complexes: Old dogs show new tricks. *BioEssays* **39**:1600264.
- Faust T, Frankel A, D'Orso I.** 2012. Transcription control by long non-coding RNAs. *Transcription* **3**:78-86.
- Fayet-Lebaron E, Atzorn V, Henry Y, 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-M, Tian H-F, Wen J-F.** 2013. Origin and Evolution of the Eukaryotic SSU Processome Revealed by a Comprehensive Genomic Analysis and Implications for the Origin of the Nucleolus. *Genome Biol Evol* **5**:2255-2267.
- Fica SM, Tuttle N, Novak T, Li N-S, Lu J, Koodathingal P, Dai Q, Staley JP, Piccirilli JA.** 2013. RNA catalyses nuclear pre-mRNA splicing. *Nature* **503**:229-234.
- Franzén O, Jerlström-Hultqvist J, Castro E, Sherwood E, Ankarklev J, Reiner DS, Palm D, Andersson JO, Andersson B, Svärd SG.** 2009. Draft genome sequencing of *Giardia intestinalis* assemblage B isolate GS: is human giardiasis caused by two different species? *PLoS Pathog* **5**:e1000560.
- Franzén O, Jerlström-Hultqvist J, Einarsson E, Ankarklev J, Ferella M, Andersson B, Svärd SG.** 2013. Transcriptome Profiling of *Giardia intestinalis* Using Strand-specific RNA-Seq. *PLoS Comput Biol* **9**:e1003000.
- Fu D, Collins K.** 2007. Purification of human telomerase complexes identifies factors involved in telomerase biogenesis and telomere length regulation. *Mol cell* **28**:773-785.
- Gagnon KT, Biswas S, Zhang X, Brown BA, 2nd, Wollenzien P, Mattos C, Maxwell ES.** 2012. Structurally conserved Nop56/58 N-terminal domain facilitates archaeal box C/D ribonucleoprotein-guided methyltransferase activity. *J Biol Chem* **287**:19418-19428.
- Gagnon KT, Zhang X, Qu G, Biswas S, Suryadi J, Brown BA, 2nd, Maxwell ES.** 2010. Signature amino acids enable the archaeal L7Ae box C/D RNP core protein to recognize and bind the K-loop RNA motif. *RNA* **16**:79-90.
- Galardi S, Fatica A, Bachi A, Scaloni A, Presutti C, Bozzoni I.** 2002. Purified Box C/D snoRNPs Are Able To Reproduce Site-Specific 2'-O-Methylation of Target RNA In Vitro. *Mol Cell Biol* **22**:6663-6668.
- Ganot P, Bortolin M-L, Kiss T.** 1997. Site-Specific Pseudouridine Formation in Preribosomal RNA Is Guided by Small Nucleolar RNAs. *Cell* **89**:799-809.
- Ganot P, Caizergues-Ferrer M, Kiss T.** 1997. 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, Cavaillé J, Erauso G, Bachellerie J-P.** 2000. Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the *Pyrococcus* genomes. *J Mol Biol* **297**:895-906.
- Gebert LFR, MacRae IJ.** 2019. Regulation of microRNA function in animals. *Nat Rev Mol Cell Bio* **20**:21-37.
- Genuth NR, Barna M.** 2018. The Discovery of Ribosome Heterogeneity and Its Implications for Gene Regulation and Organismal Life. *Mol Cell* **71**:364-374.
- Ghalei H, Hsiao HH, Urlaub H, Wahl MC, Watkins NJ.** 2010. A novel Nop5-sRNA interaction that is required for efficient archaeal box C/D sRNP formation. *RNA* **16**:2341-2348.
- Gibbs SP.** 1978. The chloroplasts of *Euglena* may have evolved from symbiotic green algae. *Canadian Journal of Botany* **56**:2883-2889.
- Gilbert W.** 1986. Origin of life: The RNA world. *Nature* **319**:618-618.
- Gill T, Cai T, Aulds J, Wierzbicki S, Schmitt ME.** 2004. RNase MRP Cleaves the *CLB2* mRNA To Promote Cell Cycle Progression: Novel Method of mRNA Degradation. *Mol Cell Biol* **24**:945-953.
- Golovanov AP, Hautbergue GM, Wilson SA, Lian L-Y.** 2004. A Simple Method for Improving Protein Solubility and Long-Term Stability. *J Am Chem Soc* **126**:8933-8939.
- Goodrich JA, Kugel JF.** 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.
- Gorynia S, Bandejas TM, Pinho FG, McVey CE, Vonnrhein C, Round A, Svergun DI, Donner P, Matias PM, Carrondo MA.** 2011. Structural and functional insights into a dodecameric molecular machine – The RuvBL1/RuvBL2 complex. *J Struct Biol* **176**:279-291.
- Granneman S, Kudla G, Petfalski E, Tollervey D.** 2009. Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc Natl Acad Sci U S A* **106**:9613-9618.
- Greenwood SJ, Schnare MN, Gray MW.** 1996. Molecular characterization of U3 small nucleolar RNA from the early diverging protist, *Euglena gracilis*. *Curr Genet* **30**:338-346.
- Greider CW, Blackburn EH.** 1989. A telomeric sequence in the RNA of *Tetrahymena* telomerase required for telomere repeat synthesis. *Nature* **337**:331-337.
- Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S.** 1983. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**:849-857.



- Gumińska N, Plecha M, Zakryś B, Milanowski R.** 2018. Order of removal of conventional and nonconventional introns from nuclear transcripts of *Euglena gracilis*. *PLoS Genet* **14**:e1007761.
- Gupta S, Roy M, Dey D, Bhakta K, Bhowmick A, Chattopadhyay D, Ghosh A.** 2021. Archaeal SRP RNA and SRP19 facilitate the assembly of SRP54-FtsY targeting complex. *Biochem Biophys Res Commun* **566**:53-58.
- Gupta SK, Hury A, Ziporen Y, Shi H, Ullu E, 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.
- Gupta SK, Kolet L, Doniger T, Biswas VK, Unger R, Tzfati Y, Michaeli S.** 2013. The *Trypanosoma brucei* telomerase RNA (TER) homologue binds core proteins of the C/D snoRNA family. *FEBS Lett* **587**:1399-1404.
- Hafner M, Katsantoni M, Köster T, Marks J, Mukherjee J, Staiger D, Ule J, Zavolan M.** 2021. CLIP and complementary methods. *Nat Rev Methods Primers* **1**:20.
- Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, Spielmann A, Stutz E.** 1993. Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Res* **21**:3537-3544.
- Hartshorne T, Toyofuku W.** 1999. Two 5'-ETS regions implicated in interactions with U3 snoRNA are required for small subunit rRNA maturation in *Trypanosoma brucei*. *Nucleic Acids Res* **27**:3300-3309.
- Hasan G, Turner MJ, Cordingley JS.** 1984. Ribosomal RNA genes of *Trypanosoma brucei*: Mapping the regions specifying the six small ribosomal RNAs. *Gene* **27**:75-86.
- Hashem Y, Georges Ad, Fu J, Buss SN, Jossinet F, Jobe A, Zhang Q, Liao HY, Grassucci RA, Bajaj C, et al.** 2013. High-resolution cryo-electron microscopy structure of the *Trypanosoma brucei* ribosome. In: Single-Particle Cryo-Electron Microscopy. p. 456-462.
- Hashimoto C, Steitz JA.** 1984. U4 and U6 RNAs coexist in a single small nuclear ribonucleoprotein particle. *Nucleic Acids Res* **12**:3283-3293.
- He S, Wurtzel O, Singh K, Froula JL, Yilmaz S, Tringe SG, Wang Z, Chen F, Lindquist EA, Sorek R, et al.** 2010. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat Methods* **7**:807-812.
- Henras A, Dez C, Noaillac-Depeyre J, Henry Y, Caizergues-Ferrer M.** 2001. Accumulation of H/ACA snoRNPs depends on the integrity of the conserved central domain of the RNA-binding protein Nhp2p. *Nucleic Acids Res* **29**:2733-2746.

**Henras A, Henry Y, Bousquet-Antonelli C, Noaillac-Depeyre J, Gélugne JP, Caizergues-Ferrer M.** 1998. Nhp2p and Nop10p are essential for the function of H/ACA snoRNPs. *EMBO J* **17**:7078-7090.

**Henras AK, Plisson-Chastang C, Humbert O, Romeo Y, Henry Y.** 2017. Synthesis, Function, and Heterogeneity of snoRNA-Guided Posttranscriptional Nucleoside Modifications in Eukaryotic Ribosomal RNAs. In: Chanfreau GF, editor. *The Enzymes*: Academic Press. p. 169-213.

**Hille F, Richter H, Wong SP, Bratovič M, Ressel S, Charpentier E.** 2018. The Biology of CRISPR-Cas: Backward and Forward. *Cell* **172**:1239-1259.

**Hirakata S, Siomi MC.** 2016. piRNA biogenesis in the germline: From transcription of piRNA genomic sources to piRNA maturation. *Biochim Biophys Acta Gene Regul Mech* **1859**:82-92.

**Hoagland MB, Stephenson ML, Scott JF, Hecht LI, Zamecnik PC.** 1958. A soluble ribonucleic acid intermediate in protein synthesis. *J Biol Chem* **231**:241-257.

**Hoepfner MP, Poole AM.** 2012. Comparative genomics of eukaryotic small nucleolar RNAs reveals deep evolutionary ancestry amidst ongoing intragenomic mobility. *BMC Evol Biol* **12**:183.

**Hopper AK, Pai DA, Engelke DR.** 2010. Cellular dynamics of tRNAs and their genes. *FEBS Lett* **584**:310-317.

**Hori H.** 2014. Methylated nucleosides in tRNA and tRNA methyltransferases. *Front Genet* **5**.

**Houry WA, Bertrand E, Coulombe B.** 2018. The PAQosome, an R2TP-Based Chaperone for Quaternary Structure Formation. *Trends Biochem Sci* **43**:4-9.

**Huang C, Shi J, Guo Y, Huang W, Huang S, Ming S, Wu X, Zhang R, Ding J, Zhao W, et al.** 2017. A snoRNA modulates mRNA 3' end processing and regulates the expression of a subset of mRNAs. *Nucleic Acids Res* **45**:8647-8660.

**Huang L, Lilley DMJ.** 2018. The kink-turn in the structural biology of RNA. *Q Rev Biophys* **51**:e5.

**Huang L, Lilley DMJ.** 2016. The Kink Turn, a Key Architectural Element in RNA Structure. *J Mol Biol* **428**:790-801.

**Hudson AJ.** 2014. Spliceosomal intron and spliceosome evolution in *Giardia lamblia* and other diplomonads. [Ph.D.]. [Lethbridge]: University of Lethbridge.

**Hudson AJ, McWatters DC, Bowser BA, Moore AN, Larue GE, Roy SW, Russell AG.** 2019. Patterns of conservation of spliceosomal intron structures and spliceosome

divergence in representatives of the diplomonad and parabasalid lineages. *BMC Evol Biol* **19**:162.

**Hudson AJ, Moore AN, Elniski D, Joseph J, Yee J, Russell AG.** 2012. Evolutionarily divergent spliceosomal snRNAs and a conserved non-coding RNA processing motif in *Giardia lamblia*. *Nucleic Acids Res* **40**:10995-11008.

**Hudson AJ, Stark MR, Fast NM, Russell AG, Rader SD.** 2015. Splicing diversity revealed by reduced spliceosomes in *C. merolae* and other organisms. *RNA Biol* **12**:1-8.

**Hughes JM, Ares M, Jr.** 1991. Depletion of U3 small nucleolar RNA inhibits cleavage in the 5' external transcribed spacer of yeast pre-ribosomal RNA and impairs formation of 18S ribosomal RNA. *EMBO J* **10**:4231-4239.

**Hughes JM, Konings DA, Cesareni G.** 1987. The yeast homologue of U3 snRNA. *EMBO J* **6**:2145-2155.

**Hughes JMX.** 1996. Functional Base-pairing Interaction Between Highly Conserved Elements of U3 Small Nucleolar RNA and the Small Ribosomal Subunit RNA. *Jol Mol Biol* **259**:645-654.

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

**Jackson AP, Otto TD, Aslett M, Armstrong SD, Bringaud F, Schlacht A, Hartley C, Sanders M, Wastling JM, Dacks JB, et al.** 2016. Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism. *Curr Biol* **26**:161-172.

**Jády BE, Bertrand E, Kiss T.** 2004. Human telomerase RNA and box H/ACA scaRNAs share a common Cajal body-specific localization signal. *J Cell Biol* **164**:647-652.

**Jády BE, Darzacq X, Tucker KE, Matera AG, Bertrand E, Kiss T.** 2003. Modification of Sm small nuclear RNAs occurs in the nucleoplasmic Cajal body following import from the cytoplasm. *EMBO J* **22**:1878-1888.

**Jády BE, 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.

**Jarrous N.** 2017. Roles of RNase P and Its Subunits. *Trends Genet* **33**:594-603.

**Jerlström-Hultqvist J, Einarsson E, Xu F, Hjort K, Ek B, Steinhauf D, Hultenby K, Bergquist J, Andersson JO, Svärd SG.** 2013. Hydrogenosomes in the diplomonad *Spironucleus salmonicida*. *Nat Commun* **4**:2493.

**Jerlström-Hultqvist J, Franzén O, Ankarklev J, Xu F, Nohýnková E, Andersson JO, Svärd SG, Andersson B.** 2010. Genome analysis and comparative genomics of a *Giardia intestinalis* assemblage E isolate. *BMC Genomics* **11**:543.

**Jerlström-Hultqvist J, Stadelmann B, Birkestedt S, Hellman U, Svärd SG.** 2012. Plasmid Vectors for Proteomic Analyses in *Giardia*: Purification of Virulence Factors and Analysis of the Proteasome. *Eukaryot Cell* **11**:864-873.

**Jex AR, Svard S, Hagen KD, Starceovich H, Emery-Corbin SJ, Balan B, Nosala C, Dawson SC.** 2020. Recent advances in functional research in *Giardia intestinalis*. *Adv Parasitol* **107**:97-137.

**Jiménez-García LF, Zavala G, Chávez-Munguía B, Ramos-Godínez Mdel P, López-Velázquez G, Segura-Valdez Mde L, Montañez C, Hehl AB, Argüello-García R, Ortega-Pierres G.** 2008. Identification of nucleoli in the early branching protist *Giardia duodenalis*. *Int J Parasitol* **38**:1297-1304.

**Joardar A, Malliahgari SR, Skariah G, Gupta R.** 2011. 2'-O-methylation of the wobble residue of elongator pre-tRNAMet in *Haloferax volcanii* is guided by a box C/D RNA containing unique features. *RNA Biol* **8**:782-791.

**Jorjani H, Kehr S, Jedlinski DJ, Gumienny R, Hertel J, Stadler PF, Zavolan M, Gruber AR.** 2016. An updated human snoRNAome. *Nucleic Acids Res* **44**:5068-5082.

**Karnkowska A, Treitli SC, Brzoň O, Novák L, Vacek V, Soukal P, Barlow LD, Herman EK, Pipaliya SV, Pánek T, et al.** 2019. The Oxymonad Genome Displays Canonical Eukaryotic Complexity in the Absence of a Mitochondrion. *Mol Biol Evol* **36**:2292-2312.

**Karnkowska A, Vacek V, Zubáčová Z, Treitli SC, Petrželková R, Eme L, Novák L, Žárský V, Barlow LD, Herman EK, et al.** 2016. A Eukaryote without a Mitochondrial Organelle. *Curr Biol* **26**:1274-1284.

**Kass S, Tyc K, Steitz JA, Sollner-Webb B.** 1990. The U3 small nucleolar ribonucleoprotein functions in the first step of preribosomal RNA processing. *Cell* **60**:897-908.

**Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE.** 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**:845-858.

**Khoshnevis S, Dreggors RE, Hoffman TFR, Ghalei H.** 2019. A conserved Bcd1 interaction essential for box C/D snoRNP biogenesis. *J Biol Chem* **294**:18360-19371.

**Kikovska E, Svärd SG, Kirsebom LA.** 2007. Eukaryotic RNase P RNA mediates cleavage in the absence of protein. *Proc Natl Acad Sci U S A* **104**:2062-2067.

**Kishore S, Khanna A, Zhang Z, Hui J, Balwierz PJ, Stefan M, Beach C, Nicholls RD, Zavolan M, 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, Stamm S.** 2006. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science* **311**:230-232.

- Kiss-László Z, Henry Y, Bachellerie J-P, Caizergues-Ferrer M, Kiss T.** 1996. Site-Specific Ribose Methylation of Preribosomal RNA: A Novel Function for Small Nucleolar RNAs. *Cell* **85**:1077-1088.
- Kiss AM, Jádý BE, Darzacq X, Verheggen C, Bertrand E, Kiss T.** 2002. A Cajal body-specific pseudouridylation guide RNA is composed of two box H/ACA snoRNA-like domains. *Nucleic Acids Res* **30**:4643-4649.
- Klein DJ, Schmeing TM, Moore PB, Steitz TA.** 2001. The kink-turn: a new RNA secondary structure motif. *EMBO J* **20**:4214-4221.
- Koonin EV, Bork P, Sander C.** 1994. A novel RNA-binding motif in omnipotent suppressors of translation termination, ribosomal proteins and a ribosome modification enzyme? *Nucleic Acids Res* **22**:2166-2167.
- Kos M, Tollervey D.** 2010. Yeast pre-rRNA processing and modification occur cotranscriptionally. *Mol Cell* **37**:809-820.
- Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR.** 1982. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell* **31**:147-157.
- Kudla G, Granneman S, Hahn D, Beggs JD, Tollervey D.** 2011. Cross-linking, ligation, and sequencing of hybrids reveals RNA–RNA interactions in yeast. *Proc Natl Acad Sci U S A* **108**:10010-10015.
- Kuhn JF, Tran EJ, Maxwell ES.** 2002. Archaeal ribosomal protein L7 is a functional homolog of the eukaryotic 15.5kD/Snu13p snoRNP core protein. *Nucleic Acids Res* **30**:931-941.
- Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL.** 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5**:R12.
- Lan P, Zhou B, Tan M, Li S, Cao M, Wu J, Lei M.** 2020. Structural insight into precursor ribosomal RNA processing by ribonuclease MRP. *Science* **369**:656-663.
- Langhendries J-L, Nicolas E, Doumont G, Goldman S, Lafontaine DLJ.** 2016. The human box C/D snoRNAs U3 and U8 are required for pre-rRNA processing and tumorigenesis. *Oncotarget* **7**:59519-59534.
- Lara-Martínez R, De Lourdes Segura-Valdez M, De La Mora-De La Mora I, López-Velázquez G, Jiménez-García LF.** 2016. Morphological Studies of Nucleologenesis in *Giardia lamblia*. *Anat Rec* **299**:549-556.
- Laslett D, Bjorn C.** 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* **32**:11-16.

- Lemay V, Hossain A, Osheim YN, Beyer AL, Dragon F.** 2011. Identification of novel proteins associated with yeast snR30 small nucleolar RNA. *Nucleic Acids Res* **39**:9659-9670.
- Li HD, Zagorski J, Fournier MJ.** 1990. Depletion of U14 small nuclear RNA (snR128) disrupts production of 18S rRNA in *Saccharomyces cerevisiae*. *Mol Cell Biol* **10**:1145-1152.
- Li L, Ye K.** 2006. Crystal structure of an H/ACA box ribonucleoprotein particle. *Nature* **443**:302-307.
- Li S, Duan J, Li D, Yang B, Dong M, Ye K.** 2011. Reconstitution and structural analysis of the yeast box H/ACA RNA-guided pseudouridine synthase. *Genes Dev* **25**:2409-2421.
- Li W, Saraiya AA, Wang CC.** 2011. Gene Regulation in *Giardia lamblia* Involves a Putative MicroRNA Derived from a Small Nucleolar RNA. *PLoS Negl Trop Dis* **5**:e1338.
- Liang B, Zhou J, Kahen E, Terns RM, Terns MP, Li H.** 2009. Structure of a functional ribonucleoprotein pseudouridine synthase bound to a substrate RNA. *Nat Struct Mol Biol* **16**:740-746.
- Liang WQ, Fournier MJ.** 1995. U14 base-pairs with 18S rRNA: a novel snoRNA interaction required for rRNA processing. *Genes Dev* **9**:2433-2443.
- Liang X-h, Hury A, Hoze E, Uliel S, Myslyuk I, Apatoff A, Unger R, Michaeli S.** 2007. Genome-Wide Analysis of C/D and H/ACA-Like Small Nucleolar RNAs in *Leishmania major* Indicates Conservation among Trypanosomatids in the Repertoire and in Their rRNA Targets. *Eukaryot Cell* **6**:361-377.
- Liang X-h, Liu Q, Fournier MJ.** 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, Xu Y-X, Michaeli S.** 2002. 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.
- Liang XH, Uliel S, Hury A, Barth S, Doniger T, Unger R, Michaeli S.** 2005. A genome-wide analysis of C/D and H/ACA-like small nucleolar RNAs in *Trypanosoma brucei* reveals a trypanosome-specific pattern of rRNA modification. *RNA* **11**:619-645.
- Liao Y, Smyth GK, Shi W.** 2019. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res* **47**:1-9.
- Lilley DMJ.** 2012. The structure and folding of kink turns in RNA. *WIREs RNA* **3**:797-805.



- Lindahl L, Bommankanti A, Li X, Hayden L, Jones A, Khan M, Oni T, Zengel JM.** 2009. RNase MRP is required for entry of 35S precursor rRNA into the canonical processing pathway. *RNA* **15**:1407-1416.
- Lingner J, Hendrick LL, Cech TR.** 1994. Telomerase RNAs of different ciliates have a common secondary structure and a permuted template. *Genes Dev* **8**:1984-1998.
- Liu S, Li P, Dybkov O, Nottrott S, Hartmuth K, Lührmann R, Carlomagno T, Wahl MC.** 2007. Binding of the Human Prp31 Nop Domain to a Composite RNA-Protein Platform in U4 snRNP. *Science* **316**:115-120.
- Logeswaran D, Li Y, Podlevsky JD, Chen JJ-L.** 2020. Monophyletic Origin and Divergent Evolution of Animal Telomerase RNA. *Mol Biol Evol* **38**:215-228.
- Lovejoy AF, Riordan DP, Brown PO.** 2014. Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in *S. cerevisiae*. *PLoS One* **9**:e110799.
- Lowe TM, Eddy SR.** 1999. A computational screen for methylation guide snoRNAs in yeast. *Science* **283**:1168-1171.
- Lu Z, Zhang Qiangfeng C, Lee B, Flynn Ryan A, Smith Martin A, Robinson James T, Davidovich C, Gooding Anne R, Goodrich Karen J, Mattick John S, et al.** 2016. RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell* **165**:1267-1279.
- Lui LM, Uzilov AV, Bernick DL, Corredor A, Lowe TM, Dennis PP.** 2018. Methylation guide RNA evolution in archaea: structure, function and genomic organization of 110 C/D box sRNA families across six *Pyrobaculum* species. *Nucleic Acids Res* **46**:5678-5691.
- Lukowiak AA, Narayanan A, Li ZH, Terns RM, Terns MP.** 2001. The snoRNA domain of vertebrate telomerase RNA functions to localize the RNA within the nucleus. *RNA* **7**:1833-1844.
- Luo J, Zhou H, Chen C, Li Y, Chen Y, Qu L.** 2006. Identification and evolutionary implication of four novel box H/ACA snoRNAs from *Giardia lamblia*. *Chinese Sci Bull* **51**:2451-2456.
- Lygerou Z, Allmang C, Tollervey D, Séraphin B.** 1996. Accurate processing of a eukaryotic precursor ribosomal RNA by ribonuclease MRP in vitro. *Science* **272**:268-270.
- MacRae IJ, Zhou K, Li F, Repic A, Brooks AN, Cande WZ, Adams PD, Doudna JA.** 2006. Structural Basis for Double-Stranded RNA Processing by Dicer. *Science* **311**:195-198.
- Madhani HD, Guthrie C.** 1992. A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome. *Cell* **71**:803-817.

- Makarova OV, Makarov EM, Liu S, Vornlocher H-P, Lührmann R.** 2002. Protein 61K, encoded by a gene (PRPF31) linked to autosomal dominant retinitis pigmentosa, is required for U4/U6\*U5 tri-snRNP formation and pre-mRNA splicing. *EMBO J* **21**:1148-1157.
- Malik HS, Burke WD, Eickbush TH.** 2000. Putative telomerase catalytic subunits from *Giardia lamblia* and *Caenorhabditis elegans*. *Gene* **251**:101-108.
- Malinová A, Cvačková Z, Matějů D, Hořejší Z, Abéza C, Vandermoere F, Bertrand E, Staněk D, Verheggen C.** 2017. Assembly of the U5 snRNP component PRPF8 is controlled by the HSP90/R2TP chaperones. *J Cell Biol* **216**:1579-1596.
- Manning G, Reiner DS, Lauwaet T, Dacre M, Smith A, Zhai Y, Svard S, Gillin FD.** 2011. The minimal kinome of *Giardia lamblia* illuminates early kinase evolution and unique parasite biology. *Genome Biol* **12**:R66.
- Marmier-Gourrier N, Cléry A, Schlotter F, Senty-Ségault V, Branlant C.** 2011. A second base pair interaction between U3 small nucleolar RNA and the 5'-ETS region is required for early cleavage of the yeast pre-ribosomal RNA. *Nucleic Acids Res* **39**:9731-9745.
- Marnef A, Richard P, Pinzón N, Kiss T.** 2014. Targeting vertebrate intron-encoded box C/D 2'-O-methylation guide RNAs into the Cajal body. *Nucleic Acids Res* **42**:6616-6629.
- Martin M.** 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**:3.
- Marz M, Stadler PF.** 2009. Comparative analysis of eukaryotic U3 snoRNA. *RNA Biol* **6**:503-507.
- Massenet S, Bertrand E, Verheggen C.** 2017. Assembly and trafficking of box C/D and H/ACA snoRNPs. *RNA Biol* **14**:680-692.
- Matera AG, Terns RM, Terns MP.** 2007. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol* **8**:209-220.
- Mattick JS, Makunin IV.** 2006. Non-coding RNA. *Hum Mol Genet* **15 Spec No 1**:R17-29.
- McCormick-Graham M, Romero DP.** 1995. Ciliate telomerase RNA structural features. *Nucleic Acids Res* **23**:1091-1097.
- McInally SG, Hagen KD, Nosala C, Williams J, Nguyen K, Booker J, Jones K, Dawson SC.** 2019. Robust and stable transcriptional repression in *Giardia* using CRISPRi. *Mol Biol Cell* **30**:119-130.
- McPhee SA, Huang L, Lilley DMJ.** 2014. A critical base pair in k-turns that confers folding characteristics and correlates with biological function. *Nat Commun* **5**:5127.



- McWatters DC, Russell AG.** 2017. Euglena Transcript Processing. In: Schwartzbach SD, Shigeoka S, editors. *Euglena: Biochemistry, Cell and Molecular Biology*. Advances in Experimental Medicine and Biology Springer. p. 141-158.
- Mefford MA, Staley JP.** 2009. Evidence that U2/U6 helix I promotes both catalytic steps of pre-mRNA splicing and rearranges in between these steps. *RNA* **15**:1386-1397.
- Meier UT.** 2017. RNA modification in Cajal bodies. *RNA Biol* **14**:693-700.
- Mercer TR, Dinger ME, Mattick JS.** 2009. Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10**:155-159.
- Méreau A, Fournier R, Grégoire A, Mougín A, Fabrizio P, Lührmann R, Branlant C.** 1997. An in vivo and in vitro structure-function analysis of the *Saccharomyces cerevisiae* U3A snoRNP: protein-RNA contacts and base-pair interaction with the pre-ribosomal RNA. *J Mol Biol* **273**:552-571.
- Michaeli S, Doniger T, Gupta SK, Wurtzel O, Romano M, Visnovezky D, Sorek R, Unger R, Ullu E.** 2012. RNA-seq analysis of small RNPs in *Trypanosoma brucei* reveals a rich repertoire of non-coding RNAs. *Nucleic Acids Res* **40**:1282-1298.
- Michel CI, Holley CL, Scruggs BS, Sidhu R, Brookheart RT, Listenberger LL, Behlke MA, Ory DS, Schaffer JE.** 2011. Small nucleolar RNAs U32a, U33, and U35a are critical mediators of metabolic stress. *Cell Metab* **14**:33-44.
- Mihalusova M, Wu JY, Zhuang X.** 2011. Functional importance of telomerase pseudoknot revealed by single-molecule analysis. *Proc Natl Acad Sci U S A* **108**:20339-20344.
- Milanowski R, Karnkowska A, Ishikawa T, Zakryś B.** 2014. Distribution of conventional and nonconventional introns in tubulin ( $\alpha$  and  $\beta$ ) genes of euglenids. *Mol Biol Evol* **31**:584-593.
- Mitchell JR, Cheng J, 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.
- Moore AN.** 2015. Characterization of small nucleolar RNAs in the protist organism *Euglena gracilis*. [Ph.D.]. [Lethbridge]: University of Lethbridge.
- Moore AN, Russell AG.** 2012. Clustered organization, polycistronic transcription, and evolution of modification-guide snoRNA genes in *Euglena gracilis*. *Mol Genet Genomics* **287**:55-66.
- Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, Olsen GJ, Best AA, Cande WZ, Chen F, Cipriano MJ, et al.** 2007. Genomic Minimalism in the Early Diverging Intestinal Parasite *Giardia lamblia*. *Science* **317**:1921-1926.

- Morrissey JP, Tollervey D.** 1993. Yeast snR30 is a small nucleolar RNA required for 18S rRNA synthesis. *Mol Cell Biol* **13**:2469-2477.
- Mottram J, Perry KL, Lizardi PM, Luhrmann R, Agabian N, Nelson RG.** 1989. Isolation and sequence of four small nuclear U RNA genes of *Trypanosoma brucei* subsp. *brucei*: identification of the U2, U4, and U6 RNA analogs. *Mol Cell Biol* **9**:1212-1223.
- Mount SM, Pettersson I, Hinterberger M, Karmas A, Steitz JA.** 1983. The U1 small nuclear RNA-protein complex selectively binds a 5' splice site in vitro. *Cell* **33**:509-518.
- Muchhal US, Schwartzbach SD.** 1994. Characterization of the unique intron-exon junctions of *Euglena* gene(s) encoding the polypeptide precursor to the light-harvesting chlorophyll a/b binding protein of photosystem II. *Nucleic Acids Res* **22**:5737-5744.
- Muñoz-Hernández H, Pal M, Rodríguez CF, Fernandez-Leiro R, Prodromou C, Pearl LH, Llorca O.** 2019. Structural mechanism for regulation of the AAA-ATPases RUVBL1-RUVBL2 in the R2TP co-chaperone revealed by cryo-EM. *Sci Adv* **5**:eaaw1616.
- Musgrove C, Jansson LI, Stone MD.** 2018. New perspectives on telomerase RNA structure and function. *WIREs RNA* **9**:1-15.
- Myslyuk I, Doniger T, Horesh Y, Hury A, Hoffer R, Ziporen Y, Michaeli S, 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.
- Narcisi EM, Glover CV, Fechheimer M.** 1998. Fibrillarin, a conserved pre-ribosomal RNA processing protein of *Giardia*. *J Eukaryot Microbiol* **45**:105-111.
- Nawrocki EP, Eddy SR.** 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**:2933-2935.
- Nguyen THD, Galej WP, Bai X-c, Oubridge C, Newman AJ, Scheres SHW, Nagai K.** 2016. Cryo-EM structure of the yeast U4/U6.U5 tri-snRNP at 3.7 Å resolution. *Nature* **530**:298-302.
- Nguyen THD, Tam J, Wu RA, Greber BJ, Toso D, Nogales E, Collins K.** 2018. Cryo-EM structure of substrate-bound human telomerase holoenzyme. *Nature* **557**:190-195.
- Ni J, Tien AL, Fournier MJ.** 1997. Small Nucleolar RNAs Direct Site-Specific Synthesis of Pseudouridine in Ribosomal RNA. *Cell* **89**:565-573.
- Nolivos S, Carpousis AJ, Clouet-d'Orval B.** 2005. The K-loop, a general feature of the *Pyrococcus* C/D guide RNAs, is an RNA structural motif related to the K-turn. *Nucleic Acids Res* **33**:6507-6514.
- Nottrott S, Hartmuth K, Fabrizio P, Urlaub H, Vidovic I, Ficner R, Lührmann R.** 1999. Functional interaction of a novel 15.5kD[U4/U6·U5] tri-snRNP protein with the 5' stem-loop of U4 snRNA. *EMBO J* **18**:6119-6133.

- O'Keefe RT, Newman AJ.** 1998. Functional analysis of the U5 snRNA loop 1 in the second catalytic step of yeast pre-mRNA splicing. *EMBO J* **17**:565-574.
- O'Neil D, Glowatz H, Schlumpberger M.** 2013. Ribosomal RNA depletion for efficient use of RNA-seq capacity. *Curr Protoc Mol Biol* **Chapter 4**:Unit 4 19.
- O'Neill EC, Trick M, Hill L, Rejzek M, Dusi RG, Hamilton CJ, Zimba PV, Henrissat B, Field RA.** 2015. The transcriptome of *Euglena gracilis* reveals unexpected metabolic capabilities for carbohydrate and natural product biochemistry. *Mol Biosyst* **11**:2808-2820.
- Ogbonna JC, Ichige E, Tanaka H.** 2002. Interactions between photoautotrophic and heterotrophic metabolism in photoheterotrophic cultures of *Euglena gracilis*. *Appl Microbiol Biotechnol* **58**:532-538.
- Okamura K, Lai EC.** 2008. Endogenous small interfering RNAs in animals. *Nat Rev Mol Cell Biol* **9**:673-678.
- Omer AD, Lowe TM, Russell AG, Ebhardt H, Eddy SR, Dennis PP.** 2000. Homologs of Small Nucleolar RNAs in Archaea. *Science* **288**:517-522.
- Omer AD, Ziesche S, Ebhardt H, Dennis PP.** 2002. *In vitro* reconstitution and activity of a C/D box methylation guide ribonucleoprotein complex. *Proc Natl Acad Sci U S A* **99**:5289-5294.
- Ono M, Scott MS, Yamada K, Avolio F, Barton GJ, Lamond AI.** 2011. Identification of human miRNA precursors that resemble box C/D snoRNAs. *Nucleic Acids Res* **39**:3879-3891.
- Palade GE.** 1955. A small particulate component of the cytoplasm. *J Biophys Biochem Cytol* **1**:59-68.
- Pannucci JA, Haas ES, Hall TA, Harris JK, Brown JW.** 1999. RNase P RNAs from some Archaea are catalytically active. *Proc Natl Acad Sci U S A* **96**:7803-7808.
- Parker R, Siliciano PG, Guthrie C.** 1987. Recognition of the TACTAAC box during mRNA splicing in yeast involves base pairing to the U2-like snRNA. *Cell* **49**:229-239.
- Paul A, Tiotiu D, Bragantini B, Marty H, Charpentier B, Massenet S, Labialle S.** 2019. Bcd1p controls RNA loading of the core protein Nop58 during C/D box snoRNP biogenesis. *RNA* **25**:496-506.
- Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, Berney C, Bowser SS, Cepicka I, Decelle J, Dunthorn M, et al.** 2012. CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. *PLoS Biol* **10**:e1001419.
- Peano C, Pietrelli A, Consolandi C, Rossi E, Petiti L, Tagliabue L, De Bellis G, Landini P.** 2013. An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. *Microb Inform Exp* **3**:1.

- Peculis BA.** 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 BA, Steitz JA.** 1993. Disruption of U8 nucleolar snRNA inhibits 5.8S and 28S rRNA processing in the *Xenopus* oocyte. *Cell* **73**:1233-1245.
- Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N.** 1996. Requirement for Xist in X chromosome inactivation. *Nature* **379**:131-137.
- Podlevsky JD, Chen JJJL.** 2016. Evolutionary perspectives of telomerase RNA structure and function. *RNA Biol* **13**:720-732.
- Podlevsky JD, Li Y, Chen JJJL.** 2016. The functional requirement of two structural domains within telomerase RNA emerged early in eukaryotes. *Nucleic Acids Res* **44**:9891-9901.
- Pogacić V, Dragon F, Filipowicz W.** 2000. Human H/ACA small nucleolar RNPs and telomerase share evolutionarily conserved proteins NHP2 and NOP10. *Mol Cell Biol* **20**:9028-9040.
- Quinternet M, Chagot M-E, Rothé B, Tiotiu D, Charpentier B, Manival X.** 2016. Structural Features of the Box C/D snoRNP Pre-assembly Process Are Conserved through Species. *Structure* **24**:1693-1706.
- Quinternet M, Rothé B, Barbier M, Bobo C, Saliou J-M, Jacquemin C, Back R, Chagot M-E, Cianférani S, Meyer P, et al.** 2015. Structure/Function Analysis of Protein–Protein Interactions Developed by the Yeast Pih1 Platform Protein and Its Partners in Box C/D snoRNP Assembly. *J Mol Biol* **427**:2816-2839.
- Rajan KS, Chikne V, Decker K, Waldman Ben-Asher H, Michaeli S.** 2019. Unique Aspects of rRNA Biogenesis in Trypanosomatids. *Trends Parasitol* **35**:778-794.
- Rederstorff M, Huttenhofer A.** 2011. cDNA library generation from ribonucleoprotein particles. *Nat Protoc* **6**:166-174.
- Reichow SL, Hamma T, Ferré-D'Amaré AR, Varani G.** 2007. The structure and function of small nucleolar ribonucleoproteins. *Nucleic Acids Res* **35**:1452-1464.
- Richard P, Darzacq X, Bertrand E, Jádý BE, Verheggen C, Kiss T.** 2003. A common sequence motif determines the Cajal body-specific localization of box H/ACA scaRNAs. *EMBO J* **22**:4283-4293.
- Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, et al.** 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**:1311-1323.
- Rivera-Calzada A, Pal M, Muñoz-Hernández H, Luque-Ortega JR, Gil-Carton D, Degliesposti G, Skehel JM, Prodromou C, Pearl LH, Llorca O.** 2017. The Structure of

the R2TP Complex Defines a Platform for Recruiting Diverse Client Proteins to the HSP90 Molecular Chaperone System. *Structure* **25**:1145-1152.e1144.

**Rothé B, Back R, Quinternet M, Bizarro J, Robert M-C, Blaud M, Romier C, Manival X, Charpentier B, Bertrand E, et al.** 2013. Characterization of the interaction between protein Snu13p/15.5K and the Rsa1p/NUFIP factor and demonstration of its functional importance for snoRNP assembly. *Nucleic Acids Res* **42**:2015-2036.

**Rothé B, Manival X, Rolland N, Charron C, Senty-Ségault V, Branlant C, Charpentier B.** 2017. Implication of the box C/D snoRNP assembly factor Rsa1p in U3 snoRNP assembly. *Nucleic Acids Res* **45**:7455-7473.

**Rothé B, Saliou J-M, Quinternet M, Back R, Tiotiu D, Jacquemin C, Loegler C, Schlotter F, Peña V, Eckert K, et al.** 2014. Protein Hit1, a novel box C/D snoRNP assembly factor, controls cellular concentration of the scaffolding protein Rsa1 by direct interaction. *Nucleic Acids Res* **42**:10731-10747.

**Roy SW, Hudson AJ, Joseph J, Yee J, Russell AG.** 2011. Numerous Fragmented Spliceosomal Introns, AT–AC Splicing, and an Unusual Dynein Gene Expression Pathway in *Giardia lamblia*. *Mol Biol Evol* **29**:43-49.

**Rozhdestvensky TS, Tang TH, Tchirkova IV, Brosius J, Bachellerie J-P, Hüttenhofer 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 AG, Schnare MN, Gray MW.** 2006. A large collection of compact box C/D snoRNAs and their isoforms in *Euglena gracilis*: structural, functional and evolutionary insights. *J Mol Biol* **357**:1548-1565.

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

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

**Sagolla MS, Dawson SC, Mancuso JJ, Cande WZ.** 2006. Three-dimensional analysis of mitosis and cytokinesis in the binucleate parasite *Giardia intestinalis*. *J Cell Sci* **119**:4889-4900.

**Samarsky DA, Fournier MJ.** 1998. Functional mapping of the U3 small nucleolar RNA from the yeast *Saccharomyces cerevisiae*. *Mol Cell Biol* **18**:3431-3444.

**Sambrook J, Russell DW.** 2001. Molecular cloning: A laboratory manual.

- Saraiya AA, Li W, Wang CC.** 2011. A microRNA derived from an apparent canonical biogenesis pathway regulates variant surface protein gene expression in *Giardia lamblia*. *RNA* **17**:2152-2164.
- Saraiya AA, Wang CC.** 2008. snoRNA, a Novel Precursor of microRNA in *Giardia lamblia*. *PLoS Pathog* **4**:e1000224.
- Savino R, Gerbi SA.** 1990. In vivo disruption of *Xenopus* U3 snRNA affects ribosomal RNA processing. *EMBO J* **9**:2299-2308.
- Schattner P, Decatur WA, Davis CA, Ares M, Jr, Fournier MJ, Lowe TM.** 2004. Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res* **32**:4281-4296.
- Schmitt ME, Clayton DA.** 1993. Nuclear RNase MRP is required for correct processing of pre-5.8S rRNA in *Saccharomyces cerevisiae*. *Mol Cell Biol* **13**:7935-7941.
- Schnare MN, Cook JR, Gray MW.** 1990. Fourteen internal transcribed spacers in the circular ribosomal DNA of *Euglena gracilis*. *J Mol Biol* **215**:85-91.
- Schnare MN, Gray MW.** 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.
- Schnare MN, Gray MW.** 1990. Sixteen discrete RNA components in the cytoplasmic ribosome of *Euglena gracilis*. *J Mol Biol* **215**:73-83.
- Schubert T, Pusch MC, Diermeier S, Benes V, Kremmer E, Imhof A, Längst G.** 2012. Df31 protein and snoRNAs maintain accessible higher-order structures of chromatin. *Mol Cell* **48**:434-444.
- Schultz A, Nottrott S, Hartmuth K, Lührmann R.** 2006. RNA Structural Requirements for the Association of the Spliceosomal hPrp31 Protein with the U4 and U4atac Small Nuclear Ribonucleoproteins\*. *J Biol Chem* **281**:28278-28286.
- Schwartz S, Bernstein DA, Mumbach MR, Jovanovic M, Herbst RH, León-Ricardo BX, Engreitz JM, Guttman M, Satija R, Lander ES, et al.** 2014. Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell* **159**:148-162.
- Scott MS, Avolio F, Ono M, Lamond AI, Barton GJ.** 2009. Human miRNA Precursors with Box H/ACA snoRNA Features. *PLoS Comput Biol* **5**:e1000507.
- Sergiev PV, Aleksashin NA, Chugunova AA, Polikanov YS, Dontsova OA.** 2018. Structural and evolutionary insights into ribosomal RNA methylation. *Nat Chem Biol* **14**:226-235.



**Sharma S, Langhendries J-L, Watzinger P, Kötter P, Entian K-D, Lafontaine DLJ.** 2015. Yeast Kre33 and human NAT10 are conserved 18S rRNA cytosine acetyltransferases that modify tRNAs assisted by the adaptor Tan1/THUMP1. *Nucleic Acids Res* **43**:2242-2258.

**Sharma S, Yang J, van Nues R, Watzinger P, Kötter P, Lafontaine DLJ, Granneman S, Entian K-D.** 2017. Specialized box C/D snoRNPs act as antisense guides to target RNA base acetylation. *PLoS Genet* **13**:e1006804.

**Sibbald SJ, Archibald JM.** 2017. More protist genomes needed. *Nat Ecol Evol* **1**:0145.

**Sloan KE, Warda AS, Sharma S, Entian K-D, Lafontaine DLJ, Bohnsack MT.** 2017. Tuning the ribosome: The influence of rRNA modification on eukaryotic ribosome biogenesis and function. *RNA Biol* **14**:1138-1152.

**Soltys BJ, Falah M, Gupta RS.** 1996. Identification of endoplasmic reticulum in the primitive eukaryote *Giardia lamblia* using cryoelectron microscopy and antibody to Bip. *J Cell Sci* **109**:1909-1917.

**Song J, Logeswaran D, Castillo-González C, Li Y, Bose S, Aklilu BB, Ma Z, Polkhovskiy A, Chen JJ-L, Shippen DE.** 2019. The conserved structure of plant telomerase RNA provides the missing link for an evolutionary pathway from ciliates to humans. *Proc Nat Acad Sci* **116**:24542-24550.

**Sun Q, Zhu X, Qi J, An W, Lan P, Tan D, Chen R, Wang B, Zheng S, Zhang C, et al.** 2017. Molecular architecture of the 90S small subunit pre-ribosome. *eLife* **6**:e22086.

**Szewczak LB, DeGregorio SJ, Strobel SA, Steitz JA.** 2002. Exclusive interaction of the 15.5 kD protein with the terminal box C/D motif of a methylation guide snoRNP. *Chem Biol* **9**:1095-1107.

**Taft RJ, Glazov EA, Lassmann T, Hayashizaki Y, Carninci P, Mattick JS.** 2009. Small RNAs derived from snoRNAs. *RNA* **15**:1233-1240.

**Tang TH, Bachellerie JP, Rozhdestvensky T, Bortolin ML, Huber H, Drungowski M, Elge T, Brosius J, 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.

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

**Tessier LH, Paulus F, Keller M, Vial C, Imbault P.** 1995. Structure and expression of *Euglena gracilis* nuclear rbcS genes encoding the small subunits of the ribulose 1,5-bisphosphate carboxylase/oxygenase: A novel splicing process for unusual intervening sequences? *J Mol Biol* **245**:22-33.

- Tian XF, Yang ZH, Shen H, Adam RD, Lu SQ.** 2010. Identification of the nucleoli of *Giardia lamblia* with TEM and CFM. *Parasitol Res* **106**:789-793.
- Tollervey D, Lehtonen H, Jansen R, Kern H, Hurt EC.** 1993. Temperature-sensitive mutations demonstrate roles for yeast fibrillarin in pre-rRNA processing, pre-rRNA methylation, and ribosome assembly. *Cell* **72**:443-457.
- Torgerson PR, Devleesschauwer B, Praet N, Speybroeck N, Willingham AL, Kasuga F, Rokni MB, Zhou X-N, Fèvre EM, Sripa B, et al.** 2015. World Health Organization Estimates of the Global and Regional Disease Burden of 11 Foodborne Parasitic Diseases, 2010: A Data Synthesis. *PLoS Med* **12**:e1001920.
- Touz MC, Zamponi N.** 2017. Sorting without a Golgi complex. *Traffic* **18**:637-645.
- Tovar J, León-Avila G, Sánchez LB, Sutak R, Tachezy J, van der Giezen M, Hernández M, Müller M, Lucocq JM.** 2003. Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation. *Nature* **426**:172-176.
- Tran EJ, Zhang X, Maxwell ES.** 2003. Efficient RNA 2'-O-methylation requires juxtaposed and symmetrically assembled archaeal box C/D and C'/D' RNPs. *EMBO J* **22**:3930-3940.
- Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L.** 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**:562-578.
- Tůmová P, Hofštetrová K, Nohýnková E, Hovorka O, Král J.** 2007. Cytogenetic evidence for diversity of two nuclei within a single diplomonad cell of *Giardia*. *Chromosoma* **116**:65-78.
- Tůmová P, Uzlíková M, Jurczyk T, Nohýnková E.** 2016. Constitutive aneuploidy and genomic instability in the single-celled eukaryote *Giardia intestinalis*. *MicrobiologyOpen* **5**:560-574.
- Turner B, Melcher SE, Wilson TJ, Norman DG, Lilley DM.** 2005. Induced fit of RNA on binding the L7Ae protein to the kink-turn motif. *RNA* **11**:1192-1200.
- Uliel S, Liang X-h, Unger R, Michaeli S.** 2004. Small nucleolar RNAs that guide modification in trypanosomatids: repertoire, targets, genome organisation, and unique functions. *Int J Parasitol* **34**:445-454.
- Uzlíková M, Fulnečková J, Weisz F, Sýkorová E, Nohýnková E, Tůmová P.** 2017. Characterization of telomeres and telomerase from the single-celled eukaryote *Giardia intestinalis*. *Mol Biochem Parasit* **211**:31-38.
- Valadkhan S, Mohammadi A, Jaladat Y, Geisler S.** 2009. Protein-free small nuclear RNAs catalyze a two-step splicing reaction. *Proc Natl Acad Sci U S A* **106**:11901-11906.



**van Nues RW, Granneman S, Kudla G, Sloan KE, Chicken M, Tollervey D, Watkins NJ.** 2011. Box C/D snoRNP catalysed methylation is aided by additional pre-rRNA base-pairing. *EMBO J* **30**:2420-2430.

**van Nues RW, Watkins NJ.** 2016. Unusual C'/D' motifs enable box C/D snoRNPs to modify multiple sites in the same rRNA target region. *Nucleic Acids Res* **45**:2016-2028.

**Venema J, Vos HR, Faber AW, van Venrooij WJ, Raué HA.** 2000. Yeast Rrp9p is an evolutionarily conserved U3 snoRNP protein essential for early pre-rRNA processing cleavages and requires box C for its association. *RNA* **6**:1660-1671.

**Vestheim H, Jarman SN.** 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, Kiss T.** 2019. Cooperative 2'-O-methylation of the wobble cytidine of human elongator tRNA(Met)(CAT) by a nucleolar and a Cajal body-specific box C/D RNP. *Genes Dev* **33**:741-746.

**Vogan JM, Zhang X, Youmans DT, Regalado SG, Johnson JZ, Hockemeyer D, Collins K.** 2016. Minimized human telomerase maintains telomeres and resolves endogenous roles of H/ACA proteins, TCAB1, and Cajal bodies. *eLife* **5**:e18221.

**Vos TJ, Kothe U.** 2020. snR30/U17 Small Nucleolar Ribonucleoprotein: A Critical Player during Ribosome Biogenesis. *Cells* **9**:2195.

**Wagner EGH, Romby P.** 2015. Chapter Three - Small RNAs in Bacteria and Archaea: Who They Are, What They Do, and How They Do It. In: Friedmann T, Dunlap JC, Goodwin SF, editors. *Advances in Genetics*: Academic Press. p. 133-208.

**Wang C, Meier UT.** 2004. Architecture and assembly of mammalian H/ACA small nucleolar and telomerase ribonucleoproteins. *EMBO J* **23**:1857-1867.

**Watkins NJ, Bohnsack MT.** 2012. The box C/D and H/ACA snoRNPs: key players in the modification, processing and the dynamic folding of ribosomal RNA. *WIREs RNA* **3**:397-414.

**Watkins NJ, Dickmanns A, Lührmann R.** 2002. Conserved Stem II of the Box C/D Motif Is Essential for Nucleolar Localization and Is Required, Along with the 15.5K Protein, for the Hierarchical Assembly of the Box C/D snoRNP. *Mol Cell Biol* **22**:8342-8352.

**Watkins NJ, Gottschalk A, Neubauer G, Kastner B, Fabrizio P, Mann M, 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 NJ, Ségault V, Charpentier B, Nottrott S, Fabrizio P, Bachi A, Wilm M, Rosbash M, Branlant C, Lührmann R.** 2000. A Common Core RNP Structure Shared

between the Small Nucleolar Box C/D RNPs and the Spliceosomal U4 snRNP. *Cell* **103**:457-466.

**Watson JD, Crick FH.** 1953. The structure of DNA. *Cold Spring Harb Symp Quant Biol* **18**:123-131.

**Wilkinson ME, Charenton C, Nagai K.** 2020. RNA Splicing by the Spliceosome. *Annu Rev Biochem* **89**:1-30.

**Wilusz JE.** 2015. Removing roadblocks to deep sequencing of modified RNAs. *Nat Methods* **12**:821-822.

**Wu J, Niu S, Tan M, Huang C, Li M, Song Y, Wang Q, Chen J, Shi S, Lan P, et al.** 2018. Cryo-EM Structure of the Human Ribonuclease P Holoenzyme. *Cell* **175**:1393-1404.e1311.

**Xu F, Jiménez-González A, Einarsson E, Ástvaldsson Á, Peirasmaki D, Eckmann L, Andersson JO, Svärd SG, Jerlström-Hultqvist J.** 2020. The compact genome of *Giardia muris* reveals important steps in the evolution of intestinal protozoan parasites. *Microb Genom* **6**:mgen000402.

**Yang C-Y, Zhou H, Luo J, Qu L-H.** 2005. Identification of 20 snoRNA-like RNAs from the primitive eukaryote, *Giardia lamblia*. *Biochem Bioph Res Co* **328**:1224-1231.

**Yang JH, Zhang XC, Huang ZP, Zhou H, Huang MB, Zhang S, Chen YQ, Qu LH.** 2006. snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucleic Acids Res* **34**:5112-5123.

**Yang X, Duan J, Li S, Wang P, Ma S, Ye K, Zhao XS.** 2012. Kinetic and thermodynamic characterization of the reaction pathway of box H/ACA RNA-guided pseudouridine formation. *Nucleic Acids Res* **40**:10925-10936.

**Yang Z, Lin J, Ye K.** 2016. Box C/D guide RNAs recognize a maximum of 10 nt of substrates. *Proc Natl Acad Sci U S A* **113**:10878-10883.

**Yang Z, Wang J, Huang L, Lilley DMJ, Ye K.** 2020. Functional organization of box C/D RNA-guided RNA methyltransferase. *Nucleic Acids Res*.

**Ye K, Jia R, Lin J, Ju M, Peng J, Xu A, Zhang L.** 2009. Structural organization of box C/D RNA-guided RNA methyltransferase. *Proc Natl Acad Sci U S A* **106**:13808-13813.

**Yoshida Y, Tomiyama T, Maruta T, Tomita M, Ishikawa T, Arakawa K.** 2016. De novo assembly and comparative transcriptome analysis of *Euglena gracilis* in response to anaerobic conditions. *BMC Genomics* **17**:182.

**Yu G, Zhao Y, Li H.** 2018. The multistructural forms of box C/D ribonucleoprotein particles. *RNA* **24**:1625-1633.

**Yu Y-T, Meier UT.** 2014. RNA-guided isomerization of uridine to pseudouridine—pseudouridylation. *RNA Biol* **11**:1483-1494.

**Yu YT, Maroney PA, Darzynkiwicz E, Nilsen TW.** 1995. U6 snRNA function in nuclear pre-mRNA splicing: a phosphorothioate interference analysis of the U6 phosphate backbone. *RNA* **1**:46-54.

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

**Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JHD, Noller HF.** 2001. Crystal Structure of the Ribosome at 5.5 Å Resolution. *Science* **292**:883-896.

**Zamudio JR, Mittra B, Chattopadhyay A, Wohlschlegel JA, Sturm NR, Campbell DA.** 2009. *Trypanosoma brucei* Spliced Leader RNA Maturation by the Cap 1 2'-O-Ribose Methyltransferase and SLA1 H/ACA snoRNA Pseudouridine Synthase Complex. *Mol Cell Biol* **29**:1202-1211.

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

**Zhang L, Vielle A, Espinosa S, Zhao R.** 2019. RNAs in the spliceosome: Insight from cryoEM structures. *WIREs RNA* **10**:e1523.

**Zhao R, Kakihara Y, Gribun A, Huen J, Yang G, Khanna M, Costanzo M, Brost ReL, Boone C, Hughes TR, et al.** 2008. Molecular chaperone Hsp90 stabilizes Pih1/Nop17 to maintain R2TP complex activity that regulates snoRNA accumulation. *J Cell Biol* **180**:563-578.

**Zhong F, Zhou N, Wu K, Guo Y, Tan W, Zhang H, Zhang X, Geng G, Pan T, Luo H, et al.** 2015. A SnoRNA-derived piRNA interacts with human interleukin-4 pre-mRNA and induces its decay in nuclear exosomes. *Nucleic Acids Res* **43**:10474-10491.

**Zimorski V, Ku C, Martin WF, Gould SB.** 2014. Endosymbiotic theory for organelle origins. *Curr Opin Microbiol* **22**:38-48.

## **Appendix 1 – Supplementary Material for Chapter 2:**

**Analysis of *Giardia muris* ncRNAs reveals highly divergent spliceosomal, telomerase, and U3 RNAs and uncovers an RNase P-snoRNA diRNP**

**GlsR1** -----GTCCACTGGCCTCTCCTGAGGCAG**ATGATG**--**ACTTTGCGACGGGCGGA****CG** 49  
**GmsR1 (nc157)** CGACTTTACCGCGACCGGCCCTCCCCAAAAAGCA**ATGAGGA**TGCT**CGACGGGCGG**--- 57  
\* \* \* \* \*

**GlsR1** **GAGGGACGC****GTGACGA**AGTTTGTCTATT**CTGA**ATTCCTT 89  
**GmsR1 (nc157)** -**ACTGA**CTT**CTGAGGA**TGCAAGTCGATT**CTGA**ACCTT-- 94  
\* \* \* \* \*

**GmsR1 (nc157)** CGACTTTACCGCGACCGGCCCTCCCCAAAAAGCAATGAGGATGCTCGACGGGCGGACT 60  
**GmsR1 (nc162)** CGACTTTACCGCGACCGGCCCTCCCCAAAAAGCAATGAGGATGCTCGACGGGCGGACT 60  
\*\*\*\*\*

**GmsR1 (nc157)** GACTTCTGAGGATGCAAGTCGATTGCTGAACCTTCACTAAATGGCCTC 108  
**GmsR1 (nc162)** GACTTCTGAGGATGCAAGTCGATTGCTGAACCTTCACTAAATGGCCTC 108  
\*\*\*\*\*

**GlsR2** ---TGTAGCGAAC--CCACGCGCAAGCGTTGCTACGAGGCGA**TGGAGA**CAAAAGCAGTTA 55  
**GmsR2 (nc033)** GAGAAAAAGAAATGCGAGAACCCCAATTGACTCGCAGGAGTGA**TGCTGA**CAAGGGATTTTA 60  
\* \* \* \* \*

**GlsR2** CGTT-CGCAACTCT**CTGAGGGTTCC****TGATGC**TTCCCTGGATGT**CCGAGCCTT**---- 106  
**GmsR2 (nc033)** CGTCCGCCGACGAA**CTGA**GTGAGCA**TGATGC**TTCCCTGGATGT**TCGA**AGCTCCCTT 116  
\* \* \* \* \*

**GlsR4** TGTCTCCA**TGACGA**GAATTACGCCGCCCCAGT**CTGA**CCCC**TGAC**-GAACGGCTTCT**CTGA** 59  
**GmsR4 (nc135)** ----TTCGA**TGATGA**AGTATGCCGCCCCAGT**CTGA**TATCC**TGCTGA**CTGGCTTTT**CTGA** 55  
\* \* \* \* \*

**GlsR4** TCATT 64  
**GmsR4 (nc135)** CCTT- 59  
\* \*

**GlsR5** AATTAAAAGCTG**TGATGA**CAGGTTCTTGCCCCGT**ATGA**CCCTG**CGATGA**GTTATACAAA 60  
**GmsR5 (nc199)** -----A**TGATGA**GAATGCTTCTTGCCCCGT**CTGA**CACGA**CGATGA**GTGACGCCAAG 51  
\* \* \* \* \*

**GlsR5** GAACGCATCCAAGCCAACCG**CTGA**GCTCCTT--- 92  
**GmsR5 (nc199)** AA--GACTACAAGCCAACCG**CTGA**GCCACCCCTT 84  
\* \* \* \* \*

**GlsR6** ----A**ATGATGG**CTTGTTATCCCTGT**CTGA**GGTCAATACCT**TGATTA**GACGAT**TGACA** 55  
**GmsR6 (nc189)** AAAT**ATGATGA**GCGTTGTTATCCCTGT**CTGA**TAT**GTGACGA**-----AAACGA-----CT 49  
\* \* \* \* \*

**GlsR6** GAGCATCCTT 65  
**GmsR6 (nc189)** **CCTGA**CCCTT 59  
\* \* \* \*

**GlsR7** --CCGCGA**TGATTA**CCGAATCACAG**GCGA**TACACG**ATGAAGC**ACTCATAGTTACT**CTGAGC** 58  
**GmsR7 (nc013)** GATGGTGA**TGACGA**G--CCTAATCAAC**CTGA**CTC**CTGATGG**TGTCATAGTTACT**CGGAGC** 58  
\* \* \* \* \*

**GlsR7** GGTCTT-- 65  
**GmsR7 (nc013)** CCTCGCTC 67

**GlsR8** CGT **AGATGA** AGAGAGATAAATCAGCTACCG **CTGA** GCCCAACG **TGAGGA** --AGAAACCGCC 58  
**GmsR8 (nc160)** ACT **GAA** -- **TGAT** TGACGGTT **ACCAGCTACCGCTGA** GCCCAGCG **TGATGA** GCACAAACCACC 58  
 \* \* \* \* \*

**GlsR8** TTTCGT **CTGA** CCCTT 73  
**GmsR8 (nc160)** TTTCGT **CTGA** GCCTT 73  
 \* \* \* \* \*

**GlsR9** TAGCAACCCG **TGATTT** GCAACGCTTAGTCCGTGTT **CGGA** GTGTCTTGACGC **TGATGA** G 60  
**GmsR9 (nc018)** -CTACGGCCGA **TGACTG** TCAAGCTTGCTCCGTGTT **CGGA** GTGCTTCGTGCTT **TGATGA** G 59  
 \* \* \* \* \*

**GlsR9** TGAAAGCACAC **ATGA** AGGTTCCCTT--- 83  
**GmsR9 (nc018)** ATCAACAAGCT **AAGA** GGCCTTACCTT 85  
 \* \* \* \* \*

**GlsR13** ATCCATTTCGTATGAGATA **TGATGA** TTGGGAGCGACCTATC- **TTGA** GGACGACGCGCCGCC 59  
**GmsR13 (nc029)** ----- **GTGATGA** TTGGACTCTTCTGATCC **TTGA** TGAGGTCGGCTGCCT 43  
 \* \* \* \* \*

**GlsR13** GTCTTACCTTGTGACGTTTGCCGTCTTACAATGCT **CTGA** CCCTT 103  
**GmsR13 (nc029)** ACTTCACCTTGTAGTAG **TTGCCGA** -TGGAGGTGTT **CTGA** CCCTT- 85  
 \* \* \* \* \*

**GlsR14** AAA **TGATGA** CAATGCGCATTTGTGAGAAGG **CTCA** CTTCT **TGATGA** TTCCTCTGTCCATTCC 60  
**GmsR14 (nc093)** AAA **TGATGA** GAAAGTGCAATTTGTCTAAGGG **CTTA** CCGCT **TGAAGA** CACAAAGATCCATTCT 60  
 \* \* \* \* \*

**GlsR14** **CTGA** TCCTT 70  
**GmsR14 (nc093)** **CTGA** CC--- 67  
 \* \* \* \* \*

**GlsR15** AACCCGATTCAGACTACTCCTTGGTTCCTCGCAGA **ATGAT** --- **TA** TCTGTCTCCGAG--- 54  
**GmsR15 (nc092)** -----ACTCCTTACTTTCTCAAACG **ATGATGA** TTCTCTATAACTACGGCT 45  
 \* \* \* \* \*

**GlsR15** -----CAAG---CACGACTATGAGCTTACTTATGAGAT **CTGA** CTCCTT 94  
**GmsR15 (nc092)** ATGTGCACCGTGGAAGGGTGTGTGCCGATGGCTTATGAGAT **CTGA** CCCTT- 94  
 \* \* \* \* \*

**Candidate1** -----CAAAAGCAGACGAAAAAATAAA **TGAAGA** CAGAAC **CACAGACCTGTA** 46  
**GmCandidate1 (nc096)** CGGTGAGACGTGGAACGTGTCCGCCATGAAAAAA- **TGAGGA** CTCAAC **CGCAGACCTGTG** 59  
 \* \* \* \* \*

**Candidate1** **CTGA** CCCTT **TGATGT** TAGTGTGCGCT **CTGA** TATCCTT 82  
**GmCandidate1 (nc096)** **CTGA** CCTTTGC **CTGATGA** CG-CGTTGTACT **CTGA** CCTT 94  
 \* \* \* \* \*

**Candidate13** ----- **TGAT** -- **TA** CTCCAACACGACGGTCTA **CTGA** GAACCCAGTATCTTTAGACTGCTG 52  
**GmCandidate13 (nc012)** GTGAT **GCGATGA** TTTCAACACACGACGGTCTT **CTGA** GTGAACAACAG-----CCTGCTG 54  
 \* \* \* \* \*

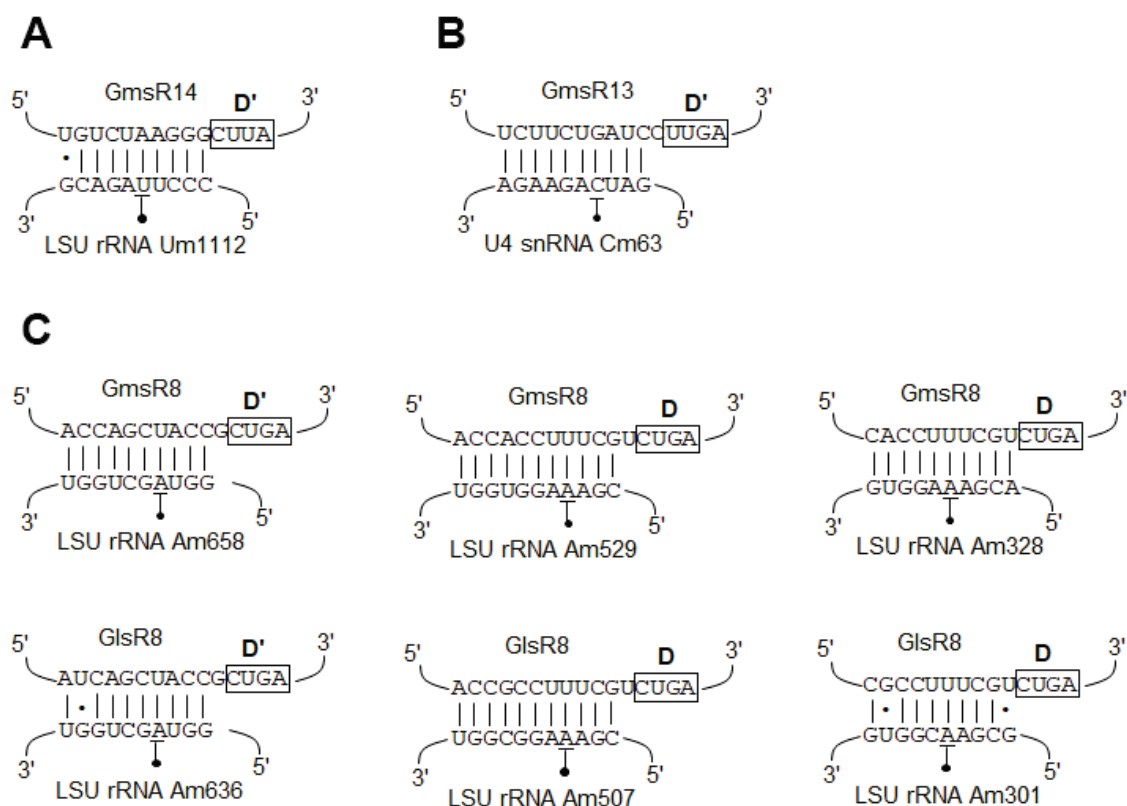
**Candidate13** AGACAGTGTAT **ATGA** TTTCCTT 75  
**GmCandidate13 (nc012)** ACCCTGTGTTT **ATGA** GCCTT-- 75  
 \* \* \* \* \*

## H/ACA snoRNAs

<b>GlsR17</b>	GTGAGGATCCGGGGCACTGAGCAATCCCCAGGACACAGGC	60
<b>GmsR17 (nc140)</b>	ATAGTGATGCGGGGCACTGAGGGATCCGTAGGGGAGAGGAGGGCTGGTGGGCATGGTCGC	60
	*    ***    *****    *****    *    ***    *    *    *    *    *	
<b>GlsR17</b>	GCCACGCAGCCTAATCACC GCCCCTATAGTCCTT	94
<b>GmsR17 (nc140)</b>	ATC-GGCGGCCAATCACCAGTCTCACATCCTT	92
	*    *    *    *    *    *    *    *    *    *    *	
<b>GlsR18</b>	-CCGCTGGCGCTTGCAGC-----GTGCACAGGCCTACATCCAGGGTCATAGGTGGGGAG	54
<b>GmsR18 (nc141)</b>	CCGGCGACAGACTGCAATGCGCGCCCCGCTGGCCAATAGG-ATGGACCTTGGTGGGT--	57
	*    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *	
<b>GlsR18</b>	CGGATCCCGTCCATCCTCAATCCGGGCCCGCACATCCTT	94
<b>GmsR18 (nc141)</b>	CGGACGCAGTCTGGGCTCAATCCAGTCCGCAAACCTT---	94
	*****    *    ***    *****    *    *    *    *	
<b>GlsR19</b>	AAAAGCAAGCAGAAGCCCAGTTTGGTCTCTACCGGCGTATGCATGTGCATAGGCTGGCCA	60
<b>GmsR19 (nc202)</b>	GTACTCAACCGTGACCTCCGTTCTGGAGCTATCGGGCGGGCGTGGCGTGGGACGGGC-G	59
	*    ***    *    *    *    *    *    *    *    *    *    *    *    *    *	
<b>GlsR19</b>	AGCATCGTT-----GATAGAAAGCTGCTCTTGGTCACCGAGGGTCTCCGGTTTC	109
<b>GmsR19 (nc202)</b>	AGGAGACTTGCGCCTGCACGACGAAGATCCGCTGTCTTTGATCACCGCGGAGCGGTTTC	119
	**    *    **    *    *    *    *    *    *    *    *    *    *    *	
<b>GlsR19</b>	ATACGCAGAGACATCCTT 127	
<b>GmsR19 (nc202)</b>	AGCGGCAGGAACACCTT- 136	
	*    ****    ***    *    *	
<b>GlsR20</b>	-----AAAAATGCCAGCTGAGTTACGTCTGTGTGCACAGGCGCGTCAGAGGCCGG-CTAG	54
<b>GmsR20 (nc011)</b>	GGCCGCGTTGCGCTCACAAAGATGCTTCTGTGAG-----CGCGTCAACGCGGGCGAG	54
	*    *    *    *    *    *    *    *    *    *    *    *    *    *	
<b>GlsR20</b>	AGCGCGACTGGTTGAGTTCCCAGA--GCGATCTGGGTGATTAGCAGTCATACATCCTT	111
<b>GmsR20 (nc011)</b>	ACCTCGGCTGGTGAGTATCTGCGCCTTTGGCGGAGACAATTGACGCGCCGATAGACCTT	113
	*    *    *    *    *    *    *    *    *    *    *    *    *    *	
<b>GlsR22</b>	ACGCAAGCCCTCTAGCAAGATGCAGGCCGAGCCTGTGTCTCGTTCCCTGGGGCGATAGC	60
<b>GmsR22 (nc192)</b>	--TGGAGCCCTACAGCAAGATGCATGCCGGAGCATGTGTCTCGTTCCCTTGGGCCAGAGC	58
	*****    *****    *****    *****    *****    *****    *    *    *	
<b>GlsR22</b>	TCTTGCTGGCAGGTCTTGCAGTGTCCATACCCGGCAACACGTTTCCAGCTACACCTT	119
<b>GmsR22 (nc192)</b>	GATGCCTTGTAACCTTTGAGTGCTGGGCA-----CACACCTT-----	94
	*    *    *    *    *    *    *    *    *    *    *    *    *	
<b>GlsR27</b>	AGCTCACCCAAAGTCAACGGAGCG-CCAGCTACGTGTTATGGGCAGCGAAAGTACCAGAG	59
<b>GmsR27 (nc095)</b>	-ATGCATGTGCAGTCAAGCAGCTCCTGGAGATCGTGTTAGGCCATGCGATAGGCGCCAGAG	59
	*    *    *    *    *    *    *    *    *    *    *    *    *	
<b>GlsR27</b>	CCAAAGAGTTCCTCT--GATCGCCTGGCCGGAGCACATTTGTG-----ATCTCCTA	108
<b>GmsR27 (nc095)</b>	CCGCCCCGGTCTTCCGGGAATGTTGAGGCTCGCGG-TATCTCGAAGCCCGAGTGCCGCCA	118
	*    *    *    *    *    *    *    *    *    *    *    *    *	
<b>GlsR27</b>	TACCTT 114	
<b>GmsR27 (nc095)</b>	CACCTT 124	
	*****	

<b>Candidate16</b>	AGATCAAAAGCAAGGCTAGAGCCATGGAGCGCGGATCTGCGCTCTGCCAGATACGCCGAC	60
<b>GmCan16 (nc036)</b>	-----CCTGCCACGACGGCCGAGACCGGATCGTGCCTCGCGCAC-----GGCC	43
	* * * * * * * * * * * * * * * * * *	
<b>Candidate16</b>	AGAAAGCACCAAGGAAGGATGT-GGATCTCCATGTCTGCCGTGTGCGCGC <u>ATA</u> TCCTT--	117
<b>GmCan16 (nc036)</b>	TGCGCCCGTCTAGGAAGGGGGCCGTCTGACGACTCCTGCCAACGACGCAG <u>ACA</u> GCCTCCA	103
	* * * * * * * * * * * * * * * * * *	
<b>Candidate16</b>	-----	117
<b>GmCan16 (nc036)</b>	ATCAACTTCCCTT	116

**Figure A.1.1. Most snoRNA homologs from *Giardia muris* and *Giardia lamblia* diverge significantly outside of guide regions.** Pairwise alignments of homologous snoRNAs from *G. lamblia* (Gl) and *G. muris* (Gm). Candidate# labels are *G. lamblia* homologs. For C/D snoRNAs C, D, C', and D' boxes are highlighted in yellow where present and anti-sense element guide regions that pair to target RNAs are underlined. For H/ACA snoRNAs H box and ACA elements are highlighted in yellow where present and both sides of the bipartite guide regions that base-pair to rRNA are underlined. For both RNA classes the *G. muris* GeneID (nc#) for putative ncRNAs used during our analysis are indicated following the name of the snoRNA (These correspond to GeneID numbers found in Supplemental File 1).



**Figure A.1.2. Predicted base-pairing of *G. muris* C/D snoRNAs to RNA targets.**



**Figure A.1.2. Predicted base-pairing of *G. muris* C/D snoRNAs to RNA targets.** Base-pairing predictions for C/D snoRNAs GmsR14 (**A**) GmsR13 (**B**) that are not conserved in *G. lamblia*. The GmsR14 targeting uses an anti-sense element upstream of a degenerate D' box to guide the modification. Proposed degenerate D' boxes are boxed and labeled. (**C**) Conserved base-pairing for the three unique targets for GmsR8 and GlsR8 snoRNAs in their respective species rRNAs. Target rRNA nucleotides are underlined and indicated by a line connected to a •.



**Figure A.1.3. *G. muris* H/ACA snoRNA candidates form conserved dual hairpin structures.** Secondary structure predictions for newly identified H/ACA snoRNAs from *G. muris*. Predicted using MFOLD webserver and manual curation. Nucleotides that are part of the H box and ACA sequences are in red.

### 18S rRNA alignment

Gmuris_18SrRNA	CAUCCGGUUGAUCCUGCCG-GAGUACUACGCUACCCCCAAGGACAAAAGCCAUGCAAGCG	59
Gutheta_nucleo_18S	-ACCUGGUUGAUCCUGCCAGUAGUCAUAUGCUUGUCUUAAGGAUU-AAGCCAUGCAUGUC * * * * *	58
Gmuris_18SrRNA	-----GACACGAG-----GUAUGAAGUGGCGGACGGCUCGGUACAACGGUACGAGUCU	107
Gutheta_nucleo_18S	UCAGUACAAACAAGAUUGUACCGUGAAACUGCGAAUGGCUCAAUACAACAGUUAUGUU * * * * *	118
Gmuris_18SrRNA	GACCGGGGGUGA-AGGCUAGACGGAUACCGCUGGCAA-CCCAGCGCCAAGACG-----	158
Gutheta_nucleo_18S	AUUUGAUCGUGCGGGGCUACGUGGAUAACCGUAGGAAUUCUAGGCUAAUACAUAGCACCA * * * * *	178
Gmuris_18SrRNA	-----	158
Gutheta_nucleo_18S	UCGCUGUCAGGGACGGUCUAUACCGAGAACGCCUUUCAUCUGUAGGGCGUUCUCUGACGA	238
Gmuris_18SrRNA	-----AGUGCUC---AGAGCGGGGAAG-----G-----	179
Gutheta_nucleo_18S	ACCGGUGCCCCGUGUUUUUGGCUAGUCUCUAGAGGAGAGUUAACGAAAAGGUU * * * * *	298
Gmuris_18SrRNA	-----	179
Gutheta_nucleo_18S	GAGCCACCAAGGUCGUUUUCUAAACUUCGUUGGCAUUGAUUCGUCAACCCUUAACUGCA	358
Gmuris_18SrRNA	-----AAAGCACG-----CGAUG---GAGCGAAUGCCCCG	205
Gutheta_nucleo_18S	GCGGGAUUUAUAGAUCAAAAACCAUACGCAUCUCCCUUCCAUUGGGCGAGUGCGGGCGG * * * * *	418
Gmuris_18SrRNA	GAUGA-----	210
Gutheta_nucleo_18S	GAUGAAUCAUAUAACUUUUUUUUUUUUUUUUUCCGAACCGUAUGGCUUCCAAUUCUCA * * * * *	478
Gmuris_18SrRNA	-----GGUUCCGAGGUUUUACC--UAGUCGGUAGAGUAGUGGU	246
Gutheta_nucleo_18S	UGGCCGGCGGUAAAUCAUUCAAUUUCUGCCCUAUAACUUUCGAUGGUUGGGUAGUGGC * * * * *	538
Gmuris_18SrRNA	CUACGGAGGGGAUGAUGCCUGGCGGAGGAUCAGGGUUUGACUCCGGAGAACGGGCCUGAG	306
Gutheta_nucleo_18S	CAACCAUGGUGUUGACGGGU-ACGGGGAUUAGGGUUCGAUUCGGAGAGGGAGCCUGAG * * * * *	597
Gmuris_18SrRNA	AGACGGCCCGUACAUCGCAAGGACGGCAGCAGGCGCGGAACUUGCCCAAUGCGUGAAGGCG	366
Gutheta_nucleo_18S	AGAUGGCUACCACAUCAAGGAAGGCAGCAGGCGCGAAAAUUACCCAAUCC-UGAUUCAG * * * * *	656
Gmuris_18SrRNA	UGAGGCAGCAACGGGGGAUCCC-----AUGAA-----AUGGGAAGACUG	405
Gutheta_nucleo_18S	GGAGGUAGCGACAAGAAUAACGGUAGUGGACUCGUUCGAGUCUGCUAUCGGAA--U- * * * * *	712
Gmuris_18SrRNA	GGGGUAGAUGACCCC-----AGCA-CAAGUCGAGGGAAGGUCUGGUGCCAGCAGC	456
Gutheta_nucleo_18S	GAGGAUAGUUUACCCCCUUCUCGAGGAUCCAUGGAGG-GCAAGUCUGGUGCCAGCAGC * * * * *	771
Gmuris_18SrRNA	CGCGGUAAUCCAGCUCGGCAGGCGUCGUACGGCGCUGUUGCAGUUAACGUCGGAGU	516
Gutheta_nucleo_18S	CGCGGUAAUCCAGCUCCAAUAGCGUAUACUAAAGUUGCUGCAGUUAACGUCGUAGU * * * * *	831
Gmuris_18SrRNA	CGAGACGUCCAGCCGGGAGGAAAGAGGAGCGCU-----UAAG-----GC-----	555
Gutheta_nucleo_18S	CGAAUUGGAUCAUAGAGAGGAGUGCGCGCACUCCUCUGGUUGAGUGGACGCGCGCG * * * * *	891

Gmuris_18SrRNA	-----GGGAGUGAGU---ACGAGAAAGCC-----	576
Gutheta_nucleo_18S	CUCUUUCUGUGAUCAACCUUUCUGGGAGGGUCUCUGAAGAGGAUGCCUCUGUAUUCAGAC	951
	***** * * * * *	
Gmuris_18SrRNA	-----CGGGACGGACAUGAAGGU-----	594
Gutheta_nucleo_18S	CCCUCUUGGAUGUAUUUUACUGUGAACAAAUAAGAGUGUUAUGGCAGGCCUUUGAAUU	1011
	*** * * * *	
Gmuris_18SrRNA	-----	594
Gutheta_nucleo_18S	AGGCUUAGAUACCAUAGCAUGGAAUAUAAGAAUAGGACUUCGGUUCUGUUUGUUGGUU	1071
Gmuris_18SrRNA	-----GAAUGGGUAAGGGCAUGUGUAUUGGUGGGGACG	628
Gutheta_nucleo_18S	UUUGGAGCCGAGGUAUAUGAUUACAGGGAUAGUUGGGGCACCCGUACUCCGUUGUCAGA	1131
	* * * * * * * * *	
Gmuris_18SrRNA	GGUGAAAUAAGGAUGAUCCGACCAAGACAGACAAAGGCGUAGGCACUUGCCAAGACCAUUA	688
Gutheta_nucleo_18S	GGUGAAAUAUGGAUUUAACGGAAGACGAACAACUCGCGAAAGCAUCUGCCAAGGAUGUUU	1191
	***** ** * * * * *	
Gmuris_18SrRNA	CAGUCGAACCAGGACGAAGCCCAGGGGCGAGAAGGCGAUUAGACACCACCGUAUUCGCG	748
Gutheta_nucleo_18S	UCAUUGAUCAAGAACGAAAGUUAGGGGAUCCGAGACGAUUAAGAUACCGUCGUAGUCUUA	1251
	* * * * * * * * *	
Gmuris_18SrRNA	GCGUAAACGAUGCCACCGAGAGACUGGCCAG-----GUCGUCAGGAUC	791
Gutheta_nucleo_18S	CCAUAACUAUGCCAACUGGGGAUUGGUAGAAGUACCAAAUAAACGACUCUAUACGACCC	1311
	* * * * * * * * *	
Gmuris_18SrRNA	-GAAGGGAAACCGAUCAGGGUACGGGCUCUGGGGGGAGUAUGGCCGAAGGCUGGAACUU	850
Gutheta_nucleo_18S	CUAAGAGAAAUCAA--AGUCUUUAGGUUCUGGGGGGAGUAUGGUCGCAAGGCUGAAACUU	1369
	* * * * * * * * *	
Gmuris_18SrRNA	GAAGGCAUUGACGGAGGGGUACCACCAGACGUGGAGUCUGCGGCUCAAUUGACUCAACG	910
Gutheta_nucleo_18S	AAAGGAAUUGACGGAAGGCGACCACCAGCGUGGAGCCUGCGGCUUAAUUGACUCAACA	1429
	*** * * * * * *	
Gmuris_18SrRNA	CGA-ACACCUUACCAGGCCAGACGUACGGAGGAUCGACG-GUUGAGAGGACCUUCGUGA	968
Gutheta_nucleo_18S	CGGGGAACCUUACCAGGUCCGGACAUAAGGAGGAUUGACAGAUUGAGAGCUCUUCUUGA	1489
	** * * * * * *	
Gmuris_18SrRNA	UCGUACGAGUGGUGUGCAUGGCCGUUCACAGCCCUGGCUUGAGCCGUCUGCUUGACUG	1028
Gutheta_nucleo_18S	UUCUAUGGGUGGUGUGCAUGGCCGUUCUUAUUGGUGGAGUGAUUUGUCUGGUUAUUC	1549
	* * * * * * * * *	
Gmuris_18SrRNA	CGACAACGAGCGAGACCCUAACCUGGA-----UGGGACCGCCA-----AU	1068
Gutheta_nucleo_18S	CGUUAACGAGCGAGACCUUGACCUGCUAUUAAGCCCCGUCGGUGGCAACGCCUUCGUGAU	1609
	** * * * * * *	
Gmuris_18SrRNA	GGUGA-----AU-----UGGAGGAAGGUGGGGCGUAACAAG	1100
Gutheta_nucleo_18S	GGCUUCUAGAGGACGAUCCGCGUUAUAGUGGAUGGAGGA--AUGAGGCAUAACAAG	1667
	** * * * * * *	
Gmuris_18SrRNA	UCUGUGAUGCCCUUAGACGCCUGGGCUGCACGCGUACACACUGUGGG---AUGAAA	1156
Gutheta_nucleo_18S	UCUGUGAUGCCCUUAGAUUCCUGGGCCGCACGCGCUACAAGGUGCAGACAAUGAGU	1727
	***** * * * * *	
Gmuris_18SrRNA	CCACGUCG-----AGUUGUGAAGCUUGAUGAGAUCAAACCCACGUGGUUGG	1204
Gutheta_nucleo_18S	GCUGCUCCUGGUCCGAAAGGAUUGGGGAUCUUG-----GAACUGCAUCGUAUAGG	1780
	* ** * * * * *	
Gmuris_18SrRNA	GAUCGUGGACUGGAACG--UCCUCGUGAACCGGAUGUCUAGUAGGCGUAGGUCAUCA	1262
Gutheta_nucleo_18S	GAUAGAUGAUUGCAUUUUUCAUCUUGAACGAGGAUGCCUUGUAAGCGCGAGUCAUCA	1840
	*** * * * * *	
Gmuris_18SrRNA	UCUACGCCGGAUACGUCCCGGCCCUUGUACACACCGCCGUCGCUCCUACCGACUGGGU	1322
Gutheta_nucleo_18S	CUCGCGCUGAAUACGUCCUGCCCUUUGUACACACCGCCGUCGCUCCUACCGAUUGGAU	1900
	*** * * * * * *	
Gmuris_18SrRNA	CUUCUGGCGAGCUCCUGGGAGGGAUGAACCGAACAGGGACGAAC-----	1366
Gutheta_nucleo_18S	GGCCCGGUGAAAUGCUCGGACGAGUGUUCAGAGCCGCACGAUCGUGGCGCUCUCCUCG	1960
	* * * * * * *	

Gmuris_18SrRNA	-----CGCGAGGCU-----UGGAGGAAGGAGAAGUCGUAACAAGGUAUCCGUAGGU	1412
Gutheta_nucleo_18S	GAAGUGCAGUGAGCCUUGUCAUCUAGAGGAAGGAGAAGUCGUAACAAGGUUUCGUAGGU	2020
	* * * * *	
Gmuris_18SrRNA	GAACCUGCGGAUGGAUCCAUC	1433
Gutheta_nucleo_18S	GAACCUGCGGAAGGAUCAUG-	2040
	***** *	

## 28S rRNA alignment

Gmuris_28SrRNA	UCCCCCCCCACUCCGAUGAAGAUGACGGGUGGAACUUAAGCAUAUCAGUACGCCCAGGAG	60
Gutheta_nucleo_28S	CACGGUCUCAGGUCGGCGGGAAUACCCGUCGAACUUAAGCAUAUCAUAGACGGAGGAA	60
	* * * * *	
Gmuris_28SrRNA	AAGACACCAACCGGAUUCUUGAUUAGCGGCGAGCGAUCCAGGAGUAGUCCGACCUCG-	119
Gutheta_nucleo_28S	AAGAAACUAACCGAUUCCUUGAGUAGCGGCGAGCGAAUCGGGAUCAGUCUACAAGUUUU	120
	***** * * * * *	
Gmuris_28SrRNA	AAG-----CGAUGAG---CGUUGU-----GAGGACUGGGGG	147
Gutheta_nucleo_28S	AAUCGGCAGUGGUUCCUAUGAUCCACCGCUGAGAUGUAGUCCUAAGUGGCGCACUCGGGA	180
	** * * * *	
Gmuris_28SrRNA	GAUGUCCUAAACCGAAGAUGGGUUGAAAC-CCACACCAAGGAGGGUGCCAGUCCCGUAG	206
Gutheta_nucleo_28S	GAAGUGGUUAUUAU--AAG-UCUUUUGGAACAAAGCGCCAUAGAGGGUGAUAGCCCCGUU	237
	** * * * *	
Gmuris_28SrRNA	GGAUGGACACAA-----CCCCAACGACGAGUACCUCUGCU	241
Gutheta_nucleo_28S	GUGUAUACCAAGUCGUCUGAGCCUCUCCUGCCGAUGCGCCACUCGUGAGUCGGUUGCU	297
	* * * * *	
Gmuris_28SrRNA	UGAGAGUGUAGAGGGAAGAGGAGGUGGUACCCU--UCUAAGGCUAAAUACCGCCCCGAG	298
Gutheta_nucleo_28S	UGGGAGUGCAGCCUAAAUAG--GGUGGUUAUACUUAUCCAAAGCUAAAUAUGGACAGGAG	355
	** * * * *	
Gmuris_28SrRNA	ACCGAUAGAGGACCAAGUAGUGCGAACGAAAGGUGAGAAGGAUGCCGA-----	346
Gutheta_nucleo_28S	ACCGAUAGCG-AACAAGUACCGCGAGGGAAAGAUGAAAAGAACUUUGGAAAGAGAGUGAA	414
	***** * * * * *	
Gmuris_28SrRNA	-----CCC-----AGGCACGUCAAAAGACCCU	368
Gutheta_nucleo_28S	AUAGUGCUUGAAACAUUGGGGGAAAGCGAAUGGAGCUUCCGUAAUAUCGAAGGAUCUC	474
	* * * * *	
Gmuris_28SrRNA	GAACCCG-----G-----	376
Gutheta_nucleo_28S	GAGCCAGCGCUUUGGUUGGAUCUGGUGUAAUGACUUCGAUGAUGCGUCCGUGAGUUUUU	534
	** * * *	
Gmuris_28SrRNA	-----GA	378
Gutheta_nucleo_28S	CCCUUCUCGUGCGUGCGUUGAACACAACGAAAGAAUUCACUGGCUUCUCCUCCGUGUGA	594
	**	
Gmuris_28SrRNA	GG-----CGA-----	383
Gutheta_nucleo_28S	GGCAGAAGCUUUGAGCUGGAGUCUUUCCUUUUUGUUGUUGUUGAAUCGACACAGAAGGG	654
	** * * *	
Gmuris_28SrRNA	-----CUCCGACCUGUCGAU-----	399
Gutheta_nucleo_28S	CUCUUCGUGAGAGUUUCGAUUUUUGUUUCCACCUCUCCCUUUUUUUUCCAAAGGUUGG	714
	** * * * * *	
Gmuris_28SrRNA	-----GA-----UCGAUUGGUCGGGCCGUCUCGAACACGG	431
Gutheta_nucleo_28S	UGGAGGUCCAAGAUAUGGCAAUGGACGUAGGGGCUCCUUCGACCCGUCUUGAAACACGG	774
	* * * * *	
Gmuris_28SrRNA	ACCGAGGAGUUAGGACCGAUCGCGAGUCAUGUGGCAAA-----GGG-A	473
Gutheta_nucleo_28S	ACCAAGGAGUCCGGUAUGAGCGCGAGUCUCGGGAAAGUUACACCCAUGCGCGUAGCGAA	834
	** * * * *	
Gmuris_28SrRNA	AGUCAAGACUU-----	484
Gutheta_nucleo_28S	AGCAAUGACUGGAAUCUGUUAUCGAGAACGCUGAUAGACAUUCCCCUCGUUACGAGAGA	894
	** * * * *	

Gmuris_28SrRNA	-----CAAGGGGCCCCUACA-----A-----	501
Gutheta_nucleo_28S	UCUCUAGAGUCUUUCAUUCUCCUUGAAUAGGUAGGUUCUUAAGGUGAGAAAUUUGGC	954
	*** * **** *	
Gmuris_28SrRNA	-----GG-----GC	505
Gutheta_nucleo_28S	UAAGCUCCAUAUUGCAUCAUCGUCCAACCUAUCUGCCCUUAGGGUCGAUUGGCUUGUGAU	1014
	**	
Gmuris_28SrRNA	AGAGCGAUCGGUCCUAGACCCGAAGGUGGUGAUCUACACUUGACCAGGUGAAGCCAGA	565
Gutheta_nucleo_28S	UGAGCGCUCAUACUGGGACCCGAAGAUGGUGAACUAUGCCUGAGCAGGAUGAAGCCAGA	1074
	***** ** * ***** **** * ** * **** *****	
Gmuris_28SrRNA	CGAAAGCCUGGUGGAGGCCAUCCCGUGUGACGUGCAAUCGUCUGGUCGAGUUGAGU	625
Gutheta_nucleo_28S	GGAACUCUGGUGGAGGUCCGAAGGCGUUCUGACGUGCAAUCGAUGCACAGACUUGGGU	1134
	*** ***** ** ** ***** **** *	
Gmuris_28SrRNA	GUAGGGGCGAAAGACUCAUCGAACCACCGGUAGCUGGUUACCUCGAAUGUCUCCAG	685
Gutheta_nucleo_28S	AUAGGGCGAAAGACCAUCGAACCGUCUAGUAGCUGGUUCCUCCGAAGUUUCCUCAG	1194
	***** ***** ** ***** ***** * * * *	
Gmuris_28SrRNA	GAUAGCCGACAUCAUGAACAGCUGCCCUAGGUAAAGGC-CACGAUCGGUGGGAGUAGGGAG	744
Gutheta_nucleo_28S	GAUAGCUGGAGCUG-UAGCAGUCUUAUCCGGUAAAGCGAAUGAUUAGAGGAUCCGGAGAG	1253
	***** * * ** ***** ** * ** * * * *	
Gmuris_28SrRNA	GU--CACUCCUUAUCCACCCUCGAACCAUGAACAAAGGGUGGAGCGGGGCCAACUUGUGU	802
Gutheta_nucleo_28S	UCCAAACUCUUUGACCUAUUUCUAAACUUUAAACAGGUAAGG-----AAGUUGUGU	1304
	*** ** * ** * ** * ** *	
Gmuris_28SrRNA	GGGCCCCUCAACAG-----UGUUGA-UGUCGAGUGGGCCUCUCCUGGUAAAGCAG	851
Gutheta_nucleo_28S	UGC UUAGAUGAACACAGCUGGACGAUGGUUAGCUCUUAGUGGGCCAUUUUUGGUAAAGCAG	1364
	* * **** ** * *	
Gmuris_28SrRNA	GACGGGCAAGACGGGAUGAACCGACAGUCGGGGGAAGGUGUGGAAGUGUAUGCUCGAU--	909
Gutheta_nucleo_28S	AACUGGCGAUGCGGGAUGAACCGAGAGUGGUGUUAAGGCGCCCAACUACGCGCUCAUUAAG	1424
	** *** * ***** ** * * **** * ** * **** *	
Gmuris_28SrRNA	-----CAAACGGUGUCCGUCGAUCGUGACAGCUGGAAGGUGGCCUUAACAGUAGGAAGC	963
Gutheta_nucleo_28S	AUCCAGGAAAGGUGUAGGUGCGUUAUGGACAGCAGGACGGUGGUCAUGGAGGUCGACAUC	1484
	* * ***** * * ***** ** * ** *	
Gmuris_28SrRNA	CUCUAAGGAGUGUGUAACAACCCACCAGCCGAAUCGAGGGGCCCGGAAAAUGGAACACGC	1023
Gutheta_nucleo_28S	CGCUAAGGAGUGUGUAACACUCACCUGCCGAACGCAUCAGCCUGAAAAUGGAUGGCGC	1544
	* ***** ** * **** ***** *	
Gmuris_28SrRNA	CGAAGCAUACGAUCGGAACCCGACCG-----AGAUGUGUCUC---GGGUAGGAGG	1070
Gutheta_nucleo_28S	UCAAGCGGUGGCCGAUACACCACCGCCGCGUACAGAAUUGUCCGUGUGUAGGAGG	1604
	*** * ** * * **** ** * ** *	
Gmuris_28SrRNA	UCGUCAUGA--UCGAGUCGAAGCCAU-GGGGUGACACGUGGUGGAUCGAGUCAUGAUUGC	1127
Gutheta_nucleo_28S	GCGUCGUGCCUUUGCGUCGAAGCCUUUGCGUGAGCCUUGGUGGAGC-AGGCUCGAGUGC	1663
	*** ** * * ***** * ** * ** * ***** * * * *	
Gmuris_28SrRNA	UGAUCUCGGUAGUAGUAGCCAUUACUCC-----AUCUGGAGGCCUGACGUGGAGAUGG	1180
Gutheta_nucleo_28S	AGAUCUUGGUGGUAGUAGCAAAAUUAAGCGAGAAUCUUGAGGACUGGAGUGGAGAAGG	1723
	***** ** ***** * * ** ***** **** * ** ***** **	
Gmuris_28SrRNA	GUUCCGUAUCC-CUGUCGAUCAGAUACGGGUGACGACGAUCCUAAGUCCGAGUGUGUACGC	1239
Gutheta_nucleo_28S	UUUCCAUGCAAACUGUG-UUUGAGCAUGGUGAGUCGAUCCUAAGAGAA-UGGGUAAUUC	1781
	*** * **** * * ***** ***** ** * *	
Gmuris_28SrRNA	CGUACCA-----UGGGGUCAUCCUCACAGGACGAUAGGGCUGACCGGUUAAACAUCCCG	1292
Gutheta_nucleo_28S	CGUUAGAGUUGUCUUCGGGCAUGUCCA---UCGA-AAGGGAACGGGUUAAUUAUCCCG	1837
	** * * ** * ** * ** * ** *	
Gmuris_28SrRNA	GUGGUAUGAUGGGGUCUUGG--UGACAUCUUU--CACAAACGGUACAUGAACGAAGCAC	1347
Gutheta_nucleo_28S	GUACCAUGAUGCGGAGUUUAAAGCGCAACGCAACGAUCUUGGGGACAUCCGUGUGGACA	1897
	** ***** ** * * * * * * * * * *	
Gmuris_28SrRNA	AUGGUGGGAGUCAUCCCGUCCUACCC--CAGUC---CCCCAUGGCCAGUAGUACUUA	1401
Gutheta_nucleo_28S	CUGGAGAGAGUACUCUUUUC-UAAUUAACAGCACCAUCAACCAUGGA-----AUCA	1947
	*** * **** * * * * * * * * * *	

Gmuris_28SrRNA	GGUCGUGCAUG-----GCCAUGAUGGAUCACAUGUCCCUUCGGUC-----	1441
Gutheta_nucleo_28S	GGUUGUCUGGAGAAGUGGUGUGUCUGGAAGAGCGGUGCCUUCGCGCAUCGUCUGGU *** ** * **** * ** ***** *	2007
Gmuris_28SrRNA	-----AUCCAUGAAACCCGU-----UGUGUCCCCCAUCCUGAUCG	1478
Gutheta_nucleo_28S	AGUUCAUAACGAUCCUUGAAAACCCGAGGGAGAAAUCUGACUCCGC--GCAUGGUCG *** ***** ** ** ***** * ** ** *	2064
Gmuris_28SrRNA	UAC-CACAACCGCAACAGGACUCCAAGGUGAUCAGCCUCUAGGCGGGAGA-GACCAUGAC	1536
Gutheta_nucleo_28S	UACCCGAACCGCAUCA GUCU CAAGGUGAGUAGCCUCGAGUCGGUAGACGAAGGUAAG *** ** ***** ** ***** ***** ** ** ** * *	2124
Gmuris_28SrRNA	UCAGGGAAGUCGGCAACUAGCUCCGUAACCUCGGGAGAAGGAGUGGCUCUGAUCGA---	1593
Gutheta_nucleo_28S	UAAGGGAAGUCGGCAAAUUGGAUCCUAACUUCGGGACAAGGAUUGGCUCUAAGGUGUA * ***** * * ***** ***** ***** ***** * *	2184
Gmuris_28SrRNA	-----	1593
Gutheta_nucleo_28S	GCAUCGUCGCACGUUUCUGCGGUUCUUGAGCGCGUGUUCGAGAGACUCGAUGCUCUCGA	2244
Gmuris_28SrRNA	-----	1593
Gutheta_nucleo_28S	GUCUCUCCUGGCUCAUACCAUGCUUCUUGCAGGAGGCUAUAGAGAGAUUAGGAGUUC	2304
Gmuris_28SrRNA	-----	1593
Gutheta_nucleo_28S	UCGGAUCUCCUCUCCACUGAGAACCAUCUGAACUAUGGUGCACGAUUGUGAAUAAAAC	2364
Gmuris_28SrRNA	-----UCAGAACCGGCACGGACCGAGGAUCCCGACUGUCUACUAACAACAUAGCGUC	1646
Gutheta_nucleo_28S	AACCAGCUUAGAACUGGUACGGACAAGGGAUCCGACUGUUUAAUUAACAUAGCAU * ***** ** ***** **** ***** ** * * ***** *	2424
Gmuris_28SrRNA	GUGCCAGUCG-UCAUGGAUGCGUACACGACGUGAUUUCUGCCCAGUGCCAUGACCGUCAC	1705
Gutheta_nucleo_28S	GCGAUGGCCGGAUUCGUGUUGACGCAUUGUGAAUUCUGCCCAGUGCUCUGAAUGUUA * * * * * * * * * * * * * * * * * * * * * * * * * * * *	2484
Gmuris_28SrRNA	CAUGAAGAGGUUCAUCCAAGCCUGGUAACCGCGGGAGUA CUAUGACUCUCUUAAGGU	1765
Gutheta_nucleo_28S	UGUGAAGAAUUCAGGAAGCGCGGUAAACCGCGGGAGUA CUAUGACUCUCUUAAGGU ***** ** * * *****	2544
Gmuris_28SrRNA	AGCCAAUAGCCUCGUCGGGUAUUUCCGACGUGCAUGAAUGGAUCAACGAGGAUCCACU	1825
Gutheta_nucleo_28S	AGCCAAUAGCCUCGUCUUAUUAGUGACGCGCAUGAAUGGAUGAACGAGAUUCCACU ***** ***** **** ***** ***** *****	2604
Gmuris_28SrRNA	GUCCCAAGUCAUGCCUCCGUGAACCUACAGACCCGGGAACGGGCGGUGUUGGUCGGCGG	1885
Gutheta_nucleo_28S	GUCCUACUUAACCAUAGCGAAACACAGCAAGGGAACGGGCUUGGAAGAUACGCGG ***** * * * * * * * * * * ***** * ** ****	2664
Gmuris_28SrRNA	GCAAGAAGACCCUUUUGAGCU-UGACUCCAUCCGAUCCUGGGGAUGGGAUCAUCGGU	1944
Gutheta_nucleo_28S	GGAAAGAAGACCCUGUUGAGCUUUGACUAGUCUGGUUUUGUGAAUAGUCGAGAGGU ** ***** ***** ***** * * * * * * * * * * * *	2724
Gmuris_28SrRNA	GCAGCAUACAGGGGAGG--CCGCUUCCAUGAGAUACCCUGACGUU-GUUCACCAUCCAC	2001
Gutheta_nucleo_28S	GUAGCAUAAGUGGGAGCUUCGGCAACAGUGAAAUACCACUACUCUUGGC-UUUUCCAC * ***** ***** * * * * * * * * * * * * * * *	2783
Gmuris_28SrRNA	UCACCGACUUA-----	2012
Gutheta_nucleo_28S	UCACCCGUUAGUCGAGGCGGCCUCGUGUUAUCCUAGGCUCUCCUUUAUUGUUUGU ***** ** *	2843
Gmuris_28SrRNA	-----CAUCG-----GCACACGGU	2026
Gutheta_nucleo_28S	GUAUCUGGGAGAACAUCCCCUCGCGGAAGUUUCUCCUACCGUACGAGGACAUAA * * * * * * * *	2903
Gmuris_28SrRNA	CGGAUGGGGAGUUUGGUCGCGCGCGCCUGCUAGAUCACACCGCAGGCGUCCUAUGG	2086
Gutheta_nucleo_28S	CAGGUGGGGAGUUUGGUCGGGCGGCACAUUGCUAAAAGAUAAAGCAGGUGUCCUAAGA * * ***** * * * * * * * * * * *	2963
Gmuris_28SrRNA	UAAGCUCAGUGAGGUCGAAACCUCACGUGGAGCAUAAGGGCAUAAGCUUACUUGACUAG	2146
Gutheta_nucleo_28S	UGAGCUCAGUGAGAACAGCAAUCUUGGAGUAGAAGGGCAAAGCUAUUUGAUUUG * ***** * * * * * * * * * * ***** ***** * * * *	3023

Gmuris_28SrRNA	UACCCCCUGUCCCGGUACUUGCCGUGAAAGCGUGGCCUAACGAUCCUUAACCGCUCCGG	2206
Gutheta_nucleo_28S	GAUUUUCAGUACGAUAUACAAACUGUGAAAGCAUGGCCUAUCGAUCCUUACGACUUC--G	3081
	* * * * *	
Gmuris_28SrRNA	UAAUCCGAGCGUGGAGGUGACAGAAAAGUUACCACAGCGAUAACUGGCCUUGUGGCCGCCA	2266
Gutheta_nucleo_28S	AAGUAUGAAGCUAAGGUGUCAGAAAAGUUACCACAGCGAUAACUGGCCUUGUGGCCGCCG	3141
	* * * * *	
Gmuris_28SrRNA	AGCGCCCAUAGCGACGUGGCUUUUUGAUCCUUCGAUGUCGGCUCUUCCUACCGUCCGCAU	2326
Gutheta_nucleo_28S	AGCGUUCAUAGCGACGCUUUCUUUUGAUCCUUCGAUGUCGGCUCUUCCUUAUUAUGUGAA	3201
	* * * * *	
Gmuris_28SrRNA	GCAUCGGUGCGGAAGCGUCGGAUUGUUCACCCGU-CUAAGGGAUCGUGAGCUGGGUUUAG	2385
Gutheta_nucleo_28S	GCAGAAUUCACAAAGCGUUGUUCACCCACUGACAGGGAAGCUGCGGUUUAG	3261
	* * * * *	
Gmuris_28SrRNA	ACCGUCGUGAGACAGGUUAGUUUAUCCUACCGACCAUCUCC-AUGCGUCCAGUACGUCG	2444
Gutheta_nucleo_28S	ACCGUCGUGAGACAGGUUAGUUUAUCCUACCGACCAUCUCC-AUGCGUCCAGUACGUCG	3321
	* * * * *	
Gmuris_28SrRNA	GGUCAGUACGAGAGGAACACCCGUCGCGAGCCUCCGAUCUCCCGGUUGUUUGACUUGACA	2504
Gutheta_nucleo_28S	GCUCAGUACGAGAGGAACCGUUGAU-----UCAGAUAAU-UGGUCUUGGCACUUGGCU	3373
	* * * * *	
Gmuris_28SrRNA	GU-----GCCGGUCUGUC---GCGCUCGGGGGACUAGCACUGAACGCCUCUAAGUGUC	2554
Gutheta_nucleo_28S	GAGAAGCUAUCGGUGCGAGCUACCAUCUGCGGAUUAUGGCUGAACGCCUCUAAGCCAG	3433
	* * * * *	
Gmuris_28SrRNA	CAUCCACCUACUCGCAUGGGUUGAGGGGCUUCCCUCCGAUUCGCC-----	2600
Gutheta_nucleo_28S	AAGCCUUGCUAGAAACGACGAUAGACAAUC-UCUCCUGGUUUCGAUGUUAUCCACAGG	3492
	* * * * *	
Gmuris_28SrRNA	-----CGGGUUGACGACCCUUUUC--UCCGACCUACUGUACGGC-GUG	2640
Gutheta_nucleo_28S	GCAUAGUGUAAGAGCAAUGGGGUAAACCCUUUUGCUCUUGUGCCGAUUGUACCCGUGUC	3552
	* * * * *	
Gmuris_28SrRNA	GAGC-----UUCUUGCGACCGCCUGAGGUUUGGUUCGGGUUGGCAU--UC--	2683
Gutheta_nucleo_28S	AAUCAGAGUAACGGUGAGCGUGGGAGUGAGUGAUGACUCGUUGUGUAUUCUUACUUCUU	3612
	* * * * *	
Gmuris_28SrRNA	-----CCCCCGAG-----	2691
Gutheta_nucleo_28S	UUCGUUAAGAAUGGAGAGGACACAUCGUGUUUUCAGUAUAUCUACGAGGAUAAACGGG	3672
	* * * * *	
Gmuris_28SrRNA	-----	2691
Gutheta_nucleo_28S	AGGGUACAAAGGGGUUGCUUUUUCUUUUUUUUUUUAUCCCGCAGCCUCCUUCUCCG	3732
Gmuris_28SrRNA	-----	2691
Gutheta_nucleo_28S	UAUAACCCUCCGCGCGGAAUUCUAUCACGCUGUAGGUGGUUGCGCGUGAUGCCGGUGCAA	3792
Gmuris_28SrRNA	-----	2691
Gutheta_nucleo_28S	CUCCUCUUUUUUUGUUUAUAGGUGGAGUUGUCUUUACAUCUCAGUUUUGUUUUUCCUGA	3852
Gmuris_28SrRNA	-----	2691
Gutheta_nucleo_28S	AGAGAAAGAAAAGGAAAGGGCGGGCCUGGAGAGGGAGAUCCCUUGUAGACGACUUGGGAG	3912
Gmuris_28SrRNA	-----	2691
Gutheta_nucleo_28S	CGGGGACAGGUUUUGUAAGGAGUAGAGUAGCCUGGUUUUUUUUGCUACGAUCUCUUGAGA	3972
Gmuris_28SrRNA	-----	2691
Gutheta_nucleo_28S	UUUAGCCUGAGCCACCCAGAUUCUGUGU	3999

**Figure A.1.4. Conserved snoRNA guided modification sites in *G. muris* and the *Gu. theta* nucleomorph.**



**Figure A.1.4. Conserved snoRNA guided modification sites in *G. muris* and the *Gu. theta* nucleomorph.** Large and small subunit rRNAs from *G. muris* and *Gu. theta* Nucleomorph aligned using ClustalOmega (<https://www.ebi.ac.uk/Tools/msa/clustalo/>). Nucleotides predicted to be targeted for 2'-*O*-methylation by snoRNAs are highlighted in yellow (*Gu. theta* nucleomorph) and green (*G. muris*). Green highlighted nucleotides are also targeted in a homologous position in *G. lamblia*, while the nucleotide highlighted in blue, guided by GmsR29, is not conserved in *G. lamblia*. \* indicate identical nucleotides at that position.

GmsR2	--GAGAAAAAGAAATGCGAGAACCAATTGACTCGCAGGAGTGATGCTGACAAGGGATT	58
GtNM-R9	TTGCTATGATGATTGCG-----TCCTCAACTGTTC-GCTTATAAATTC-TT	44
	* * * * *	
GmsR2	TACGTCCGCCGACGAACTGAGTGAGCATGATGCTTCCTTGGATGTCGAAGCTCCCTT	116
GtNM-R9	GAAAAATTCTACTATAAGAATTATTCGTAAGTAATCCTTGGATGCTGAGCTTT----	98
	* * * * *	
GmsR6	---AAATATGATGAGCGTTGTTATCCCTGTCTGATATGTGACGAAAACGACTCCTGACCC	57
GtNM-R10	GTGAAATTTTAATGATGTTGTTATCCCTGTATGAAATTTTGATTGTTCAAATCTGACCTA	60
	* * * * *	
GmsR6	TT- 59	
GtNM-R10	CTT 63	
	*	
GmsR7	---GATGGTGATGACGAGCCTAATCAAC---CTGACTCCTGATGGTGTCATAGTTACTCG	54
GtNM-R8	TCACATGATGACAATAATTAATGAAATTATTAATGACACTTAACCTCATAAGTTACACT	60
	* * * * *	
GmsR7	GAGCCCTCGCCTC 67	
GtNM-R8	GATTAA----- 66	
	**	
GmsR8	ACTGAATGATGACGGTTACCAGCTACCGCTGAGCCAGCGTGATGAGCACAAACCACCTT	60
GtNM-R13	TTTTAATGATTA-----TTATTTGTGCTGAATTTCTTGAAATATTCATCTT	46
	* * * * *	
GmsR8	TCGTCTGAGCCTT 73	
GtNM-R13	TCGAATGAATAAT 59	
	* * * *	
GmsR29-m	TGCAATGACGATAACACCCAGCTCACTCTGAT-CATGATGAGGATGTGGTGTCTGAGC--	57
GtNM-R11	-ACAACTATGATG-TTCTTAGCTCACATTGATTAATGATTAAATCATATAAACTGAGCTT	58
	* * * * *	
GmsR29-m	----- 57	
GtNM-R11	TGTTAGTAT 67	

Candidate13 and GtNM-R7 conserve their ASE guide sequence but are sufficiently divergent to not properly align this region

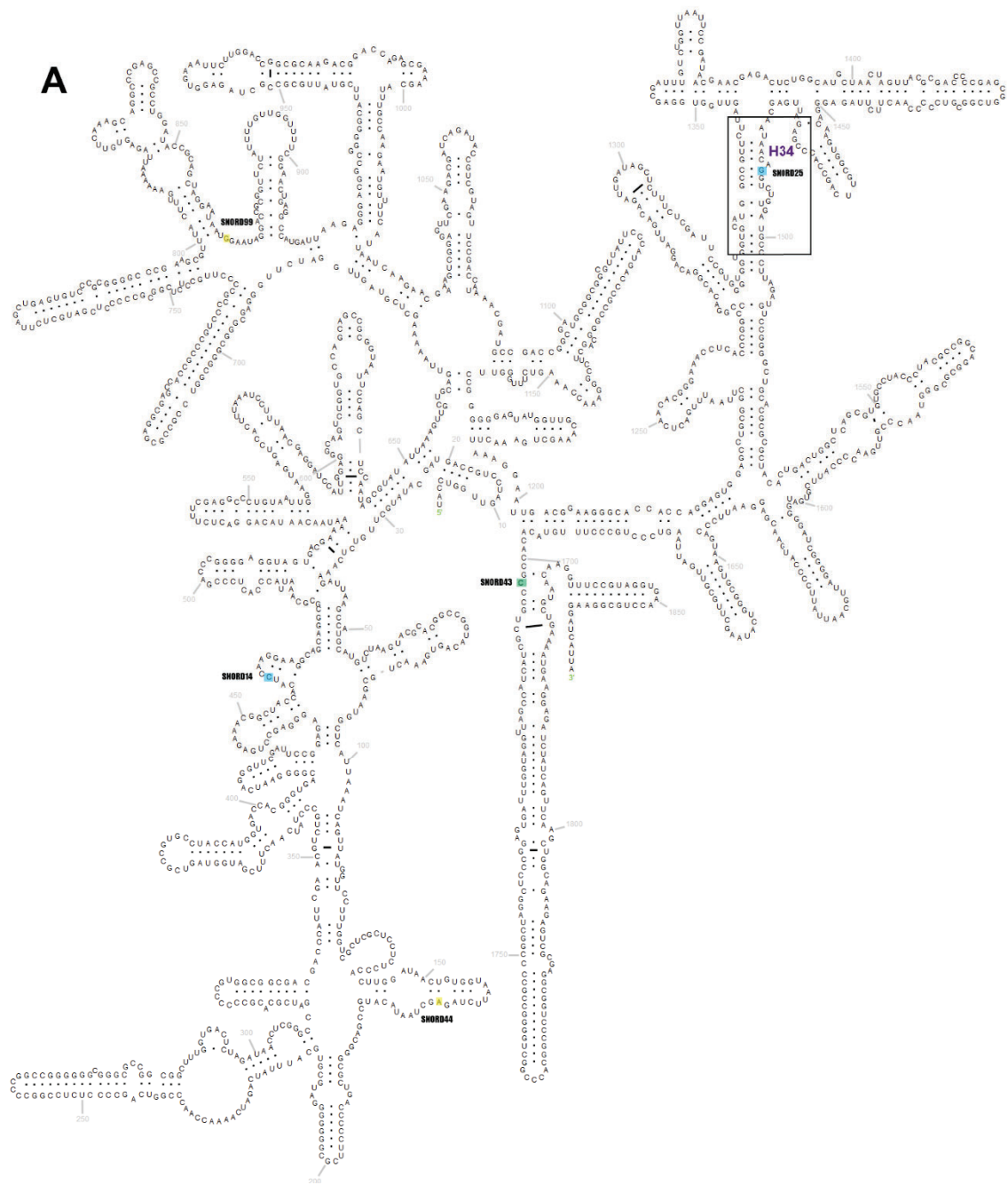
```
GmCandidate13      GTGATGCGATGATTCAACACACGACGGTCTTCTGAGTGAACAACAGCCTGC-TGACCCT      59
GtNM-R7            -----ATATAAATAAATGATGATAGTTTGACGGTCTAATGATATCAATGAAAAT      49
                   ** * ** * * * * * * * * * * * * * * * *

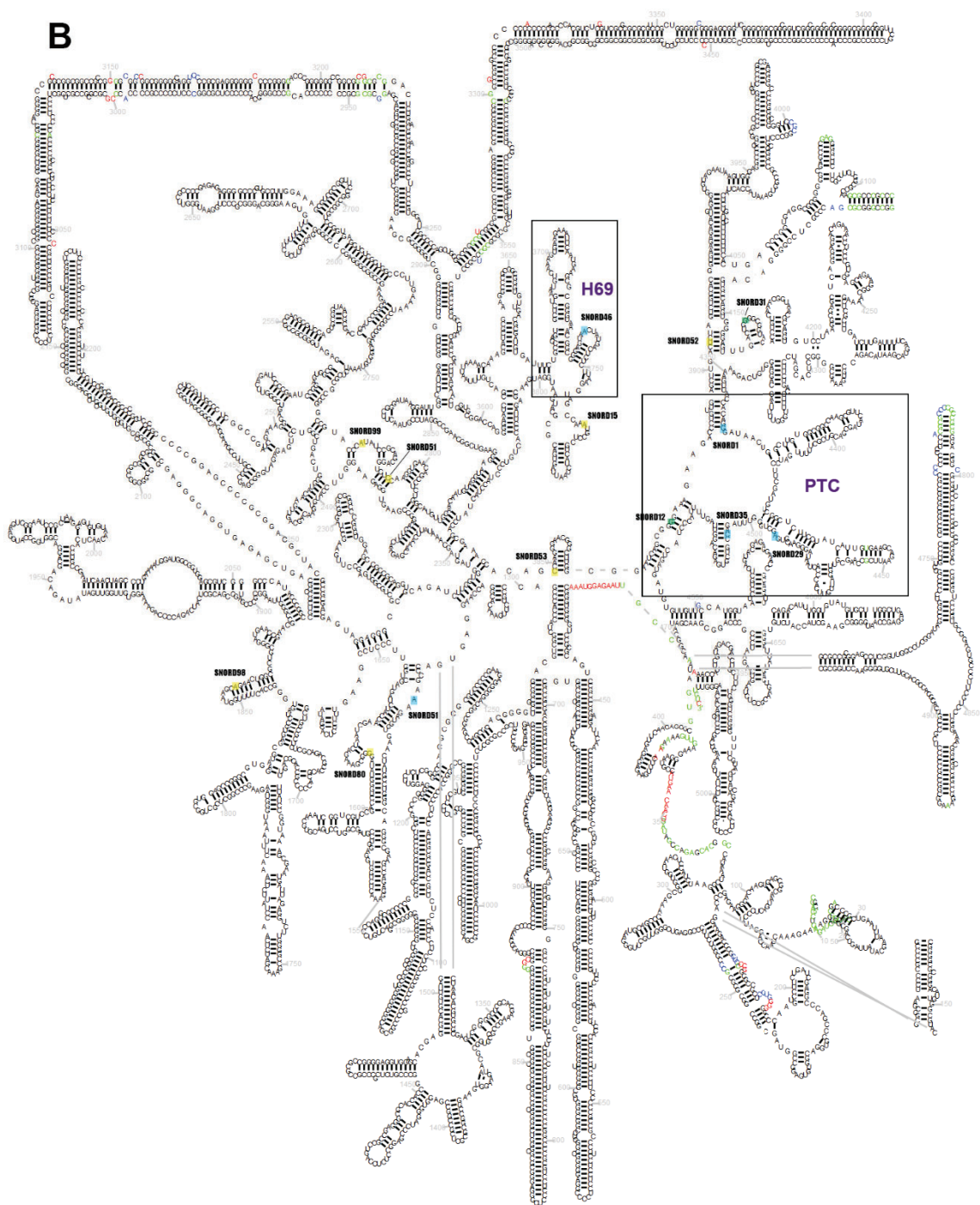
GmCandidate13      GTGTTTGATGAGCCTT      75
GtNM-R7            CAATTCACCTGAACT-      64
                   **                **

GmCandidate1       CGGTCGAGACGTGGAACGTGTCCGCATGAAAAAATGAGGACTCAACGCAGACCTGTGC      60
GtNM-R12           -----AATTAAAATTATGATTAT--AATGACGGTATCTGA      33
                   ** * ** * * ** * * * * * * * * * *

GmCandidate1       TGACCTTTGCCTGATGACGCGTTGTACTGACCTT-      94
GtNM-R12           AGTAATATGATAATAACAGACCTATAATGATTTAT      68
                   *  * * * *      *  * * * * * *
```

**Figure A.1.5. *G. muris* and *Gu. theta* nucleomorph C/D snoRNAs share little conserved sequence outside of their guide elements.** Pairwise alignment of C/D snoRNAs from *G. muris* (Gm) and the *Gu. theta* nucleomorph (GtNM) aligned using ClustalOmega (<https://www.ebi.ac.uk/Tools/msa/clustalo/>). Anti-sense guide regions that base-pair to rRNA sites are underlined for each snoRNA. \* indicates identical nucleotides at that position.





C

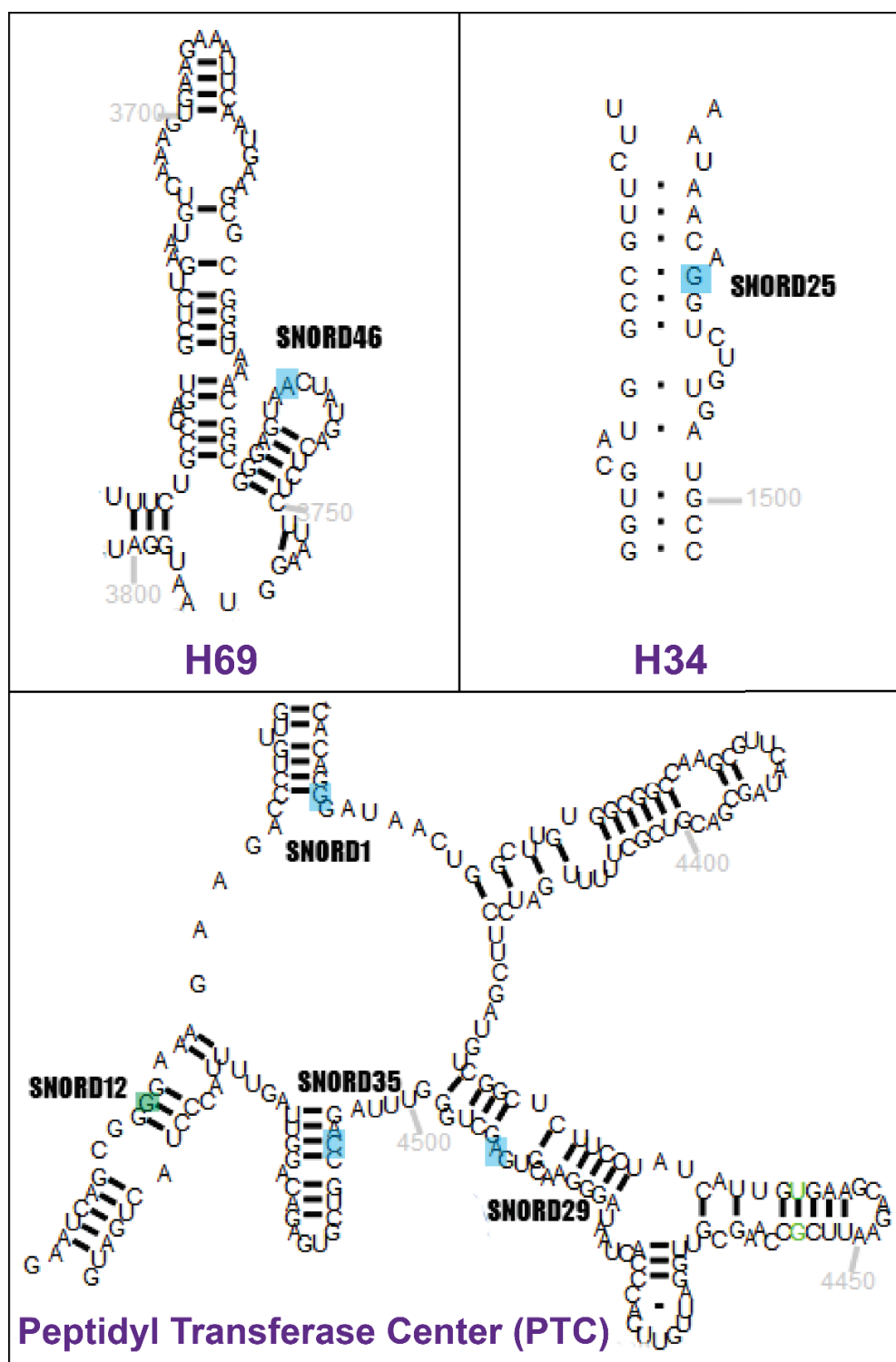


Figure A.1.6. *G. muris* and the *Gu. theta* nucleomorph snoRNAs target homologous positions to human snoRNAs in functionally important regions of the rRNA.

**Figure A.1.6. *G. muris* and the *Gu. theta* nucleomorph snoRNAs target homologous positions to human snoRNAs in functionally important regions of the rRNA.** Secondary structures of human 18S (A) and 28S (B) rRNAs. Sites homologous to 2'-*O*-methylations in *G. muris* and *Gu. theta* nucleomorph were mapped onto the human structure based on sequence alignments with human rRNAs. (C) An enhanced view of three functionally important regions of the rRNA containing conserved modifications in humans, *G. muris*, and *Gu. theta*, and boxed in (A) and (B), rRNA region names labeled in purple. Nucleotides boxed in blue are homologous to the nucleotides targeted by snoRNAs in all 3 species *G. muris*, *Gu. theta* nucleomorph, and human, nucleotides boxed in yellow are targeted by snoRNAs in *Gu. theta* nucleomorphs and humans, and those boxed in green are targeted by snoRNAs in *G. muris* and humans. Modification sites are labeled with the name of the human snoRNA targeting that position. Human rRNA structures were obtained from RNA Central (<https://rnacentral.org/>), green nucleotides indicate those modified compared to the template in RNA Central, red are inserted nucleotides, and blue are repositioned compared to the template in RNA Central.

**A**

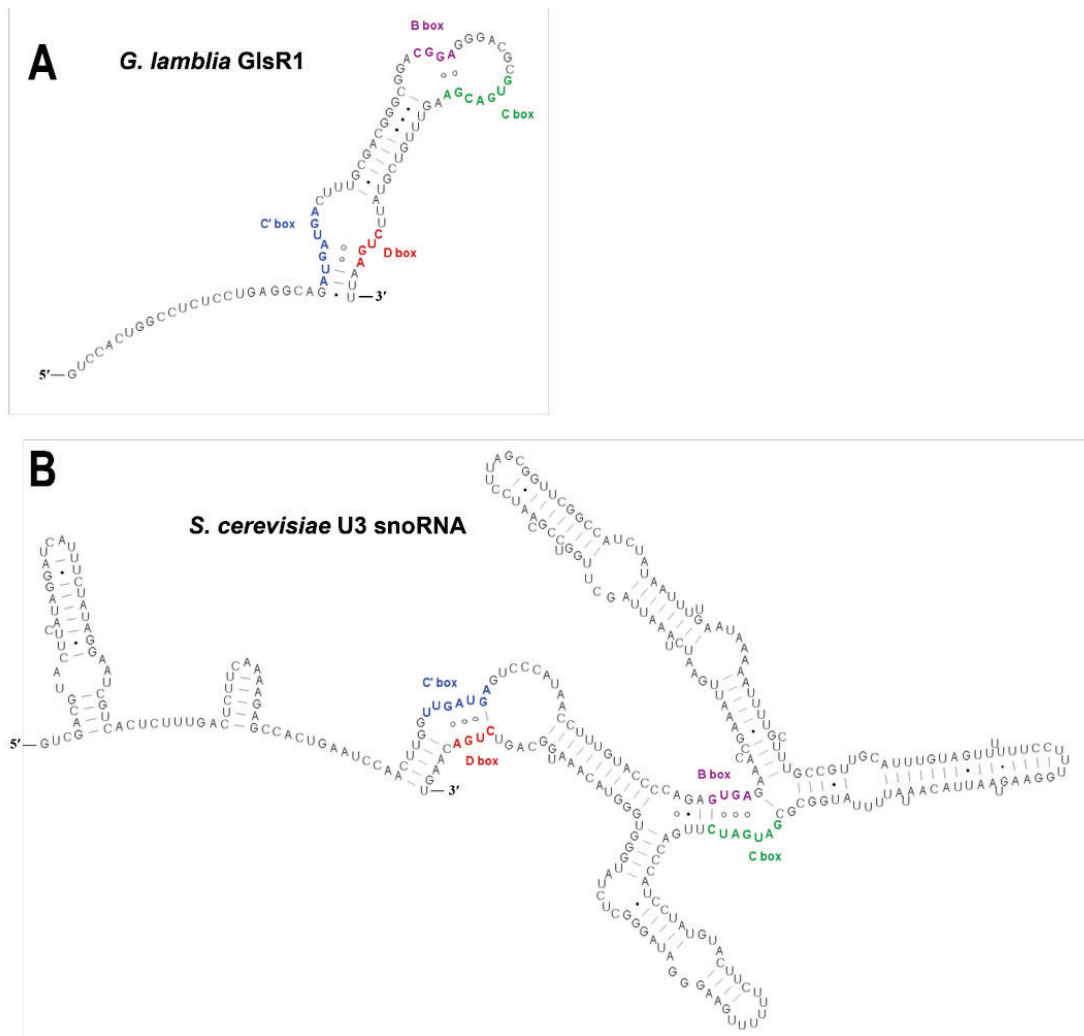
129_C2	TTCAGTTCGATGCGCCCAGGCTGACGGTAGGACGCCT	64
RNasePGlsR15	TTCAGTTCGATGCGCCCAGGCTGACGGTAGGACGCCT	240
	*****	
	↓	
129_C2	GTTCCCTCGCAGAATGATTATCTGTCTCCGAGCAAGCACGACTATGAGCTTACTTATGAGA	124
RNasePGlsR15	GTTCCCTCGCAGAATGATTATCTGTCTCCGAGCAAGCACGACTATGAGCTTACTTATGAGA	300
	*****	
129_C2	TCTGACTCAAAAAAAAAAAAAAAAAA-----	149
RNasePGlsR15	TCTGACTCCTTTACTCAATGTTAGTACTTGTTCCTTGAC	340
	***** * ** *	

**B**

GmsR15	-----	0
GmRNaseP_GmsR15	AGAAAAAGTAAGAGAGACGCTTCAGACTGTGGTCTGGGGAAGGTCCAAGGGCAAGCTCG	60
GmRNaseP	AGAAAAAGTAAGAGAGACGCTTCAGACTGTGGTCTGGGGAAGGTCCAAGGGCAAGCTCG	60
GmsR15	-----	0
GmRNaseP_GmsR15	CGAGGAGAGGCGTGGAAGCGCCAGGGCCAGACCAGAAACGCATGCCCGGGGGATCGCGAG	120
GmRNaseP	CGAGGAGAGGCGTGGAAGCGCCAGGGCCAGACCAGAAACGCATGCCCGGGGGATCGCGAG	120
GmsR15	-----	0
GmRNaseP_GmsR15	ACTCTATGAGTAGCGGCCGAGGCATCTGGAAGCGGCCTGCCGTCGTCCTTCAGTTTCGAT	180
GmRNaseP	ACTCTATGAGTAGCGGCCGAGGCATCTGGAAGCGGCCTGCCGTCGTCCTTCAGTTTCGAT	180
GmsR15	-----ACTCCTTACTTTCTC	15
GmRNaseP_GmsR15	GCGTCTACCCCTTTCCATATGGAAGAGGGACCTCAATTTGGACTACTCCTTACTTTCTC	240
GmRNaseP	GCGTCTACCCCTTTCCATATGGAAGAGGGACCTCAATTTGGACTACTCCTTACTTTCTC	240
	*****	
GmsR15	AAACGATGATGATTCTCTATAACTACGGCTATGTGCACCGTGAAGGGTGTGTGCCGATGG	75
GmRNaseP_GmsR15	AAACGATGATGATTCTCTATAACTACGGCTATGTGCACCGTGAAGGGTGTGTGCCGATGG	300
GmRNaseP	A-----	241
	*	
GmsR15	CTTATGAGATCTGACCCTT	94
GmRNaseP_GmsR15	CTTATGAGATCTGACCCTT	319
GmRNaseP	-----	241

**Figure A.1.7. The mature form of RNase P-GlsR15 is a single transcript and is conserved in *G. muris*.**

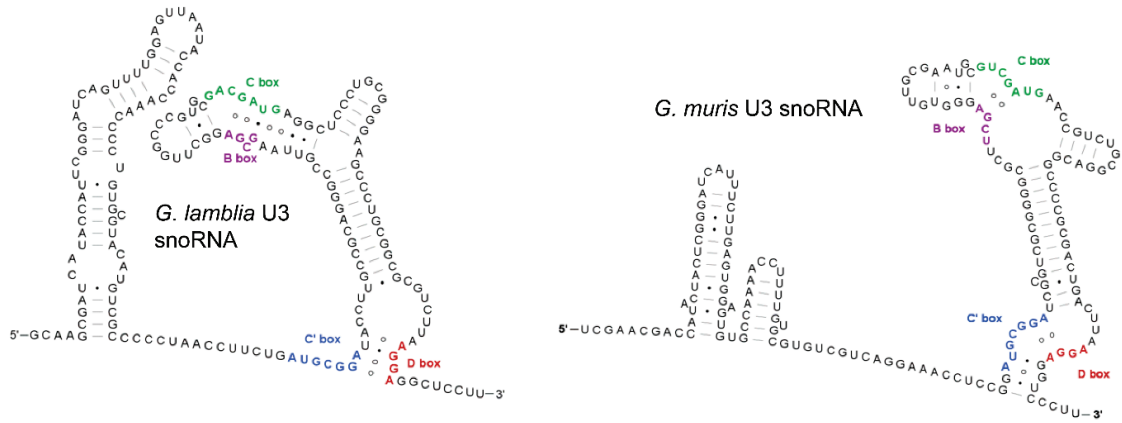
**Figure A.1.7. The mature form of RNase P–GlsR15 is a single transcript and is conserved in *G. muris*.** (A) Sequencing product of the dominant 3' RACE product aligned to the 3' end of the RNase P GlsR15 fusion from *G. lamblia* WB genome. The previously proposed overlap between RNase P and GlsR15 is highlighted in green with the previously predicted 3' end of RNase P indicated by an arrow. Guide sequence targeting tRNA<sup>Met</sup> is underlined and the D box element is highlighted in purple. (B) Alignment of the complete coding sequence for the genomic region containing RNase P and GmsR15 from *G. muris*. \* indicate the region overlapping sequence between the two genes, showing they “overlap” in the same manner as the *G. lamblia* genes.



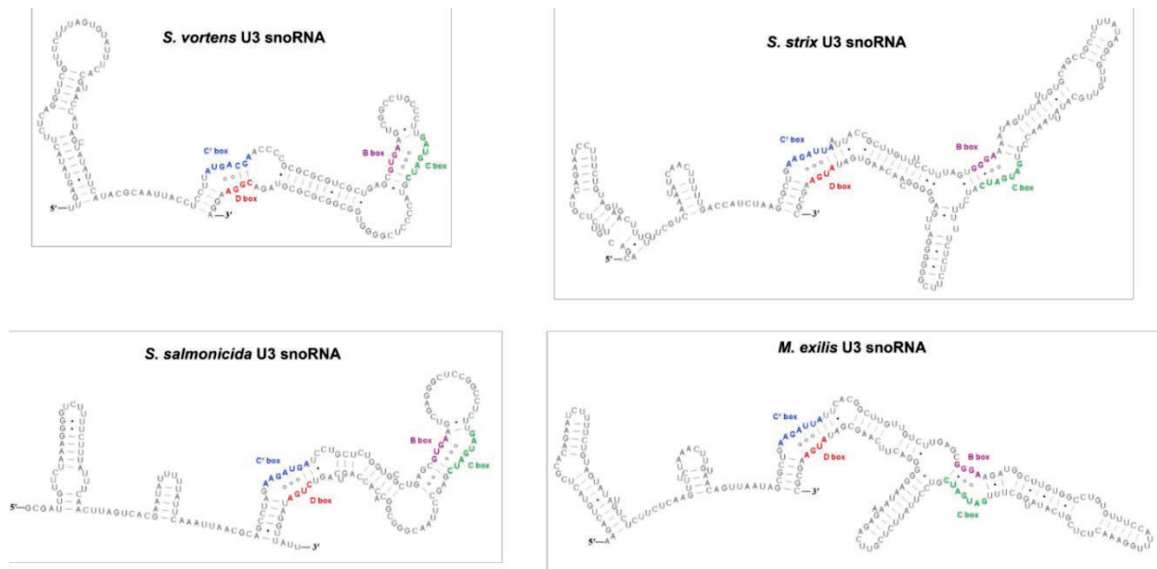
**Figure A.1.8. GlsR1 lacks many key features of U3 snoRNAs.** MFOLD predicted *G. lamblia* GlsR1 snoRNA secondary structure (A) and experimentally determined *S. cerevisiae* U3 secondary structure (B). Box elements are coloured corresponding to the colours found in Figure 2.5A for U3s. GlsR1 lacks most of the conserved features of U3 snoRNAs found in other eukaryotes.



**A**



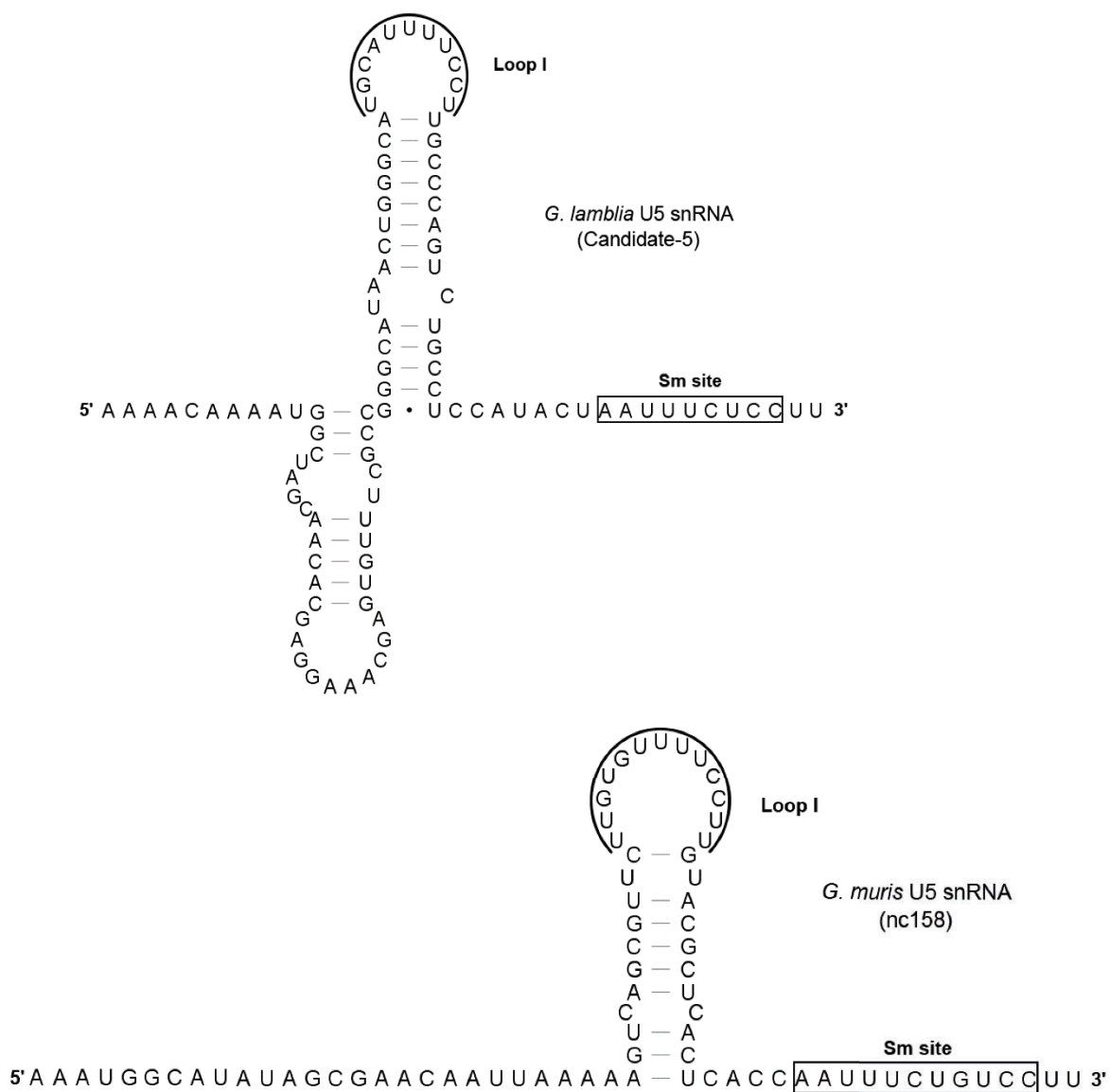
**B**



**Figure A.1.9. Secondary structure predictions for metamonad U3 snoRNAs.** Secondary structures for the *Giardia* U3 snoRNAs (**A**) and other identified metamonad U3 snoRNAs (**B**). Structures are based on MFOLD predictions and manual curation to pair the box elements. Conserved box elements are coloured following the colour scheme in Figure 2.5A.







**Figure A.1.11. Secondary structure predictions for *Giardia* U5 snRNAs.** Comparison of the secondary structures of putative *G. lamblia* and *G. muris* U5 snRNAs based on MFOLD predictions. Sm binding sites are boxed and U rich loop I indicated.

```

Gm TER      GAGAAAACAAAAGCCGCATCTCT-CGCTACCAGCCGGTGTGTCTCCCATACCCTGCAC-
Gl GS TER   -----AAAAAAGTGCACCTTGCCTCAGCC-CTGGTGTGTCTTTATTACCCTACTCT
Gl WB TER   -----AAAAAAGTGCACCTTGTGTTACT--CTGGTGTGTCTTTATTACCCTACTCT
Gl P15 TER  -----AAAAAAGTGCACCTTGCCTTACT--CCGGTGTGTCTTTATTACCCTACTCT
              **** * * * * * * * * * * * * * * * * * * * * * * * *

Gm TER      -----GCCCCACCGACAGTTCTCTCGCAGGCGTCTCTGCCAAGAAGCCTGCAC
Gl GS TER   GTCTAGTGATCTCTCACCACAGTTCTTCTCGCAGTGGTCTCTGTCAAAGACTGCCGCAT
Gl WB TER   GTCTCGTGAACCTCACCACAGTTCTTCTCGCAGTGGTCTCTGTCAAAGACTGCCGCAC
Gl P15 TER  GTCTCGTGAATTCTCACCACAGTTCTTCTCGCAGTGGTCTCTGTCAAAGACTGCCGCAT
              *   * * * * * * * * * * * * * * * * * * * * * * * *

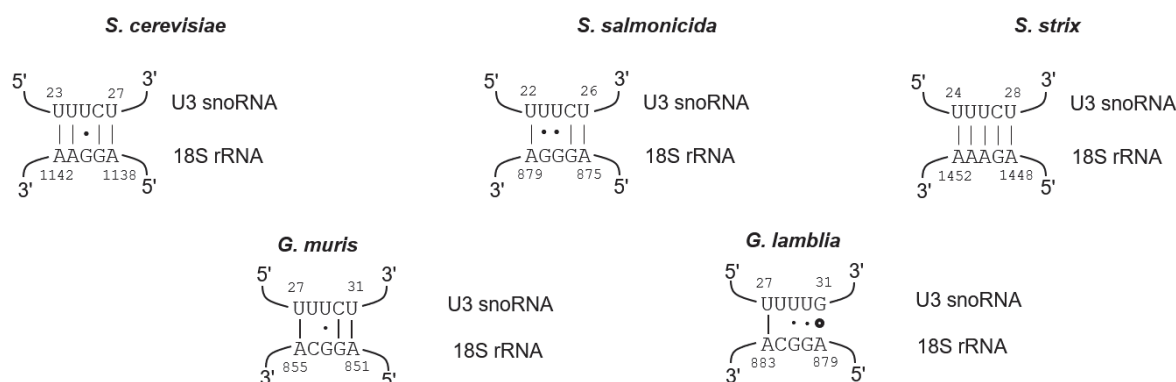
Gm TER      GAAGCACCGGACGTGAGAGAGCAGGAT--CAACCGTGCGGCCCTT-----
Gl GS TER   GGTACACCAGAAGCAAGGGAA-AGGAACCATCCACGAGTCCTT-----
Gl WB TER   GGTACACCAGAAGCAAGGGGA-AGGATCCCATCCACGAGTCCTT-----
Gl P15 TER  GGTACACCAGAAGCAAGGGGA-AGGATCCCATCCACGAGTCCTT-----
              *   **** * * * * * * * * * * * * * * * * * * * * *

```

**Figure A.1.12. Alignment of telomerase RNAs (TER) from *Giardia* species.** Alignment of telomerase RNA candidates from *G. muris* (Gm) and three *G. lamblia* (Gl) isolates: GS, WB, and P15, produced using ClustalOmega (<https://www.ebi.ac.uk/Tools/msa/clustalo/>). \* indicate identical nucleotides at that position for all TERs. The template guide sequence is highlighted in yellow.



**Figure A.1.13. *G. muris* ncRNA 3' end processing motif.** Consensus 3' processing motif derived from the 3' end of all identified *G. muris* ncRNAs and motif containing *trans*-spliced introns reported in Xu et al, 2020, (n = 40 sequences). Produced using WebLogo (Crooks et al., 2004).



**Figure A.1.14 Helix I base-pairing potential between U3 and the 18S rRNA is not conserved in all metamonads**

**Figure A.1.14. Helix I base-pairing potential between U3 and the 18S rRNA is not conserved in all metamonads.** Work preceding the yeast SSU processome cryo-EM structures predicted formation of helix I, which has since been determined not to form. • indicates a GU pair and ◦ indicates a GA pair. Start and end points of pairing interactions for each RNA are indicated. Regions targeted in the rRNA for the metamonads is based off sequence alignments with the *S. cerevisiae* rRNA. No plausible helix I base-pairing could be predicted for either *Giardia* U3, supporting the absence of this helix.

## **Appendix 2 – Supplementary Material for Chapter 3:**

**Analysis of two Snu13p homologs and their associated complexes from the diplomonad *Giardia lamblia***

## A

```

S. salmonicida α-Snu13p  --MDPRATPLATKNLEGQIYNLIETAQKQSSLKVGINEATKQAMRGQAALIIIAANASPL 58
G. lamblia α-Snu13p    MQIDPRAIPFANEELSLELLNLVKHGASLQAIKRGANEALKQVNRGKAELVIIAADADPI 60
G. muris α-Snu13p      -MVDPRATPFAPESLTVALLNATKQATSQAVRRGANEALKQINRGRAALVIAADAEP 59
                        :**** *: * :. * : * : . . .::: * *** ** **: * *:****:*. *:

S. salmonicida α-Snu13p  EVALALPLVCEDKGIPYVFVSEQEGIGRAAQVSRSGAGAVAILNSID----VSAMLHQIE 113
G. lamblia α-Snu13p    EIVLHLPLACEDKGVPYVFIGSKNALGRACNVSVPTIVASI-GKHDALGNVVAEIVGKVE 119
G. muris α-Snu13p      EIVLNIPLVCEDRGVPYVFLGSKEALGRACGVSVPTTVASL-NKHDLLNSTLEELIGQIE 118
                        *: . * :*: .***:*.****:.....:***. ** : .::: . . * : : :*:

S. salmonicida α-Snu13p  ML- 115
G. lamblia α-Snu13p    ALV 122
G. muris α-Snu13p      ALI 121

```

## B

```

S. salmonicida β-Snu13p  MTD-SRITISAKSIQDDVLAIVKQAKTINRIARGINECTKQAAKSRTRLVVIADAVPIE 59
G. lamblia β-Snu13p    MPDARAVPLASEAQSKRIYELVDLAKNSRSISRGMEVTKALNKGKARLVVLSADALPLE 60
G. muris β-Snu13p      -MDPRAIPLANEEKQVSRIYEMVMQAKGIRGSARGVNEVTALNKGARLVVIAADALPLE 59
                        * : :: : . : :* ** . :*:** ** *:****:****:*. *:

S. salmonicida β-Snu13p  IALHLPCLCEDKGIIICVFVNSRVELGRACGLGRPAVACCIKNANKETPLDKKVAEIVILKI 119
G. lamblia β-Snu13p    LVLHLPEVCEDKGIIAYIFVPSRQELGRSVGISRQAVAVAIKAPRQGTALDDKLNIFLTEL 120
G. muris β-Snu13p      LVLHLPDVCEDKGVAYIFVPSRQELGKAAGLSRSCVAIAIKNPRSGTILEDKLNALFME 119
                        :.****:****: :** ** ***: :*. * .** .** .. * *:.*: .: ::

S. salmonicida β-Snu13p  EM 121
G. lamblia β-Snu13p    GH 122
G. muris β-Snu13p      GH 121

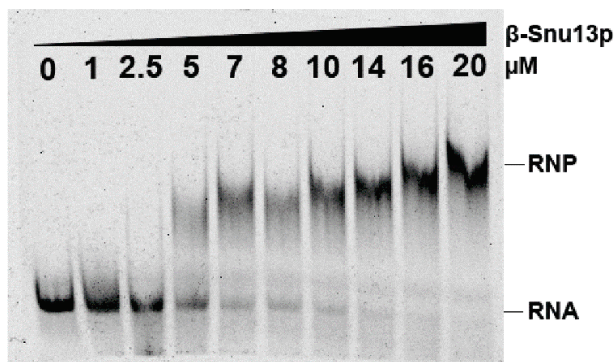
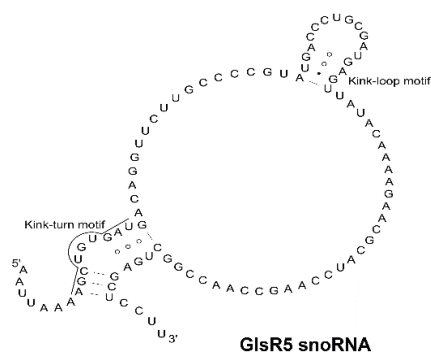
```

**Figure A.2.1. Two Snu13p homologs are present in *G. lamblia*, *G. muris* and *S. salmonicida*.** Multiple sequence alignments of the α-Snu13p (A) and β-Snu13p (B) protein sequences found in *G. lamblia*, *G. muris* and *S. salmonicida*. Residues previously determined to be at key positions involved in RNA binding in other species or that are highly conserved throughout eukaryotes or archaea are coloured. Red indicates the residue matches the general eukaryotic consensus, blue matches the archaeal consensus and green diverges from the conserved identity observed for eukaryotes or archaea. \* represents identical residues, : are residues with very similar biochemical properties and . indicates somewhat similar biochemical properties.

A. fulgidus L7Ae	-----MYVRFEPEDMQNEALILLEKVRRESG <b>K</b> IKKGTNETTKAVERGMAKLVIYA	50
S. solfataricus L7Ae	---MNAMSKASYVKFEVPQDLADKVLAEVRKAKES <b>G</b> IKKGTNETTKAVERGQAKLVIYA	57
A. pernix L7Ae	-----MSKPIYVRFEPEDLAEKAYEAVKRARETGRIKKGTNETTKAVERGLAKLVVIA	54
P. furiosus L7Ae	-----MAKPSYVKFEVPKELAEKALQAVEIARDTG <b>K</b> IRKGTNETTKAVERGQAKLVIYA	54
P. abyssi L7Ae	-----MEGWMMAPSYVKFEVPKELAEKALQAVEIARDTG <b>K</b> IRKGTNETTKAVERGQAKLVIYA	59
S. salmonicida α-Snu13p	-----MDPRATP-LATKNLEGGIYNLIETAQKQ <b>S</b> SLK <b>V</b> GINEATKQAMRGQAALIIIA	52
G. lamblia α-Snu13p	-----MQIDPRATP-FANEELSLELLNLVKGASLQ <b>A</b> IKRGANEALKQVNRGQAKLVIYA	54
G. muris α-Snu13p	-----MVDPRATP-FAPESLTVALLNATKQATSYQ <b>A</b> VRRGANEALKQINRGRAALVLIYA	53
T. thermophila Snu13p	-----MEISERATP-LADDTLSKELTNLVTSCSTQ <b>K</b> VKKGANEATKALNRGLAEIIIIA	54
T. cruzi Snu13p	---MTAEISEKAFP-LAGDRLTQTILDLVQEASNAKMIKKGANEATKALNRGIADLIVLA	56
B. bovis Snu13p	--MTGEDNTSKAFP-LATEEMNSVLLDLVQQACNY <b>K</b> QLKKGANEATKSLNRGLAEIVVLA	57
S. cerevisiae Snu13p	----MSAPNPKAFF-LADAALTQQILDVVQQAANL <b>R</b> QLKKGANEATKTLNRGISEFIIMA	55
C. albicans Snu13p	----MSAPNPKAFF-LADSALTQQILDVVQQSQNL <b>R</b> QLKKGANEATKTLNRGISEFIIMA	55
D. melanogaster Snu13p	---MTEEVNPKAFP-LADAQLTAKIMNLLQQALN <b>Y</b> QLRKGANEATKTLNRGLADIVVLA	56
Z. mays Snu13p	---MESAVNPKAYP-LADAQLTIGIIDIIQQAANY <b>K</b> QLKKGANEATKTLNRGISEFVUMA	56
B. mori Snu13p	MAETEAAVNPKAYP-LADAALTAAILNLVQQATNY <b>K</b> QLRKGANEATKTLNRGLSELIIMA	59
H. sapiens Snu13p	--MTEADVNPKAYP-LADAHLTKKLLDLVQQSCNY <b>K</b> QLRKGANEATKTLNRGISEFIVMA	57
M. musculus Snu13p	--MTEADVNPKAYP-LADAHLTKKLLDLVQQSCNY <b>K</b> QLRKGANEATKTLNRGISEFIVMA	57
X. laevis Snu13p	--MTEFEVNPKAYP-LADAQLTTLTLLDLVQQSANY <b>K</b> QLRKGANEATKTLNRGIAEFIVMA	57
S. salmonicida β-Snu13p	-----MTD-SRIT-ISAKSIQDDVLAIVKQAKTIN <b>R</b> IARGINECTKQAA <b>K</b> SRTRLVIA	52
G. lamblia β-Snu13p	-----MPDARAVP-LASEAQSKRIYELVDLAKNS <b>R</b> SISRGMNEVTALNK <b>G</b> KGARLVLS	53
G. muris β-Snu13p	-----MDPRATP-LANEKQVSRIYEMVMQAKG <b>R</b> GSARGVNEVTALNK <b>G</b> RRARLVIA	52
	* * * * * : . : : : :	
A. fulgidus L7Ae	EDVDPEEIVAHLPILCEEKNVPYIYVKSNDLGRAVG <b>I</b> EVPCASAAIINEGDLRKELGS-	109
S. solfataricus L7Ae	EDVQPEEIVAHLPILCDEKKIPYVYVSSKKALGEAC <b>L</b> QVATASAAILEPGEAKDLVDE-	116
A. pernix L7Ae	EDVDPEEIVMHLPLLCDEKKIPYVYVPSKKRLGEAAG <b>I</b> EVAAASVAIIEPGDAETLVRE-	113
P. furiosus L7Ae	EDVDPEEIVAHLPPLCEEKEIPYIYVPSKKELGAAAG <b>I</b> EVAAASVAIIEPGKARDLVEE-	113
P. abyssi L7Ae	EDVDPEEIVAHLPPLCEEKEIPYIYVPSKKELGAAAG <b>I</b> EVAAASVAIIEPGKARDLVEE-	118
S. salmonicida α-Snu13p	ANASPLEVALALPLVCEDK <b>G</b> IPVYFVSEQEGIGRAA <b>Q</b> V <b>S</b> R <b>S</b> AGAVAILNSIDVSAMLHQ-	111
G. lamblia α-Snu13p	ADADPIEIVLHLPPLCEDK <b>G</b> VPYVFIGSKNALGRAC <b>N</b> V <b>S</b> VPTIVASIGKH-DA--LGNVV	111
G. muris α-Snu13p	ADAEPLEIVLNIPLVCEDR <b>G</b> VPYVFLGSKEALGRAC <b>G</b> V <b>S</b> VPTVVASLNKH-DL--LNSTL	110
T. thermophila Snu13p	ADTTPEIIVLHLPPLCEDK <b>N</b> VPYVVFVSSKKDLGRAC <b>G</b> T <b>S</b> R <b>N</b> VVAIVAVKN-DRSNQTEKI	113
T. cruzi Snu13p	GDTNPTEIILHLPPLCEDK <b>N</b> VPYVVFVPSKTALGRAA <b>Q</b> V <b>S</b> R <b>N</b> AVAILAQ-ENSPVSAKV	115
B. bovis Snu13p	ADAEPLEIILHLPPLVCEDK <b>N</b> IPYIFVKSIALGRAC <b>G</b> V <b>S</b> R <b>P</b> VVSCAIIISR-EGSPLNQI	116
S. cerevisiae Snu13p	ADCEPIEILHLPPLCEDK <b>N</b> VPYVVFVPSRVALGRAC <b>G</b> V <b>S</b> R <b>P</b> VIAASITTN-DASAIKTQI	114
C. albicans Snu13p	ADTEPIEILHLPPLCEDK <b>N</b> VPYVVFVSSKAALGRAC <b>G</b> V <b>S</b> R <b>P</b> VIAASVTSN-DASSIKNQI	114
D. melanogaster Snu13p	GDAEPIEILHLPPLCEDK <b>N</b> VPYVVFVRSKQALGRAC <b>G</b> V <b>S</b> R <b>P</b> IVACSVTTN-EGSQLKQI	115
Z. mays Snu13p	ADTEPIEILHLPPLAEDK <b>N</b> VPYVVFVPSKQALGRAC <b>G</b> V <b>T</b> R <b>P</b> VIACSVTSN-EGSQLKQI	115
B. mori Snu13p	ADAEPLEIVLHLPPLCEDK <b>N</b> VPYVVFVRSKQALGRAC <b>G</b> V <b>S</b> R <b>P</b> IISCSITIN-EGSQLKPQI	118
H. sapiens Snu13p	ADAEPLEIILHLPPLCEDK <b>N</b> VPYVVFVRSKQALGRAC <b>G</b> V <b>S</b> R <b>P</b> VIACSVTIK-EGSQLKQI	116
M. musculus Snu13p	ADAEPLEIILHLPPLCEDK <b>N</b> VPYVVFVRSKQALGRAC <b>G</b> V <b>S</b> R <b>P</b> VIACSVTIK-EGSQLKQI	116
X. laevis Snu13p	ADAEPLEIILHLPPLCEDK <b>N</b> VPYVVFVRSKQALGRAC <b>G</b> V <b>S</b> R <b>P</b> VIACSVTIK-EGSQLKPQI	116
S. salmonicida β-Snu13p	ADAVPIEIALHLPPLCEDK <b>G</b> ICVFNVSRLVELGRAC <b>G</b> L <b>R</b> P <b>A</b> AVACCIAKNANKETPLDKKV	112
G. lamblia β-Snu13p	ADALPLEIVLHLPPEVCEDK <b>G</b> IAYIFVPSRQELGRSV <b>G</b> <b>I</b> S <b>R</b> QAVAVAIKAPRQGTALDDKL	113
G. muris β-Snu13p	ADALPLEIVLHLPDVCEDK <b>G</b> VAYIFVPSRQELGKAAG <b>L</b> S <b>R</b> SCVAIAIKNPRSGTILEDKL	112
	: * * : * * : : : : : : : : : : . : .	
A. fulgidus L7Ae	--LVEKIRGIRK----	119
S. solfataricus L7Ae	--IIKRVNEIKGKTSS	130
A. pernix L7Ae	--IVEKVKELRKAGV	127
P. furiosus L7Ae	--IAMKVKELMK----	123
P. abyssi L7Ae	--IAMKVRELMK----	128
S. salmonicida α-Snu13p	-----IEML-----	115
G. lamblia α-Snu13p	AEIVGKVEALV----	122
G. muris α-Snu13p	EELIGQIEALI----	121
T. thermophila Snu13p	KNIKDKCERLFINA--	127
T. cruzi Snu13p	QAVKLEIERLL----	126
B. bovis Snu13p	VEAKDHIERLLV----	128
S. cerevisiae Snu13p	YAVKDKIETLLI----	126
C. albicans Snu13p	YGKDKIETLLI----	126
D. melanogaster Snu13p	YSIQQEIERLLV----	127
Z. mays Snu13p	QGLKDSIEKLLI----	127
B. mori Snu13p	QTIQQEIERLLV----	130
H. sapiens Snu13p	QSIQQSIERLLV----	128
M. musculus Snu13p	QSIQQSIERLLV----	128
X. laevis Snu13p	QSVQQAIERLLV----	128
S. salmonicida β-Snu13p	AEVILKIEM-----	121
G. lamblia β-Snu13p	NIFLTELGH-----	122
G. muris β-Snu13p	NAFLMELGH-----	121

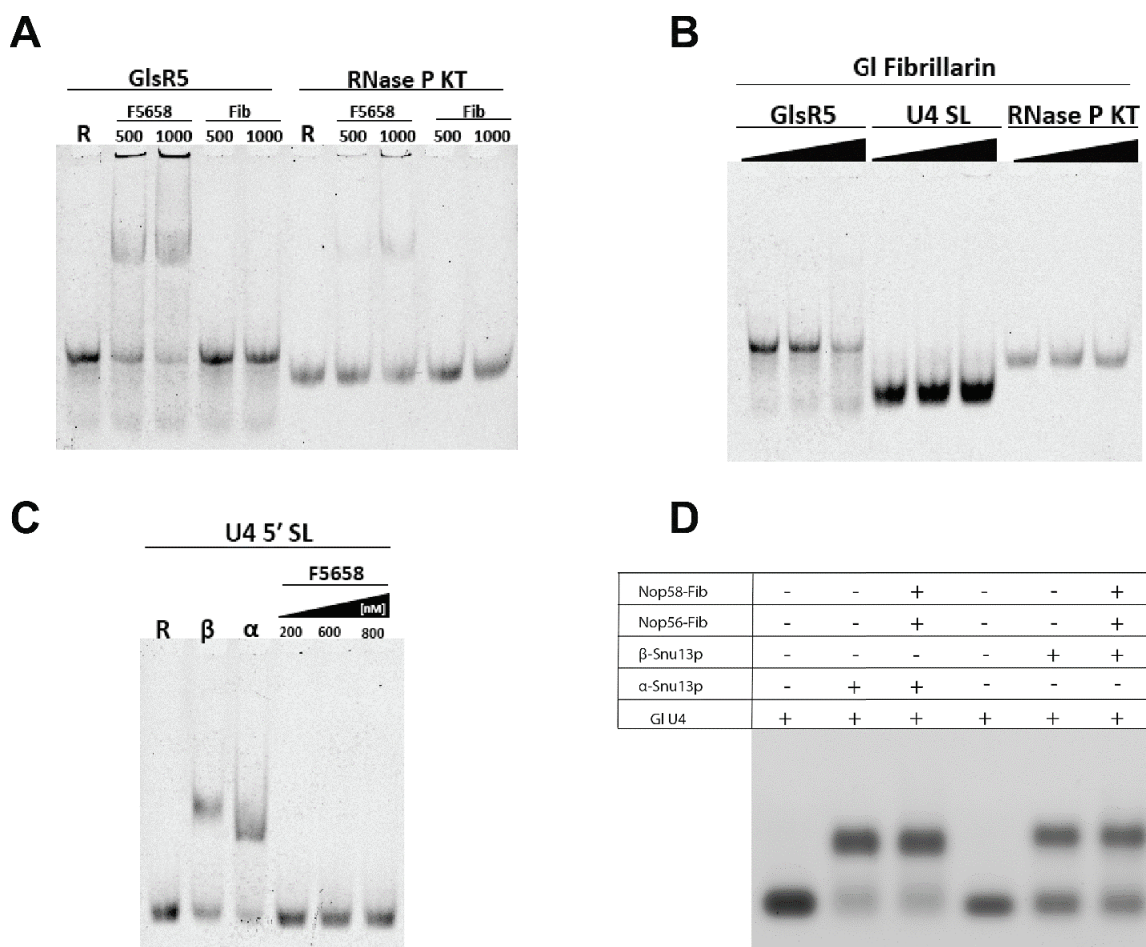
**Figure A.2.2. Diplomonad Snu13p homologs deviate from the general eukaryotic and archaeal consensus sequences at key residues.**

**Figure A.2.2. Diplomonad Snu13p homologs deviate from the general eukaryotic and archaeal consensus sequences at key residues.** Eukaryotic Snu13p and archaeal L7Ae proteins used to generate the phylogenetic tree in Figure 3.2 were aligned using MUSCLE. Residues located at key positions for RNA binding or that are very highly conserved throughout eukaryotes and archaea are coloured. Red indicates the residue matches the general eukaryotic consensus, blue matches the archaeal consensus and green diverges from the conserved identity in eukaryotes or archaea. \* represents identical residues, : are residues with very similar biochemical properties and . indicates somewhat similar biochemical properties.

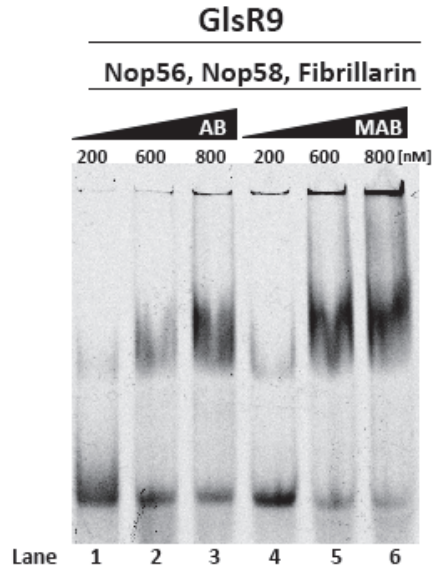


**Figure A.2.3. EMSA analysis of *G. lamblia* β-Snu13p binding to the *G. lamblia* GlsR5 C/D snoRNA.** The predicted secondary structure of GlsR5 (left) is displaced with K-turn and K-loop motifs indicated. EMSA displaying binding of GlsR5 by β-Snu13p is also shown (right) with protein concentrations indicated at the top of each lane in μM. RNP indicates protein-RNA complexes, RNA indicates free RNA.





**Figure A.2.4. Nop56-fibrillarin and Nop58-fibrillarin duplexes from *G. lamblia* specifically bind C/D snoRNAs.** EMSA analysis of Nop-fibrillarin complexes binding the C/D snoRNA GlsR5 or RNase P P12 K-turn (**A**) and GI U4 5' SL (**C**) show combinations of Nop56-fibrillarin and Nop58-fibrillarin specifically recognize C/D snoRNAs. (**A**) and (**B**) show fibrillarin on its own does not bind any of the three RNAs. (**D**) Agarose EMSA showing that larger complexes observed for Nop-fibrillarin complexes (Figure 3.7) are not detected for Nop56-fibrillarin or Nop58-fibrillarin with GI U4. F5658 = both Nop56-fibrillarin and Nop58-fibrillarin are each present at the concentration indicated at the top of the lane. α and β indicate the presence of α-Snu13p and β-Snu13p respectively at 1.5 μM concentration. R indicates an RNA only lane. Concentrations of fibrillarin protein in the lanes of (**B**) for each RNA are 0 nM, 500 nM, 1000 nM. All lanes for A-D contain 100 fmol of the respective labeled RNA transcript.



**Figure A.2.5. Nop56-fibrillarin and Nop58-fibrillarin bind to C/D snoRNAs more tightly than either Snu13p homolog.** Native PAGE EMSA showing the binding of a combined Nop56-fibrillarin and Nop58-fibrillarin complex to the GlcR9 C/D snoRNA. Comparison of binding in AB (lanes 1-3) and MAB (lanes 4-6) buffers. Concentration of Nop-fibrillarin duplexes are indicated at the top of the lane. AB = assembly buffer, MAB = modified assembly buffer (containing 50 mM Arg and 50 mM Glu – See Methods: Chapter 3). Similar results are observed using each of the two duplexes on their own (data not shown).

**A**

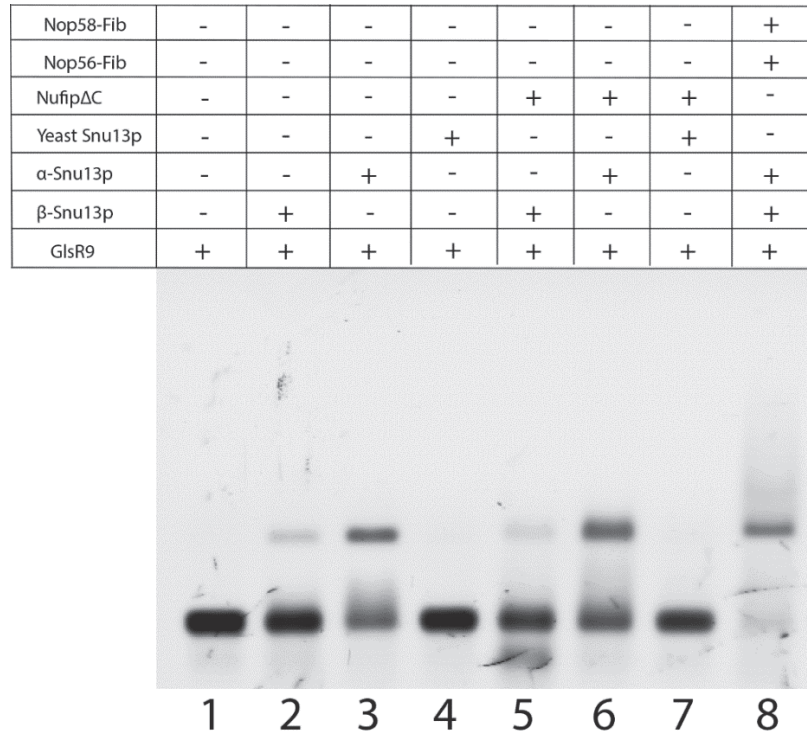
**G1-NUFIP Full length**

MSSELIPATVRRLLTPQQREGGETWTSEKRNYPPTARNKQEREDARRRLRESGAYTSPPETNRFKQGYTQR  
MVERGRRSQVLLVEKPPSLMDRLLYASIHSTDSLVLMLTFIRDSNWLEDE

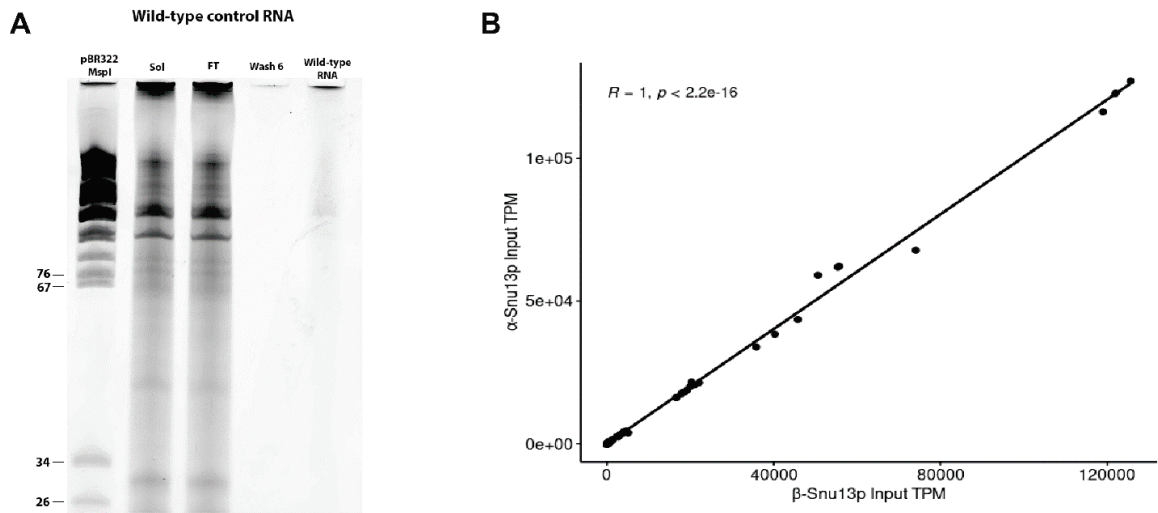
**G1-NUFIPΔC**

MSSELIPATVRRLLTPQQREGGETWTSEKRNYPPTARNKQEREDARRRLRESGAYTSPPETNRFKQGYTQR  
MVERGRRSQVLLVEKPPSLMDRLLYASIHSTDSLVL

**B**



**Figure A.2.6. Truncated NUFIPΔC cannot form a complex with either *G. lamblia* Snu13p homolog.** (A) Sequences of the full length and ΔC truncated candidate NUFIP homolog from *G. lamblia*. Residues highlighted in blue correspond to the residues found to be important for association with Snu13p in humans and yeast. (B) EMSA experiment examining the ability of NUFIPΔC to interact with *G. lamblia* α-Snu13p (lane 6), β-Snu13p (lane 5), and Yeast Snu13p (lane 7). Lane 8 contains all core C/D snoRNP proteins including both Snu13p homologs to examine the affect of both homologs being present during complex formation. + and – indicate presence and absence of the protein respectively. α-Snu13p and β-Snu13p concentrations are 3 μM, Yeast Snu13p is 400 nM, Nop56-fibrillarin and Nop58-fibrillarin duplexes are 600 nM and NUFIPΔC is 1 μM wherever present. Concentrations were selected to be near the  $K_D$  of each protein binding to a C/D snoRNA or other comparable K-turn.



**Figure A.2.7. RNA specifically precipitates with TAP tagged proteins and input libraries are consistent between transfected *G. lamblia* strains. (A)** Urea PAGE gel analysis of a control RNA co-precipitation using wild-type *G. lamblia* WB C6 lysate. Lanes contain: pBR322 MspI ladder, RNA extracted from complete soluble cell extract (Sol), flowthrough from Strep-Tactin® resin purification (FT), wash 6 from purification (Wash 6) and directly extracted from the resin following the final wash (Wild-type RNA). The length of key bands of pBR322 MspI ladder standard are indicated in nucleotides. **(B)** Pearson correlation analysis of the mean RNA abundances in TPM from control input RNA libraries from three replicates for both  $\alpha$ -Snu13p and  $\beta$ -Snu13p RNA co-precipitation experiments. Values are in transcripts per million (TPM) with *R* and *p* values indicated.

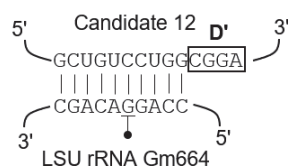
**A**

>Candidate12

**C**
**D'**
**C'**
**D**

CAACCCG**ATGACGA**ATAGCTGTCCTGG**CGGA**GGCGGTC**ATGACGA**CGAAGCC  
 ATCACGT**AGGA**TCCCTT

**B**



**Figure A.2.8. *G. lamblia* ncRNA Candidate 12 is a C/D snoRNA. (A)** Sequence of the Candidate 12 snoRNA 5' to 3' with C, D', C', and D boxes highlighted and labeled. **(B)** Predicted base-pairing between Candidate 12 (top) and the LSU rRNA (bottom). D' element is boxed and predicted target nucleotide in the rRNA is indicated by a line connected to a •.



```

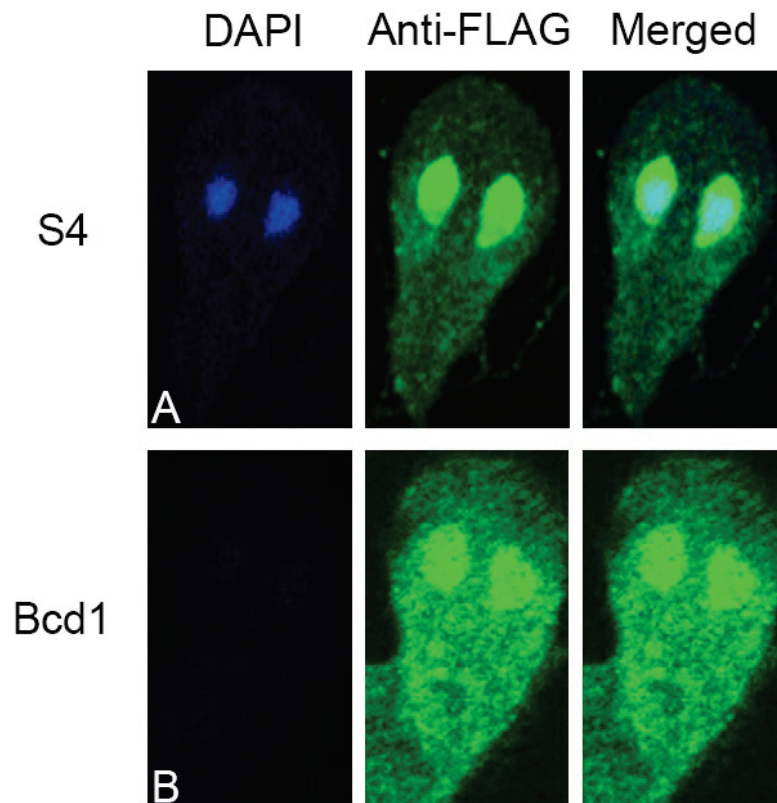
Ucp3      MTQNYRIKQLEERVDFLQSLILQQAGQGPRASVKLANQATKEGVWYVFVKNLPDGLTWRE 60
GMRT_15799 MSENNSIKALEKRISQLEQCIFSKRKGGMHGYMPI-----NQAAILVKNLPVEIDHKQ 54
          *::*  ** *:::.  *::  *::  *  .  .  ::*****  ::

Ucp3      VQDEVFSLYKKSITRVTKIQETEWALRFKSREDAQNCANKYKNAKVNDHVIKCFVKLVQD 120
GMRT_15799 VREQVFACYTKSILYAESLGKGTWRVLFNSESVANQCLAKYRDARVNGIRIHCQLEGANG 114
          *:::**: *.***  .  .:  *  : *::..  *::*  **::*::*.  *:*  ::  ...

Ucp3      EKELKVGSGSLSTTFKKRRRIKFNNRNKEKSGSSETVKKY 159
GMRT_15799 NRNPRNTRSGD---KRPRSRPN-----TATAE----- 139
          :::  :  .  .  ** *  : *  :::*

```

**Figure A.2.9. Ucp protein sequences differ significantly between *G. lamblia* and *G. muris*.** Pairwise sequence alignments of *G. lamblia* proteins Ucp1-3 with putative homologs from *G. muris*, where \* represents identical residues, : are residues with very similar biochemical properties and . indicates somewhat similar biochemical properties. *G. muris* homologs are named as ID numbers from giardiadb (<https://giardiadb.org/>).



**Figure A.2.10. Localization of ribosomal protein S4 and putative Bcd1 candidate in *G. lamblia*.** *G. lamblia* expression of TAP tagged S4 ribosomal protein (A) and putative C/D assembly factor protein Bcd1 (B) both show significant localization to both *G. lamblia* nuclei. Bcd1 also localizes throughout the cytoplasm. DNA is stained with DAPI and appears blue. The protein of interest is visualized via an Alexa Fluor™ 488 fluorophore conjugated antibody and appears in green. Cells were visualized using an Olympus Fluoview FV1200 confocal microscope using the FV10-ASW 4.2 software at 60X magnification.



**Table A.2.1. Categorization of proteins from a single replicate  $\beta$ -Snu13p protein co-precipitation experiment incubated with resin for 24 hours prior to elution.**

Complex	$\beta$ -Snu13p (24h)
Ribosome	66
C/D snoRNP	5
SSU Processome	45
Ribosome Biogenesis	25
Other	145
RNase P	1
Ucp	3
H/ACA snoRNP	3
<b>Total</b>	<b>293</b>

## A

H_sapiens_Snu13p	MTEADVNPkAYPLADAHlTKKLLDLVQQSCNYKQLRKGANeATKTLNRGISEFIVMAADA	60
Z_mays_Snu13p	-MESAVNPkAYPLADAQlTIGIIDI IQQAANYKQLKKGANeATKTLNRGISEFVMAADT	59
	*: *****:* :*:*:*.*****.*****.*****.*****:	
H_sapiens_Snu13p	EPLEIILHLPLlCEDKNVPYVFVRSKQALGRACGVSrpVIACSVTIKEGSQlKQQIQSIQ	120
Z_mays_Snu13p	EPLEIILHLPLlAEDKNVPYVFVPSKQALGRACGVTpVIACSVTSNEGSQlKQQIQGLK	119
	*****.*****.***** *****.*****.***** :*****.:::	
H_sapiens_Snu13p	QSIERLLV	128
Z_mays_Snu13p	DSIEKLLI	127
	:***:***:	

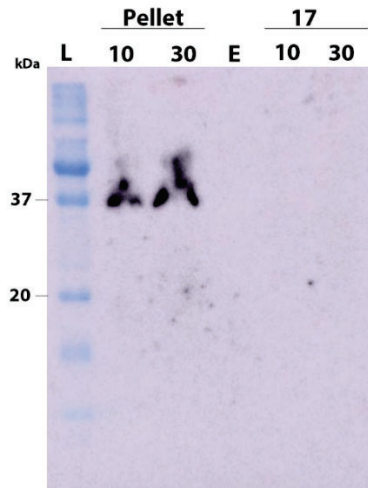
## B

G_lamblia_α_Snu13p	MQIDPRAIPFANEELSLEllNLVKHGASlQAikRGANEALKQVNRGKAELVlIAADADPI	60
S_salmonicida_α_Snu13p	--MDPRATPlATKNlEGQlYnlIETAQKQSSlKVGINeATKQAMRGQAAlIIIAANASPl	58
	:**** *:*.::*. :*:*: . . .: * * * * * . **:* *:****:*.:	
G_lamblia_α_Snu13p	EIVLHLPlACEDKGVPYVFVIGSKNALGRACNVSVPTIVASIGKHDA LGNVVAEIVGKVEA	120
S_salmonicida_α_Snu13p	EVALALPlVCEDKGIPYVFVSEQEGIGRAAQVSRsAGAVAILNSIDVS---AMlHQIEM	114
	*:.* ***.*****:****:.....:****:* * :.* * : . : : : *	
G_lamblia_α_Snu13p	LV	122
S_salmonicida_α_Snu13p	L-	115
	*	

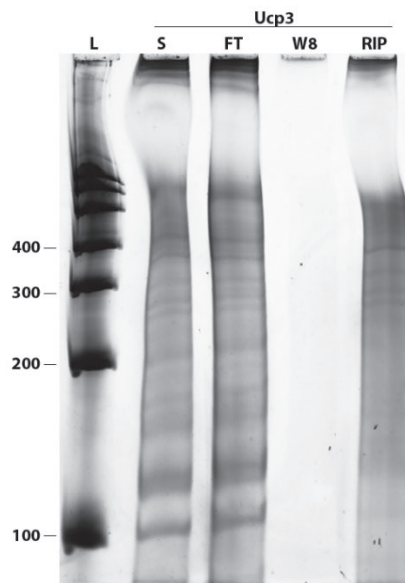
## C

H_sapiens_Snu13p	MTEADVNPkAYPLADAHlTKKLLDLVQQSCNYKQLRKGANeATKTLNRGISEFIVMAADA	60
G_lamblia_α_Snu13p	--MQIDPRAIPFANEELSLEllNLVKHGASlQAikRGANEALKQVNRGKAELVlIAADA	57
	::*: * *: * :. :*: :*: :... : :*:***** * :*** :*: :*: :*	
H_sapiens_Snu13p	EPLEIILHLPLlCEDKNVPYVFVRSKQALGRACGVSrpVIACSVTIKEGSQlKQQIQSIQ	120
G_lamblia_α_Snu13p	DPIEIVLHLPlACEDKGVPYVFVIGSKNALGRACNVSVPTIVASIGKHDA--LGNVVAEIV	115
	:*:***:***** *****.*****: **:******.* * *.***: :.. * : : *	
H_sapiens_Snu13p	QSIERLLV	128
G_lamblia_α_Snu13p	GKVEALV-	122
	.:* *:	

**Figure A.2.11. Sequence divergence between Snu13p homologs in different diplomonads and other eukaryotes.** Pairwise sequence alignments of eukaryotic Snu13p homologs generated using ClustalOmega comparing *H. sapien* and *Z. mays* (A), *G. lamblia*  $\alpha$ -Snu13p and *S. salmonicida*  $\alpha$ -Snu13p (B), and *H. sapien* Snu13p and *G. lamblia*  $\alpha$ -Snu13p (C). \* represents identical residues, : are residues with very similar biochemical properties and . indicates somewhat similar biochemical properties.



**Figure A.2.12. Tagged *G. lamblia* S4 rprotein sediments in glycerol gradients.** Western blot analysis of cell lysate from *G. lamblia* cells expressing TAP tagged S4 rprotein resolved on a 10-30% glycerol gradient for 20h at 26,000 rpm. Pellet lanes contain 10 and 30  $\mu$ L of resuspended protein pellet that ran through the gradient. '17' lanes contain 10 and 30  $\mu$ L of fraction 17, one of the lightest fractions, expected to contain free protein not found in a complex. E = empty lane and L is a pre-stained molecular weight protein reference standard. Western blot of all gradient fractions did not detect S4 in lighter fractions (data not shown).



**Figure A.2.13. Ucp3 co-precipitation enriches distinct RNAs.** RNAs associated with Ucp3 were co-precipitated in the same manner as the Snu13p homologs. RNAs extracted from the complete soluble cell lysate (S), post purification flowthrough (FT), wash 8 of the purification (W8), and precipitated protein bound to resin following all washes (RIP) were resolved on a Urea PAGE gel to examine for enriched RNAs. L = RiboRuler low range ladder.



### **Appendix 3 – Supplementary Material for Chapter 4:**

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

a)

### 5' linker & 3' tail

5' *gcugauggcggaugaagaacacugcguuugcuggcuuugaugaaa*CCAACACCCCGCCCAGACCAGGCU  
GGAUCUGCGGCGAGUGUAAGUGUUCAGAGGUUAUG*ggg*<sub>(N)</sub> 3'

### cDNA

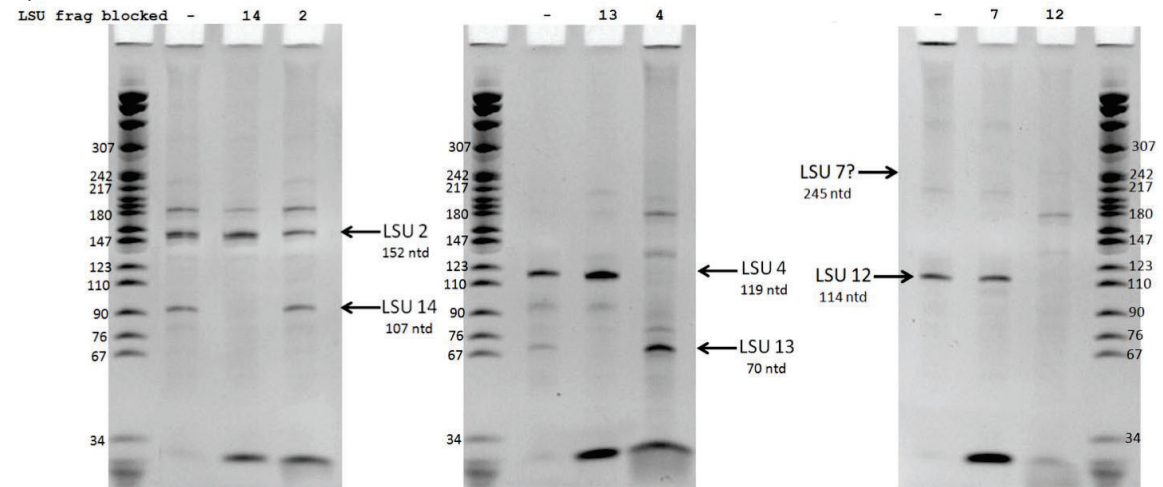
5' *ctcccgtttccagatctcgagc*<sub>(15)</sub>*g/a/t*CATAACCTCTGAACACTTACACTCGCCGCAGATCCAGC  
CTGGTCTGGGCGGGGTGTTGG*tttcatcaaagccagcaaacgcagtggttcattcatcgccatcagc* 3'

### Oligonucleotide design

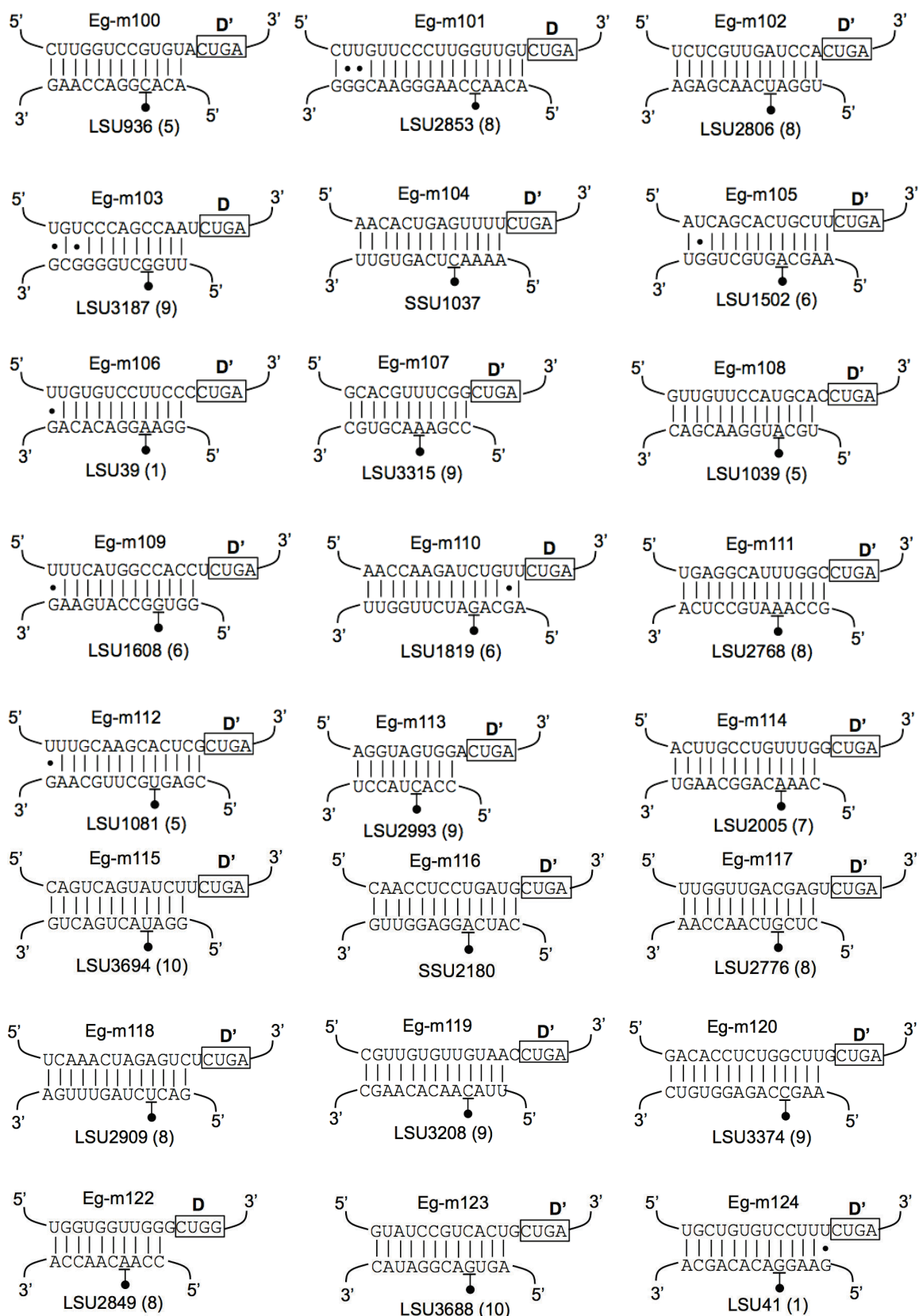
5' *ctcccgtttccagatctcgagc*<sub>(15)</sub>*g/a/t*CATAACCTCTGAACACTTACACTCGCCGCAGATCCAGC  
CTGGTCTGGGCGGGGTGTTGG*tttcatcaaagccagcaaacgcagtggttcattcatcgccatcagc* 3'

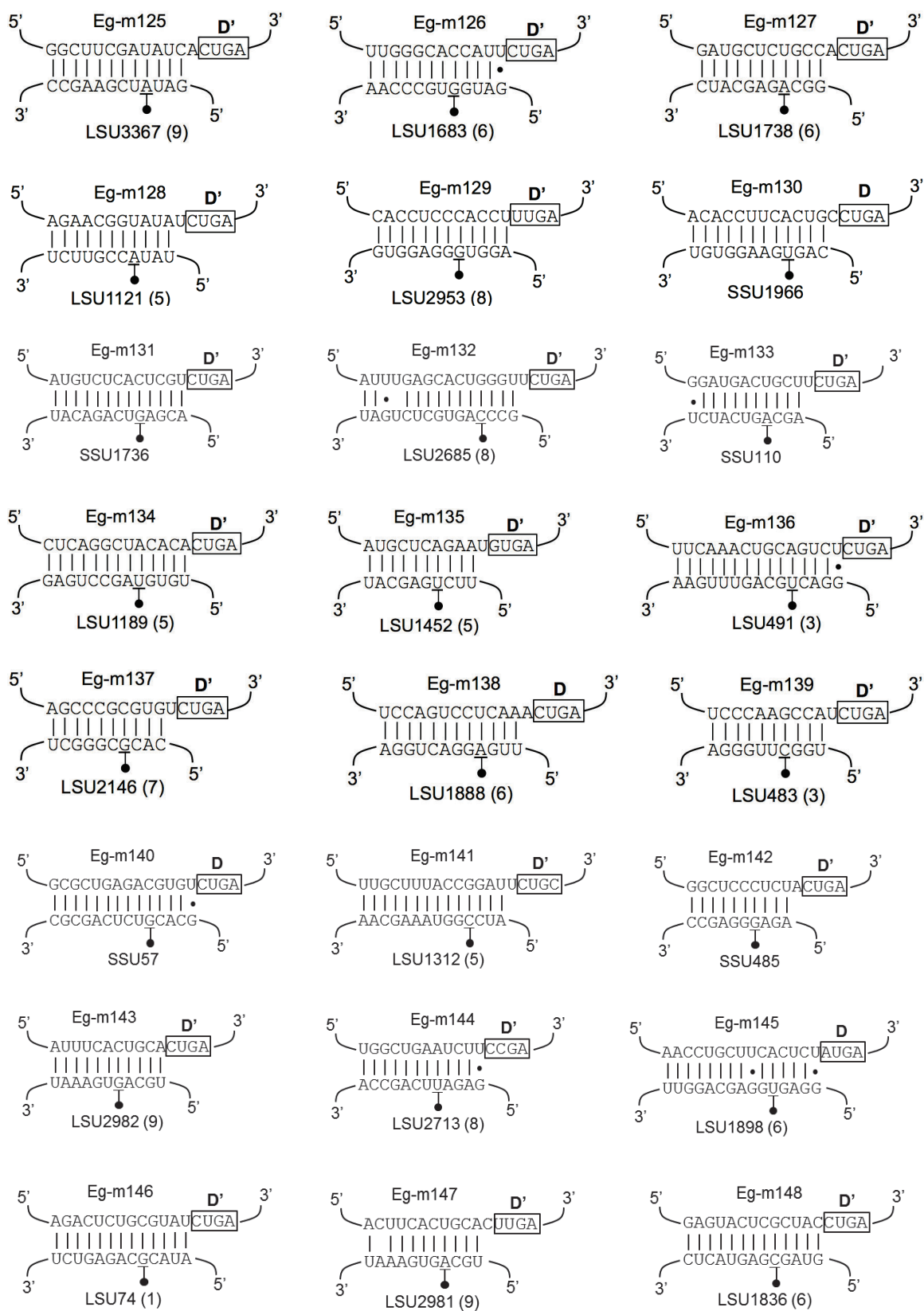
X

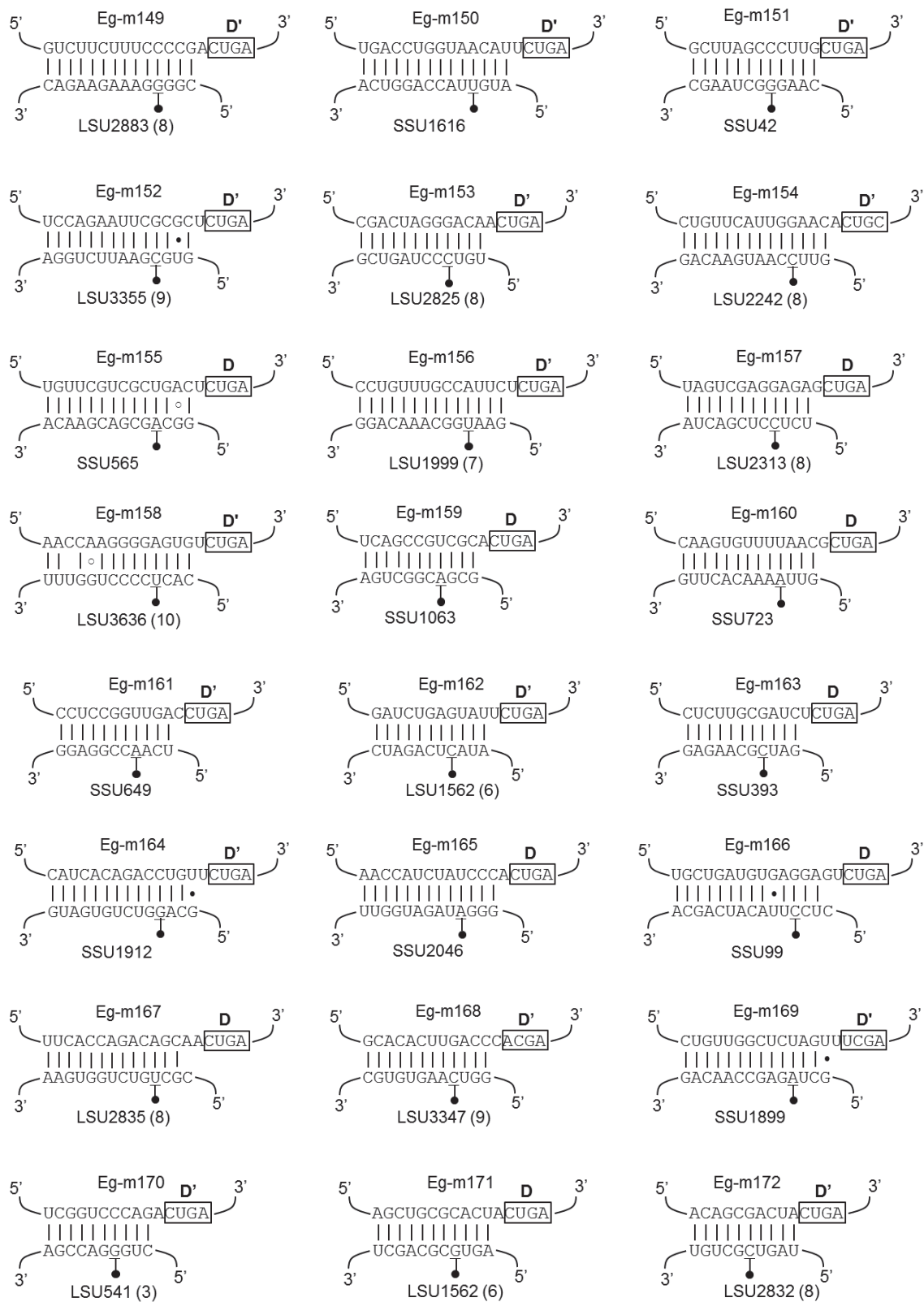
b)

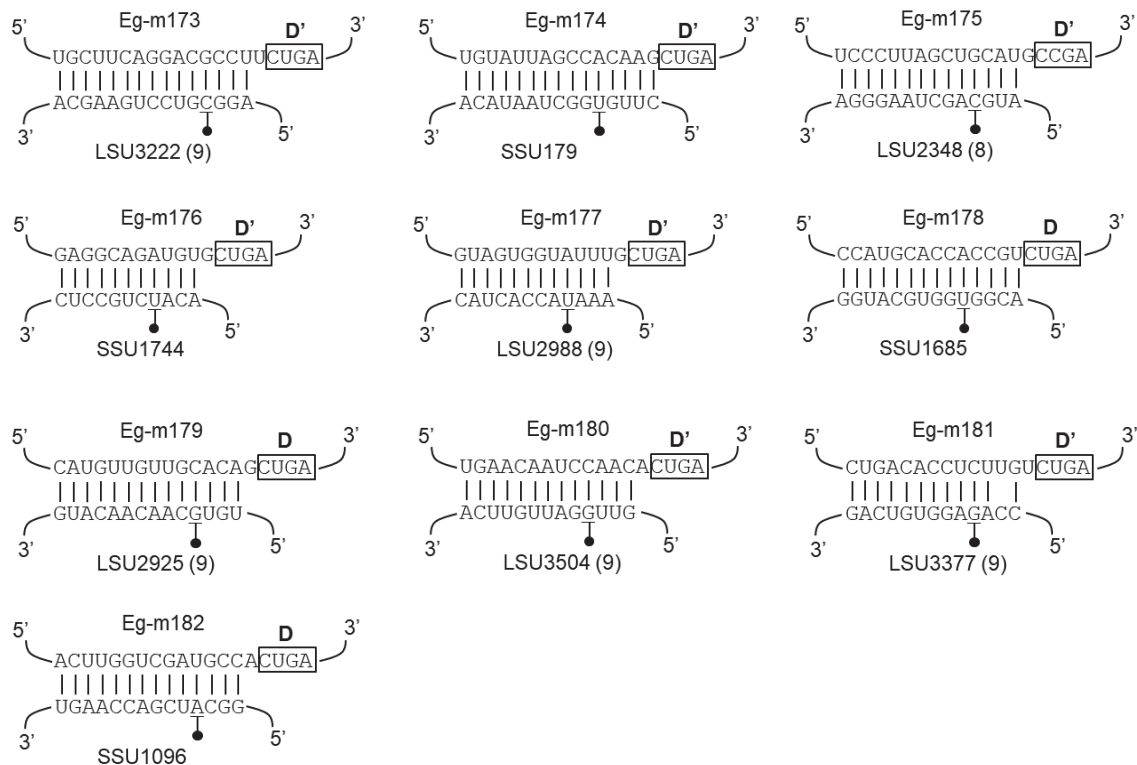


**Figure A.3.1. Illustration of the oligo blocker strategy containing sequences and experimental results.** a) After the 5' linker and 3' poly G tail are added to the RNA, it is reverse transcribed into cDNA. 'Universal' oligonucleotides that anneal to the linker and tail are used to amplify the cDNA products. A blocker oligonucleotide primer is designed to prevent extension of one of the universal oligos by annealing to the junction of the linker region of the cDNA and the rRNA species being targeted. Lower case letters represent nucleotides that were added (blue = 5' linker, red = 3' tail), upper case letters represent rRNA LSU fragment 13, in this example. Arrows indicate 'universal' oligonucleotides and the X-arrow represents the blocker oligonucleotide. b) Additional examples of primer blocking to reduce rRNA amplification during small RNA library preparation.

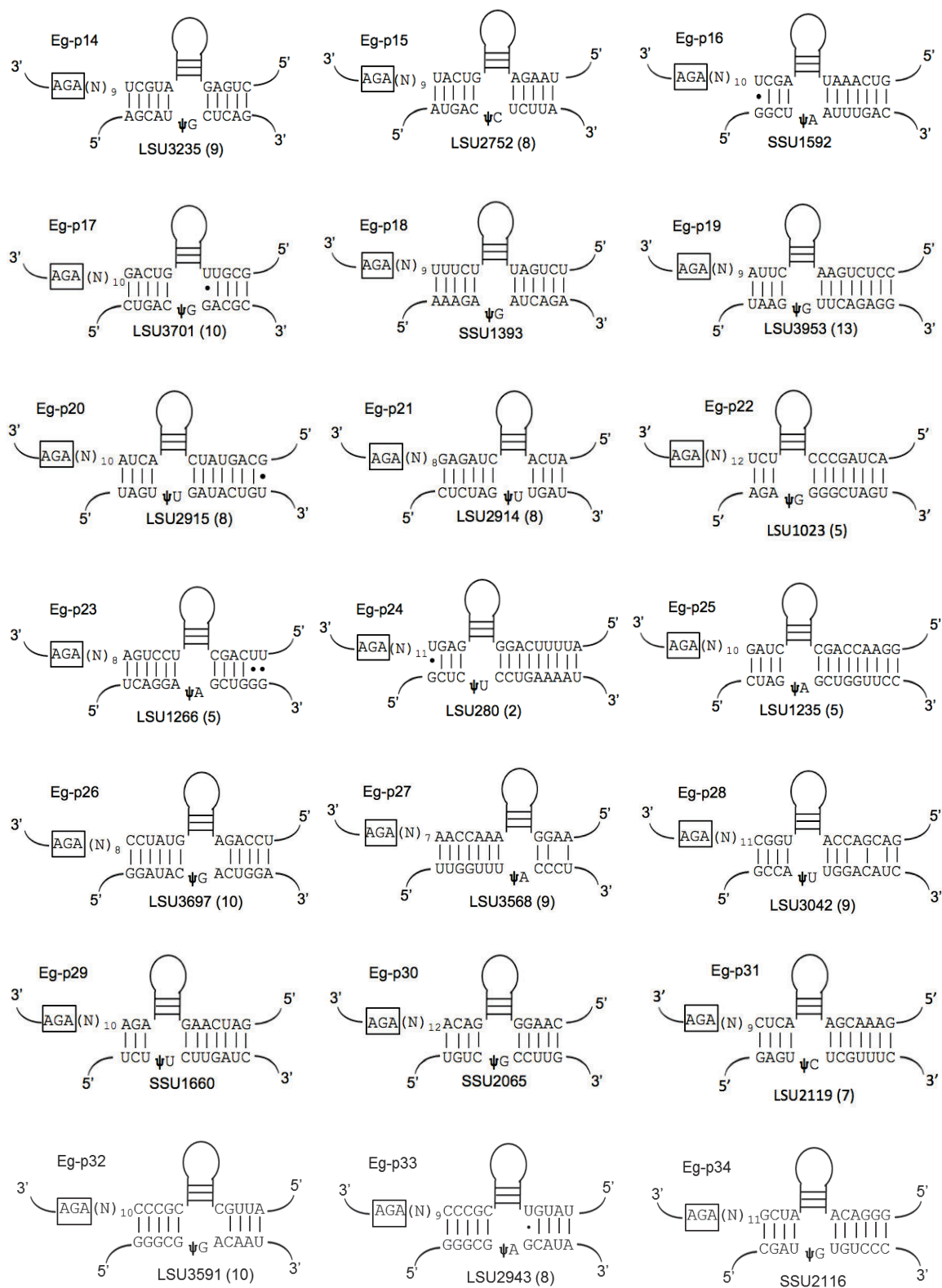




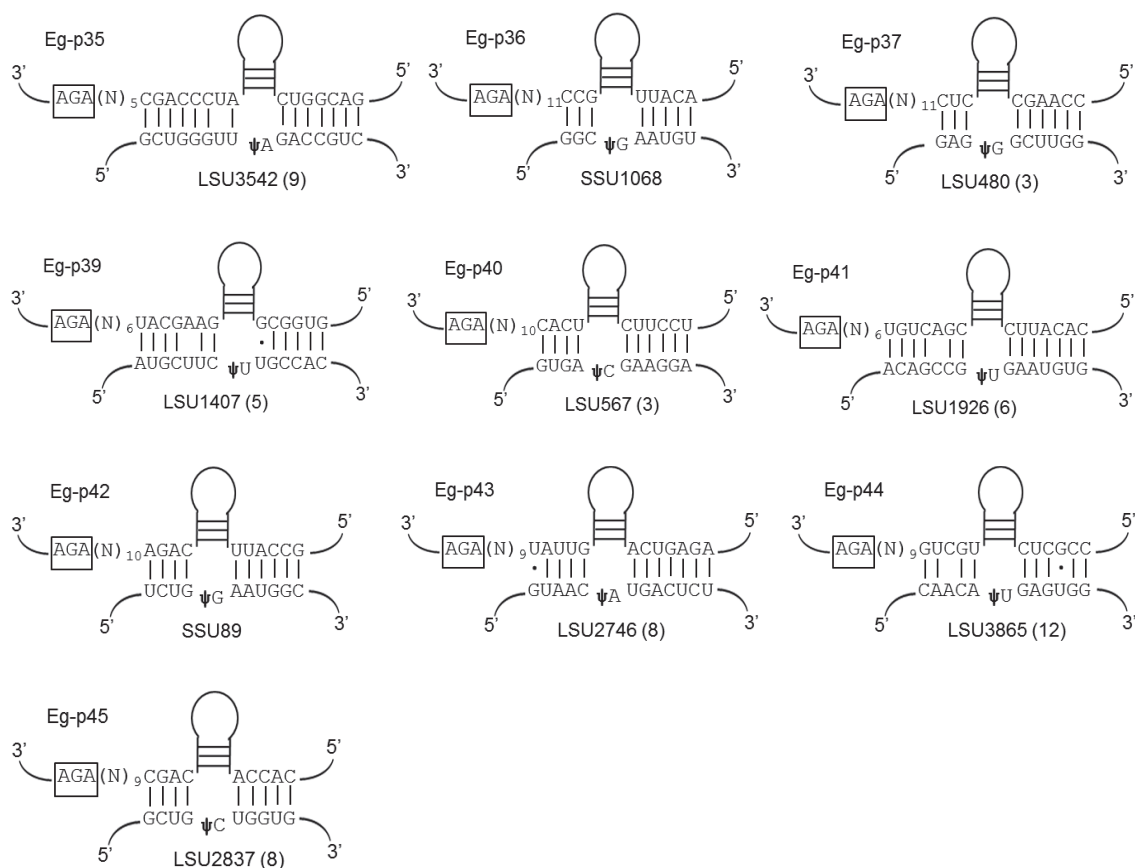




**Figure A.3.2. Identified *E. gracilis* box C/D snoRNAs and their predicted target sites in the rRNA for 2'-O-methylation events.** The predicted base-pairing interaction between the snoRNA (top strand) and the target region in the rRNA (bottom strand) is depicted. Experimentally confirmed methylation 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 A.3.4. Eg-m121 is shown in Figure 4.3.







**Figure A.3.3. Identified box AGAUGN 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” (Schnare and Gray, 2011). The AGA box element is highlighted and the number of nucleotides (N) to the base-paired region is indicated. LSU = large subunit rRNA, SSU = small subunit rRNA and the *E. gracilis* LSU “fragment” species where the modification site resides is indicated in parentheses. Full-length snoRNA sequences are shown in Figure A.3.4. Eg-p38 shown in Figure 4.3.



## Box C/D snoRNAs

The 2'-*O*-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. For Eg-m100 a 3' RACE experiment identified additional isoforms.

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

**Eg-m101 [LSU2853 (8)]**  
 TGGATGATGACTCCTTGGATGACGTGCCGTGATCTGATTCTTGTTCCCTTGTTGTCTGAC

**Eg-m102 [LSU2806 (8)]**  
 CAGATGATGTTTTGTTCTCGTTGATCCACTGACGCCCCGCTGAGGCTACCATTTTCCTGACC

**Eg-m103 [LSU3187 (9)]**  
 TTGATGATGAGCCCATTTTCTTTGCTGATCGCCAATATGCTGACAAATGTCCCAGCCAATCTGAGG

**Eg-m104 [SSU1037]**  
 Eg-m104 GCCATGATGCACT-GAACACTGAGTTTTCTGAAT--CTGAAGTCAATCCTCTCCTGAG  
 Eg-m104.1 TTCATGATGCTTTCAAACTGAGTTTTCTGATTGTGATGAAGTCAAAGCTTCTCTGA-  
 \*\*\*\*\* \* \*\*\*\*\* \*

**Eg-m105 [LSU1502 (6)]**  
 AACGTGATGAGTCTCAATCAGCACTGCTTCTGACATTGATGATTCAAACAGCTGACA

**Eg-m106 [LSU39 (1)] Eg-m106.2 D box [LSU2981 (9)]?**  
 Eg-m106 CAAGTATGTCATGCTTTGTGTCCTTCCCCTGACCACTGATGCGCTCTGTCTGTAGCTGATT  
 Eg-m106.1 --CCGTGATGTCATGTTTGTGTCCTTCCCCTGACCGATGATGTTGCTCTTTCTGCAGCTGATC  
 Eg-m106.2 GGCAATGATGGCATGCTTTGTGTCCTTCCACTGACTGGTGATGTTGTTTTTCTGCAGCTGATT  
 \*\*\*\*\* \* \* \* \* \*

**Eg-m107 [LSU3315 (9)]**  
 Eg-m107 ACAAGGATGCTGGATTAAGCACGTTTCGGCTGACGGCCTTGGGTGATTTCTAGCTCCACCGCTGAGT  
 Eg-m107 ---GGATGCTGGATTAAGCACGTTTCGGCTGACGGCCTTGGGTGATTTCTAGCTCCACCGCTGA--  
 \*\*\*\*\*

**Eg-m108 [LSU1039 (5)]**  
 Eg-m108 GCTGTGATGC--GTTGTTCCATGCACCTGACACTGCGTGATGCGCTTTTCTGGTCTGAT  
 Eg-m108 GCTGTGATGC--GTTGTTCCATGCACCTGACACTGCGTGATGCGCTTTTCTGGTCTGA-  
 Eg-m108.1 GCTATGATGTTTGTGTTCCATGCAACTGATATTG-GTGATGCGGTTTCTGGTCTGA-  
 \*\*\* \*\*\*\*\* \*

**Eg-m109 [LSU1608 (6)]**  
 Eg-m109 TGCAGGATGACACCATTTTCATGGCCACCTCTGATTGCCATGCTGAGGCCCTTTCTGCTATCCCAACTCTGAGG  
 Eg-m109.1 TGCAGGATGACACCATTTTCATGGCCACCTCTGATTGCCATGCTGAGGCCCTTTCTGCTATCCCAACGCTGAGG  
 \*\*\*\*\*

**Eg-m110 [LSU1819 (6)]**  
 Eg-m110 CGCCTGATGACCATTTGCTTTTGATT-CCTGATCATTCCAAACCAAGATCTGTTCTGATG-  
 Eg-m110.1 CGCCTGATGACTATTGCTTTTGATT-CCTGATCATTCCAAACCAAGATCTGTTCTGATCG  
 Eg-m110.2 AGATGATGACC-CGTGCTGTGATTGCCCTGATTGCTGAAACCAAGATCTGTTCTGACT-  
 \* \*\*\*\*\* \*

**Eg-m111 [LSU2768 (8)]**  
 GCAGGGATGACCCCTTTGAGGCATTTGGCCTGACCGTTCCGCATGATTTACCCGAAGCGGCTGAGC

**Eg-m112 [LSU1081 (5)]**  
Eg-m112 GGGATGATGAACCATTTGCAAGCACTCGCTGACCACGATGCCGAGTTTCATCTGAG-  
Eg-m112.1 ATGATGATGAACCATTTGCAAGCACTCGCTGACCACGATGCAGAGTTTCTTCTGATG  
\*\*\*\*\*

**Eg-m113 [LSU2993 (9)]**  
CAAAATGATGCTGTGTTAGGTAGTGGACTGATCCGATTATGCTGCAGCATTTGTACGGACTGA

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

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

**Eg-m116 [SSU2180]**  
Eg-m116 TGAATGATGTTGTATCAACCTCTGATTCTGAGTTCGCTTTGAGGAGCCTCTT--TTCTGAT-  
Eg-m116.1 CATGTGATGCTCGA--AACCTCTGATTCTGATT-CTTATGATGAGCCTTTT--TCTCTGACT  
Eg-m116.2 TTCATGATGTCCAA--AACCTCTGATGCTGAA--CTCCTGATGTGCTTTT--CTCCTGACC  
Eg-m116.3 ----TGATGTCCAA--AACCTCTGATGCTGAA--CTCCTGATGTGCTTTTTCCTCCTGA--  
\*\*\*\*\*

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

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

**Eg-m119 [LSU3208 (9)]**  
Eg-m119 CAAAATGATGTGATCAACGTTGTGTTGTAACCTGACGCTTTTGTGATGACCATGCATCCCTGATT  
Eg-m119.1 CAAAATGATGTGATCAATGTTGTGTTGTAACCTGACGCTTTTGTGATGAACATGCATCCCTGAAT  
\*\*\*\*\*

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

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

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

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

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

**Eg-m125 [LSU3367 (9)]**  
Eg-m125 CAAGTATGATCTCAAGGCTTCGATACCACTGACTGAGATCTGAAGGCTATTTCTCCCTGA--  
Eg-m125.1 CAAGTATGCTCATCTCAAGGCTTCGATATCACTGACCGAGATCTGAAGGCTATT-CTCCCTGATG  
\*\*\*\*\*

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

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

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

**Eg-m129 [LSU2953 (8)]**  
Eg-m129 CAGAGGATGTACACACCTCCCACTTTTGACGGATTGATGCAACCCTGTTTCCACTGACT  
Eg-m129.1 CAGAGGATGTGAACACCTCCCACTTTTGACGGATTGATGCAACCCTGTTTCCACTGATT  
Eg-m129.2 CAGAGGATGTAAACACCTCCCACTTTTGACGGATTGATGCAACCCTGTTTCCACTGATT  
Eg-m129.3 CAGAGGATGTAAACACCTCCCACTTTTGACGGATTGATGCAACCCTGTTTCCACTGACT  
\*\*\*\*\*

**Eg-m130** [SSU1966]  
TTTATGATGACCAGCTCCATCTTGAAAAGACATGATGCTGTTGATACACCTTCACTGCCTGATC

**Eg-m131** [SSU1736]  
TGTGTGATGGATGGCAATGTCTCACTCGTCTGACTCCTGCTGCTGTGCACATGCTCTGACTGA

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

**Eg-m133** [SSU110]  
CAAAATGATGACCTCCAGATGGATGACTGCTTCTGAGATTCTTGAAGATACATCTATCTGAAC

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

**Eg-m135** [LSU1452 (5)]  
Eg-m135 GCAATGATGTTTCATACATGCTCAGAATGTGATGACTGCATGATCC-ACTTCTTT-GCTGACG  
Eg-m135.1 ACAATGATGTT-CA---ATGCTCAGAATGTGATAATTTCATGATCTTACCTCTTTTGCTGA--  
\*\*\*\*\* \*\*

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

**Eg-m137** [LSU2146 (7)]  
Eg-m137 ---CCAAAGATGAAACACAGCCCGCGTGTCTGAAGGTTTGTATGCTTCCTTCCTGGCTGAGG  
Eg-m137.1 GATGGACGGATGAATACAGCCCGCGTGTCTGA--TTTGTATGCTCTCTCTCGGCTGATTGCGGGATTGAAGAA  
\* \*\*\*\*\*

**Eg-m138** [LSU1888 (6)]  
Eg-m138 CCCATGATGAAAAATCTGTTTCAACCTTGCTGATGAGTTTCATCCAGTCCTCAAACCTGATG  
Eg-m138.1 CAAAATGATGAACATTTGCACAGATCAGAGTATGATGAGTTTCATCCAGTCCTCAAACCTGATG  
\* \*\*\*\*\*

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

**Eg-m140** [SSU57]  
TCTGTGATGACGGTCTGAAGCAGGTGATTCCGTCAACTGCAGCGCTGAGACGTGTCTGA

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

**Eg-m142** [SSU485]  
Eg-m142 CATTTGATGACACCTGGCTCCCTCTACTGAGGTGATGAGGAGATTCTGCATCGAACCCAT  
Eg-m142.1 CTCGTGATGACACCCGGCTCCCTCTACTGAGGTAATGAGGAGACTCTCCACCGAACCCAT  
\* \*\*\*\*\*  
Eg-m142 TTCTCTGAAG  
Eg-m142.1 TTCTTCTGAGG  
\*\* \*\*\*\*\*

**Eg-m143** [LSU2982 (9)]  
Eg-m143 CACATGATGGTTACATTTCACTGCATGACGCTCGGCTGAAGA-GCGAGTTCTGCTGAGT---  
Eg-m143.1 CATATGATGGTTTCAATTTCACTGCATGACGCGAGTGTATGACACTCTTCCCTCTGATTTAGC  
\*\* \*\*\*\*\*

**Eg-m144** [LSU2713 (8)]  
ACACATGATGCAGAGCAAATGGCTGAATCTTCCGACGGCAGACTTGAGGTTACATGATGAGATTCTGAGGG

**Eg-m145** [LSU1898 (6)]  
Eg-m145 TGGATGATGTTCAATTTCTATCGCAGACCAATGACGACTTATTGACAACCTGCTTCACTCTATGACT  
Eg-m145.1 TGGATGATTCGATTTCTATCGCAGACCAATGACGACTTATTGACAACCTGCTTCACTCTATGAC-  
\*\*\*\*\*

**Eg-m146** [LSU74 (1)]  
ACGATGATGAGACCATAGACTCTGCGTATCTGACTGCTGGATGAACACAATCTGCTGAAC

**Eg-m147** [LSU2981 (9)]  
Eg-m147 CCCATGATGGTGAACCTTCACTGCACTTGACCCAGCTGATGAGCAGCCCCAAGCTGA--  
Eg-m147.1 CAGATGATGGTGAATTTCACTGCACTTGACCCAGCTGATGAGGACGCTGGTGTCTGATT  
\* \*\*\*\*\*

Eg-m148 [LSU1836 (6)]  
 GTGATGATGAACCAAGAGTACTCGCTACCTGAGACTTCTGTGAACGAGAGTCTTACCGCCTGAC

Eg-m149 [LSU2883 (8)]  
 TGCATGAGGAGAATGTCTTCTTTCCCGACTGACGGAGATGCTGCAGTTTGTACGCTCTGAGC

Eg-m150 [SSU1616]  
 TGGATGATGAGCTTTCTTTGACCTGGTAACATTCTGACATTTTGATGAGGTTTATTTCTCTGA

Eg-m151 [SSU42]  
 CTGATGATGACCATTTGTTGCTTAGCCCTTGCTGACATCTCCTGATGTTTATTTGTTTCTGA

Eg-m152 [LSU3355 (9)]  
 CTGATGATGTTTTCCAGAATTCGCGCTCTGACGGAAGCTGATGCTCTGCTATTTACTGATC

Eg-m153 [LSU2825 (8)]  
 CGCAGGATGAACCATACGACTAGGGACAACCTGAAGTGCTGTGAAGCAATCTTCTTCTCTGAGC

Eg-m154 [LSU2242 (8)]  
 CGGATGATGTTTCCATCTGTTTCAATTGGAACACTGCATACCTGATGTTTATTTTATGCCTGAC

Eg-m155 [SSU565]  
 AGCGATGATGATTGCTTTTGATGAGTCCCATGATTCAACTTTTGTTCGTCGCTGACTCTGAGC

Eg-m156 [LSU1999 (7)]  
 TTGGGATGAGATTTTGACCTGTTTGCCATTCTCTGAGCATTGATGCTGTTCTGAAGCCTGAC

Eg-m157 [LSU2313 (8)]  
 GGAATGATGACTTTTGTGCTGCGCTGGAGATCCTGACAGTTTGCATTAGTCGAGGAGAGCTGACC

Eg-m158 [LSU3636 (10)]  
 GATGTGATGATCTTTTAACCAAGGGGAGTGTCTGACCGCTGAATGCCGAATTATTTTGCTGATT

Eg-m159 [SSU1063]  
 GTGATGATGTGATTATATGTCCAGAATGAAATGATGAGATTCTCAGCCGTCGCACTGAAA

Eg-m160 [SSU723]  
 CAGATGATGACCATGCCATGTGTGATATCTGAGCCTGTCAAAAGCAAGTGTTTTAACGCTGACG

Eg-m161 [SSU649]  
 GACAGGATGAGTGATACCTCCGGTTGACCTGATTCTCTGAAGTGAACGCATGGTTTCTGACA

Eg-m162 [LSU1562 (6)]  
 CTTGGTGATGTTTCTGTTGATCTGAGTATTCTGAGCATTCGTGAAGGCAGACCATACTCTCTGA

Eg-m163 [SSU393]  
 GAGGTGATGAGCAGTGGAACCTCTGAAGGCTGATTAAATTTCTCTTGCGATCTCTGACC

Eg-m164 [SSU1912]  
 GTCCTGATGAATTTTTTTCATCACAGACCTGTTCTGAGCAATGATGTTATGTATTTCTGACA

Eg-m165 [SSU2046]  
 CACATGATGCCGCTTCTTTCCCTGACTGTCTGACGCTTCAGAAACCATCTATCCCACTGATC

Eg-m166 [SSU99]  
 TCGCTGATGACTGTGTTCCCATGACCTTCCTGGATGTATTGATGCTGATGTGAGGAGTCTGATT

Eg-m167 [LSU2835 (8)]  
 GCTGTGATGTTTTCTCAGAATCTGACCCGATTGATGTGCATTCTTCCACAGACAGCAACTGA

Eg-m168 [LSU3347 (9)]  
 TGCCTGATGACCTCTCCTTGACACTTGACCCACGACGGCGATGTGAGTTTCCCGTTTGGTCTGA

Eg-m169 [SSU1899]  
 TCCGTGATGATTTGCGGACTGTTGGCTCTAGTTTCGATCTGTCATGAACCACTGCTGGCTGAC

Eg-m170 [LSU541 (3)]  
 GACAGGATGAATTTTCGGTCCCAGACTGAACATTATGAAGGAAGCAAAGCATTCACCTGACT

Eg-m171 [LSU1562 (6)]  
GCAATGATGTGTTCTCGTCTCTGATATCTTGCTGAGTAGCTGCGCACTACTGAGT

Eg-m172 [LSU2832 (8)]  
ATCATGATGGAAGTCTCCAGCACAGCGACTACTGACATGTATGAAGGTACCTTCGATCTGACC

Eg-m173 [LSU3222 (9)]  
GTTGTGATGATTTTCTGCTTCAGGACGCCTTCTGACGCCAACATGAGAGCACCTCTGCGGTCTGACA

Eg-m174 [SSU179]  
GCGATGATGACACCCCTGTGTATTAGCCACAAGCTGAGGAGATGATGTTTCCGCATCTTTGGACTGAGC

Eg-m175 [LSU2348 (8)]  
GGCGTGATGACAACCCATCCCTTAGCTGCATGCCGATCCCCGTGAGGCATTGATTTTTCAGCCTGAGG

Eg-m176 [SSU1744]  
TGCCGTGATGTCGCTGTTTGAGGCAGATGTCTGACTTTGCACGATTAGGGCATTTCAAACCGTCTGAGC

Eg-m177 [LSU2988 (9)]  
GGACGTGATGTTATCTCGAGTAGTGGTATTTGCTGACGACGAATGAGGCCAGAGTCTGTTTCTGATC

Eg-m178 [SSU1685]  
CTGATGATGACAATATTCCTCTCTATTCTGATGAACCATGACGTCATCTTCCATGCACCACCGTCTGATG

Eg-m179 [LSU2925 (9)]  
CAGGATGATGATTTTTTTTGTGGCGGACTCGCAACGTGACAGTCTCAAACCATGTTGTTGCACAGCTGAC

Eg-m180 [LSU3504 (9)]  
GGCTATGATGTTTGCTTTCTGAACAATCCAACACTGACGGTGCCGAGGATCTTATCTTGCACTGAAG

Eg-m181 [LSU3377 (9)]  
CAAGTGATGATCCACATCTGACACCTCTTGTCTGACGCCACTGAGACTCACAACCCCTGCTGCTGACG

Eg-m182 [SSU1096]  
TCGATGATGCCAACCCACAAGTCTTCTGCTCTGTGATGCACATCTCAACTTGGTCGATGCCACTGAGG

## Box AGAUGN 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.

Eg-p14 [LSU3235 (9)]  
Eg-p14 AGGGGCACCCCTCTGAGGCATCCTGGAGGCGCGCTGGCTCTCTCCTGGGTGCATGCTGTGGTGCCCAGATGG  
Eg-p14.1 GGGGGCACCCCTCTGAGGCATCCTGGAGGCGCGCTGGCTCTCTCCTGGGTGCATGCTGTGGTGCCCAGATGG  
Eg-p14.2 -GGGGCACCCCTCTGAGGCATCCTGGAGGC--GCTGGCTCTCTCCTGGGTGCATGCTGTGGTGCCCAGATGG  
\*\*\*\*\*

Eg-p15 [LSU2752 (8)]  
Eg-p15 TGCACTTTAAAGAGCCCCGGCCCTGCCTTTCGGCATCGTGGGCTGTGGGGTGTCATAGGGTGCAGAGATGA  
Eg-p15.1 TGCACTTTAAAGAGCTCCTGGCTCTGCCCTTCGGCATCGTGGGCTGTGGGGTGTCATAGGGTGCAGAGATGA  
Eg-p15.2 TGCACTTTACGAGCCCCGGCCCTGCCTTTCGGCATCGTGGGCTGTGGGGTGTCATAGGGTGCAGAGATGA  
\*\*\*\*\*

Eg-p16 [SSU1592]  
Eg-p16 AAGCAGGACGTCAAATCCATGGGGGAGATGCTGTGGTCCCTTTCGGAGCTAAATCCTGCCAGAG-  
Eg-p16.1 -AGCAGGACGTCAAATCCCTGGGGGACATGCTGTGGTCCCTTTCGGAGCTAAATCCTGCCAGAG---  
Eg-p16.2 AAGCAGGACGTCAAATCCATGGATGCAGATGCTGTGGTCCCTTTCGGAGCTAAATCCTGCCAGAGAGA  
\*\*\*\*\*

**Eg-p17 [LSU3701 (10)]**  
Eg-p17 CAGCACAAACGCGTTACGTCGGTATCAGTCTCATGGTGCCGGGGTGTGAGATGTTGTGCCAGATCC  
Eg-p17.1 CAGCACAAACGCGTACGTCGGTATCACTCTCATGGTGCCGGGGTGTGAGATGTTGTGCCAGATCC--  
Eg-p17.2 CAGCACAAACGCGTACGTCGGTATCAGCCTCATGGTGCCGGGGTGTGAGATGTTGTGCCAGATCC--  
Eg-p17.3 CAGCACAAACGCGTTACGTCGGTATCAGTCTCATGGTGCCGGGGTGTGAGATGTTGTGCCAGATCC--  
Eg-p17.4 CAGCACAAACGCGTTACGTCGGTATCAGTCTCTTGGTGCCGGGGTGTGAGATGTTGTGCCAGATCC--  
Eg-p17.5 CAGCACAAACGCGTACGTCGGTATCAGTCTCATGATGCCGGGGTGTGAGATGTTGTGCCAGATCC--  
Eg-p17.6 CAGCACACACGCGTTACGTCGGTATCAGTCTCATGGTGCCGGGGTGTGAGATGTTGTGCCAGATCC--  
Eg-p17.7 CAGCACAGACGCGTTACGTCGGTATCAGTCTCATGGTGCCGGGGTGTGAGATGTTGTGCCAGATCC--  
Eg-p17.8 CAGCACAAACGCGTTACGTCGGTATCAGTCTCATGGTGCCGGGGTGTGAGATGTTGTGCCAGATCC--  
\*\*\*\*\*

**Eg-p18 [SSU1393]**  
Eg-p18 GTGCTGTGATCTGATGCCTGTGCCGGTGCCTTCTCTGGCAGGGTTCTTTCCACAGCAAGAT--  
Eg-p18.1 GCGCTGTGATCTGATGCCTGTGGTCCACTCTGGCAGGGTTCTTTCCACAGCAAGAT--  
Eg-p18.2 GTGCTGTGATCTGATGCCTGTGCCGGTGCCTTATCTGGCAGGGTTCTTTCCACAGCAAGATGA...  
\* \*\*\*\*\*

**Eg-p19 [LSU3953 (13)]**  
CAGTGCCTGGCCTCTGAAACAGCGCGTGGGTGGCGTCCCCGCGCTGTCTTAAGCAGGCAGAGA

**Eg-p20 [LSU2915 (8)]**  
Eg-p20 ACAGGGGTGCAGTATCGGGCAGATGGCAGCCTGGGTGAGTTCATTTGCCTACTAAACCACCCCGAGATG--  
Eg-p20.1 -CAGGGGTGCAGTATCGGGCAGATGGCAGCCTGGGTGAGTTCATTTGCCTATTAAACCACCCCGAGATGT  
Eg-p20.2 ACAGGGGTGCAGTATCGGGCAGATGGCAGCCTGGGTGAGTTCATTTGCCTACTAAACCACCCCGAGATG--  
\*\*\*\*\*

**Eg-p21 [LSU2914 (8)]**  
Eg-p21 CTGCCAGGATTTTATCACCCCTCCTGACACATTTATGTTGGAGGGGCTAGAGGTCTGGTAGATGT  
Eg-p21.1 ATGCCAGGCTTTTCTCACCCCTCCTGACAAATTTATGTTGGAGGGGCTCGAGGTCTGGTAGATGT  
Eg-p21.2 CAGCCAGGATTTTATCACCCCTCCTGACACATTTATGTTGCCGGGCTAGAGGTCTGGTAGATGT...  
Eg-p21.3 -TGCCAGGACTTTTATCACCCCTCCTGACACATTTATGTTGGAGGGGCTAGAGGTCTGGTAGATGT...  
\*\*\*\*\*

**Eg-p22 [LSU1023 (5)]**  
Eg-p22 ATCGAGGTACTAGCCAGGGCATGCGATCTTTCTGCATATCCCTTCTCAATGGCCTCGTAGACCT  
Eg-p22.1 ATCGAGGTACTAGCCAGGGCATGCGATCTTTCTGCATATCCCTTCTCAATGTCTCGTAGAT--  
\*\*\*\*\*

**Eg-p23 [LSU1266 (5)]**  
Eg-p23 CATGCACTTTTACGACCTCTCCTTAAATCTTGAGGCTGGGGTTCCCTGAGAGTGCACAGATCC  
Eg-p23.1 CATGCACTTTTAAAGCCTCTCCTTAAATCTTGAGGCTGGGGTTCCCTGAGGCTGCACAGATCC  
Eg-p23.2 CATGCACTTTTAAAGCCTCTCCTTAAATCTTGAGGCTGGGGTTCCCTGCGGGTGCCAGATCC  
Eg-p23.3 ATGCACTCCAGCCTCTCCTTAAATCTTGAGGCTGGGGTTCCCTGAGGCTGCACAGATCC  
Eg-p23.4 CATGCACTCTAAGCCTCTCCTTAAATCTTGAGGCTGGGGTTCCCTGGGGTGCACAGATCC  
Eg-p23.5 CATGCACTTTAAGCCTCTCCTTAAATCTTGAGGCTGGGGTTCCCTGAGGCTGCACAGATCC  
\*\*\*\*\*

**Eg-p24 [LSU280 (2)]**  
Eg-p24 CTGGCAATATTTTACAGCGGAGCTCAGTCCCGTCTCTGCTCCGAGTGCAATTTGCCAGATTG  
Eg-p24.1 CTGGCAATATTTTACAGCGGAGCTCAGTCCCGTCTCTGTTCCGAGTGCAATTTGCCAGATTG  
\*\*\*\*\*

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

**Eg-p26 [LSU3697 (10)]**  
Eg-p26 AAGGCAGCTCCAGAGCACCTTGAAGCCAGTGCTTTTCCATGGTGCGTATCCTGCTGCCAAGATTG  
Eg-p26.1 AAGGCAGCTCCAGCGCACCTTGAAGCCAGTGCTTTTCCATGGTGCGTATCCTGCTGCCAAGATTG...  
Eg-p26.2 AAGGCAGCTCCCGAGCACCTTGAAGCCAGTGCTTTTCCATGGTGCGTATCCTGCTGCCAAGATTG...  
\*\*\*\*\*

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

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

**Eg-p29 [SSU1660]**  
Eg-p29 CCGCCTGTGATCAAGGGTGTTCGGGCTTTTGCATGGCCGCTACATCAGAAAGGCAGGCAAGAATC  
Eg-p29.1 CCGCCTGTGATCAAGGGTGTTCGGGCTTTTGCATGGCCGCTACATCAGAAAGGCAGGCAAGAATC  
\*\*\*\*\*



**Eg-p41 [LSU1926 (6)]**  
CAGGTGACGCACATTCCCTGGGTAGCATGAACTTGCTTCCTTGCGCACTGTTACCGAGATT

**Eg-p42 [SSU89] (related to Eg-p36)**  
Eg-p42 TTGCACTCTGCCATTGGATCAGGCGATTATGCTTGTCGCCTGGTCCCAGAAAGGAGTGCCAGATTT...  
Eg-p42.1 GTGCACTCTGCCATTGGATCAGGCAATTATGCTTGTCGCCTGGTCCCAGAAAGGAGTGCCAGATTT...  
Eg-p42.2 TTGCACTCTGCCATTGGATCAGGCGATTATGCTTGTCGCCTGGTCCCAGAAAGGAGTGCCAGATTT...  
Eg-p42.3 TTGCACTCTGCCATTGGATCAGGCGATTACGCTTGTCGCCTGGTCCCAGAAAGGAGTGCCAGATTT...  
Eg-p42.4 TTGCACTCTGCCATTGGATCAGACGATTATGTATGTCGCCTGGTCCCAGAAAGGAGTGCCAGATTT...  
\*\*\*\*\* \* \*\*\*\*\*

**Eg-p43 [LSU2746 (8)]**  
Eg-p43 ACGCACGTGCCAGAGTCAGCAGGCCTGCCGGGAGTGCAAGTCCATGTGTTATGCATGTGCCAGAAGG  
Eg-p43.1 ACGCACGTGCCAGAGTCAGCCGGCCTGCCGGGAGTGCAAGTCCATGTGTTATGCATGTGCCAGA---  
Eg-p43.2 ACGCACGTGCCAGAGTCAGCAGGCCTGCCGGGAGTGCAAGTCCATGTGTTATGCATGTGCCAGAAGG  
Eg-p43.3 ATGCACGTGCCAGAGTCAGCAGGCCTGCCGGGAGTGCAAGTCCATGTGTTATGCATGTGCCAGA---  
\* \*\*\*\*\*

**Eg-p44 [SSU1624]**  
AAGCCTTGTTATCCTCCAGCCGCCTTCTCTGGGCTCTGGCCTTCCCTCAAGGCAAGATTTC

**Eg-p45 [LSU2837 (8)]**  
Eg-p45 GCGCATGACTCACCAGCCCGAGGCTCTCGTGGATGTCGCGCCATGCGGCCAGCAGGAATGCAAGAGGT  
Eg-p45.1 GCGCATGACTCACCAGCCCGAGGCCCTCGTGGATGTCGCGCCATGCGGCCAGCAGGAATGCAAGAGGT  
\*\*\*\*\*

**Figure A.3.4. Newly identified *Euglena gracilis* snoRNA sequences.** Description for C/D and AGAUGN RNAs appear at the start of their respective list of RNAs above.



## Orphan C/D box snoRNAs

>Eg-m-Orphan1

TGATGACNATTGAAGCATTGGACGCCAGATGACATTCTGAAACAATCCAACCTCTTTCTGA

> Eg-m-Orphan2

TCTTGATGATGATGGCGACGATGATAATGATGATGATGTTGATGCTGATGAT

> Eg-m-Orphan3

GCGATGATGATTGCTTTTGATGAGTCCCATGATTCTGCCCAACTTTTGTTCGTCGCTGACTCTGA  
G

> Eg-m-Orphan4

TGATGACCTAACGATGCATGCCGACCATGTCTGATGTTTCGAGACCCTGA

> Eg-m-Orphan5

CAGAGGATGGATGTGAATCTCGTTGAACCGAACGCATTGTGAAGGAGATGGCTTGGATTGCTGAC  
CN

> Eg-m-Orphan6

TGATGTTTCTGTTGATCTGAGTATTCTGAGCATTCGTGAAGGCAGACCATACTCTCTGA

> Eg-m-Orphan7

GGATGTCCACTGGTTTGCACCGCTGATGAGGTTTGATGTGATTCTCTTTTGA CTGA

> Eg-m-Orphan8

GGATGAATTCATTTGTAGTGTTCTCTGACTCCGATGACGTCCACCAGGATCTGA

> Eg-m-Orphan9

TGATGTTTTTAATTCAAGTTCTGATTTCGAGTTCATCTGATTCCGATTCCAATTCCGATTCTGA

> Eg-m-Orphan10

TGATGTTTTTCCAGAATTCGCGCTCTGACGGAAGCTGATGCTCTGCTATTTACTGA

> Eg-m-Orphan11

CAAAATGATGTGTCACATTTCCCTGGCACTGAGGTGCACGCCACCAC TGATGCTTTTCATCCCTGA  
CCC

> Eg-m-Orphan12

TGATGGCCATTGCTGACATCCTTACTGAATCTGTTGATCACCCAGACTGA

> Eg-m-Orphan13

TGATGAGAGCGGTGGGTGGACTGTCTCTGAATTTCCACTTCTTTTGCTTGCTTTGCTGA

> Eg-m-Orphan14

TGATGATCCACATCCTGACACCTCTTGTCGCCGCGCACTTGGGCCCTGA

```

> Eg-m-Orphan15
TGATGGCACGCCCCAAACGTGAGCAGTGAAATTCTTGAATGAAGGCTCACTGGCTGA

> Eg-m-Orphan16
TGATGATCACCCCCACCGCCCCCGAGTTCCGAATCCAA TGCTGGACGTCACCCTTGGCTGA

> Eg-m-Orphan17
TGATGCCGCACTTCGTCTGCTGAATGCACACTCCTCTTGCTGGCTTTGCTGA

```

### Orphan box AGAUGN snoRNAs

```

>Eg-p-Orphan1
TGGCACTGCCACAGGCTCCACCTGCGCCACGCTCTCGCTTGTGGAGAAACCAACGGTGCCAGAT
GCTCGCT

>Eg-p-Orphan2
GTGCAGCTTCAAAGCGACTGCTCCATGTGTGTGTGGATGCCAGTTTCATCCTGAGCTGCAAGAGG
T

>Eg-p-Orphan3
CCCGAGGCAACTTGGTTGGGACCAGCTTGCTCTCATGCTTGTCACCGGCTTCATACCTCGAAGAT
CT

>Eg-p-Orphan4
AAGCCTTGTTTTGCGTTCCATGGCTTCTCATGAGCTTTGGACCTTCCCTCAAGGCAAGATTT

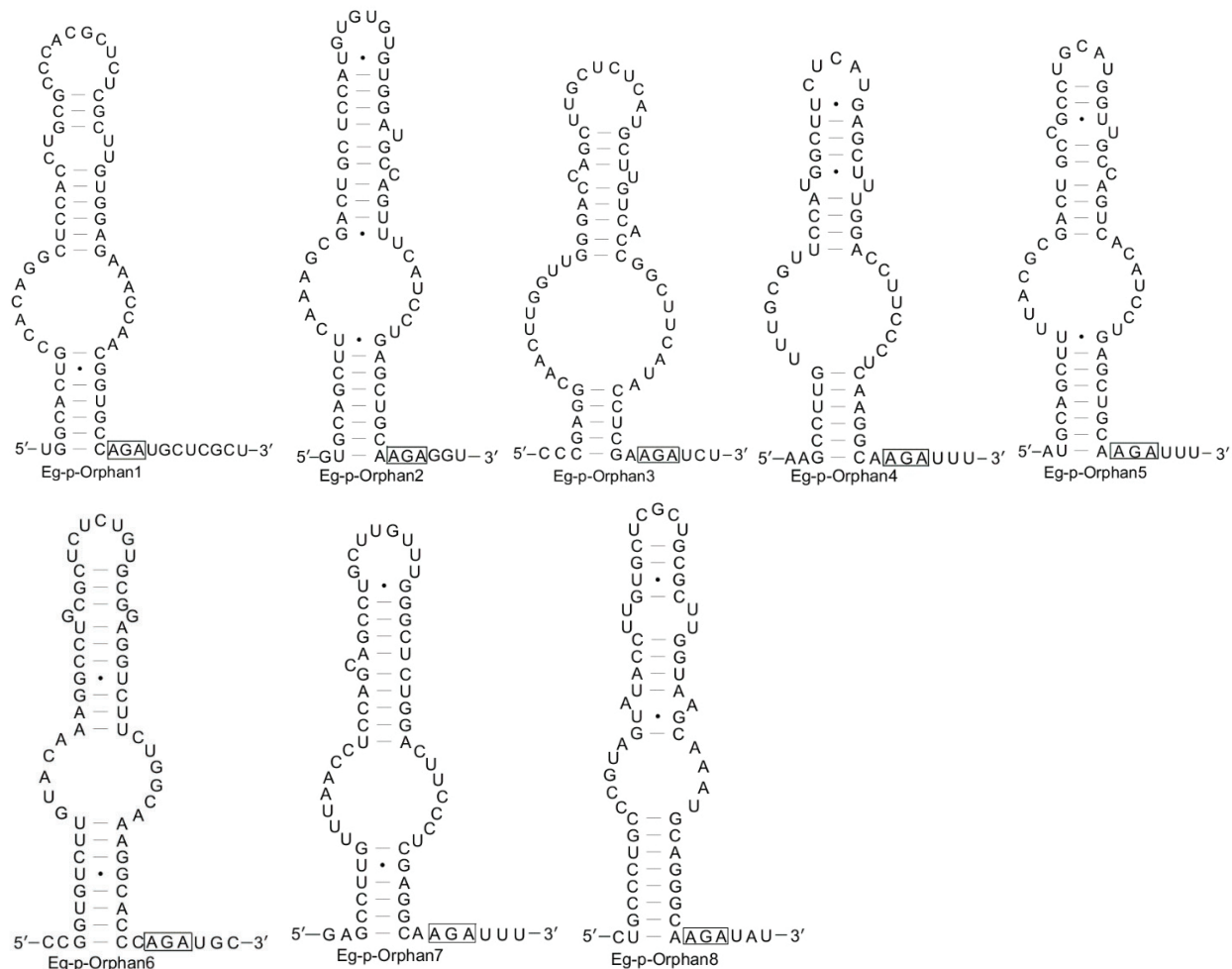
>Eg-p-Orphan5
ATGCAGCTTTTACGCGACTGCCGCCTGCATGGTTGCCAGTCACATCCTGAGCTGCAAGATTT

>Eg-p-Orphan6
CCGGTGTCTTGTACAAAGGCCTGCGCTCTCTGTGCGGAGGTCTTCTGGCAAAGGCACCCAGATGC

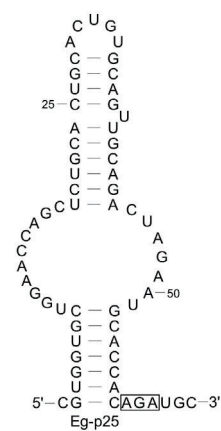
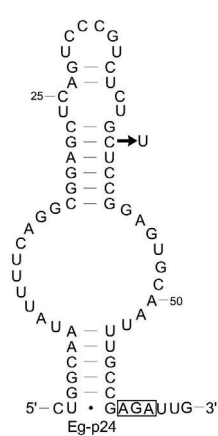
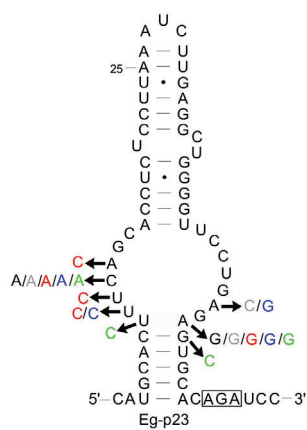
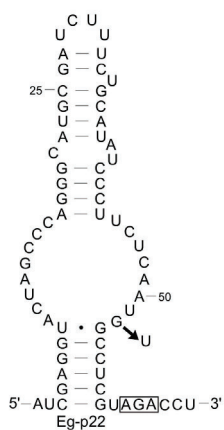
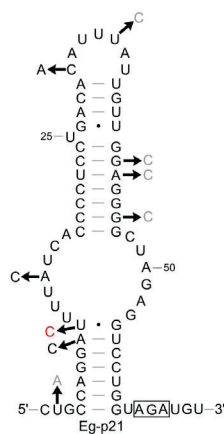
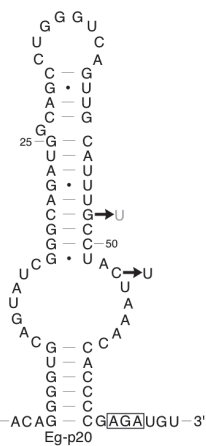
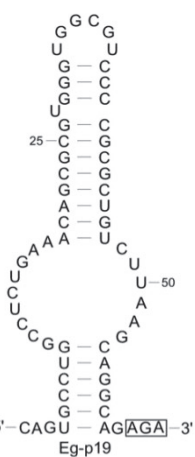
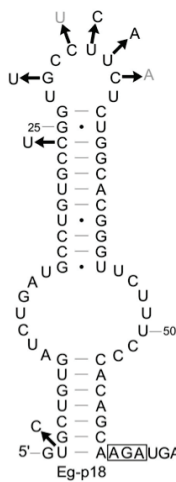
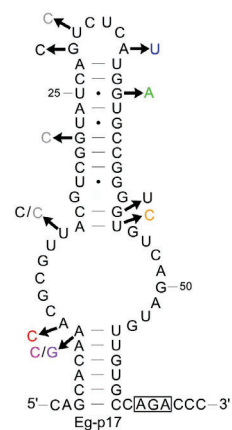
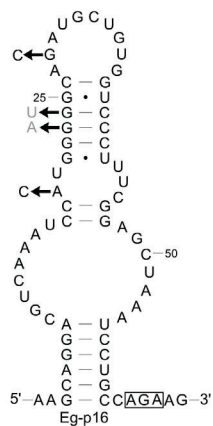
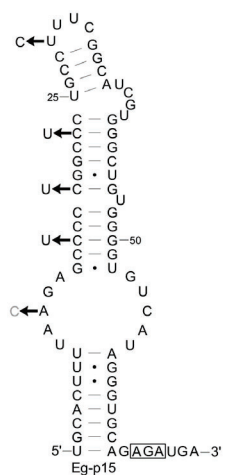
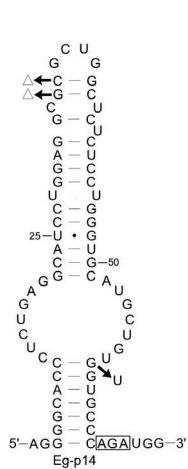
>Eg-p-Orphan7
GAGCCTTGTTTTAACCTCCAGCAGCCTGCTTGTTTGGGCTCTGGACTTCCCTCGAGGCAAGATTT

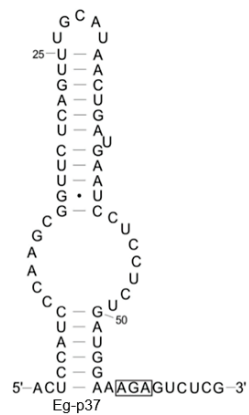
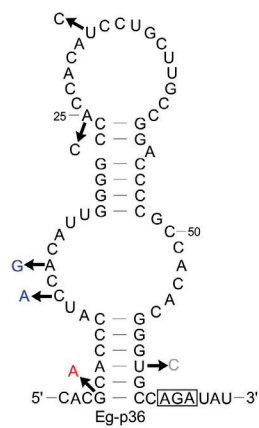
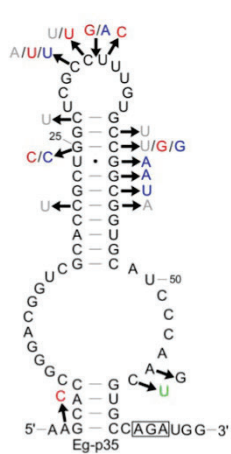
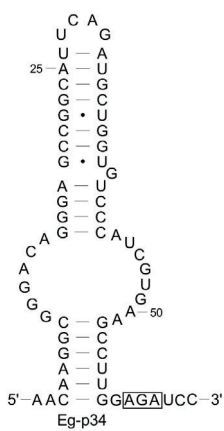
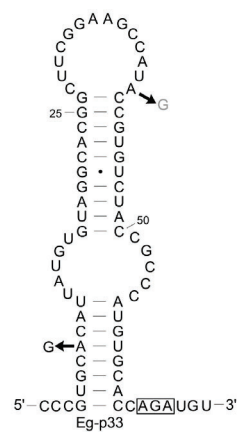
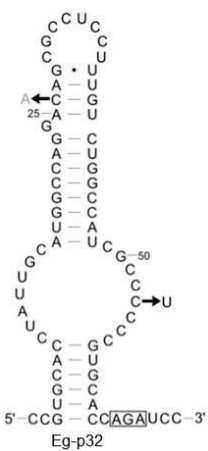
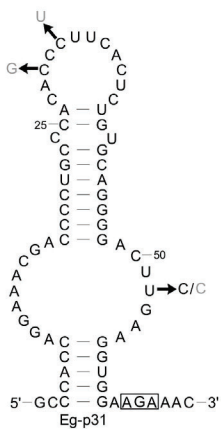
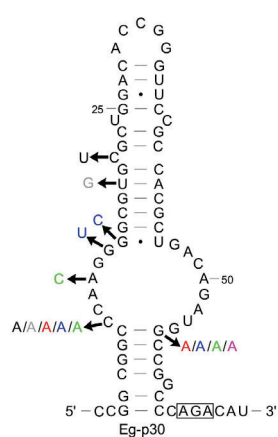
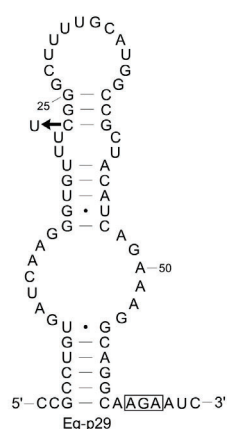
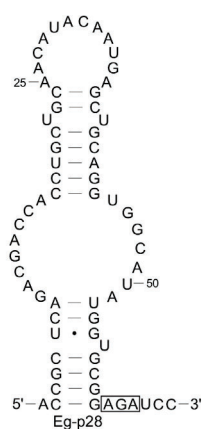
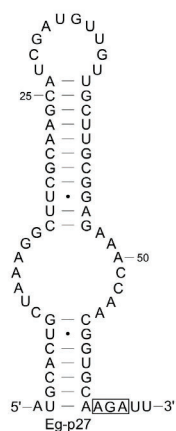
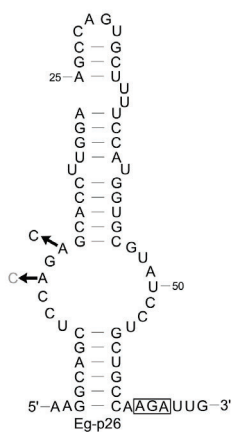
>Eg-p-Orphan8
CTGCCCTGCCCCGTAGTATACCTTGTGCTCGCTGCGCTTGTAAGCAAATGCAGGGCAAGATAT

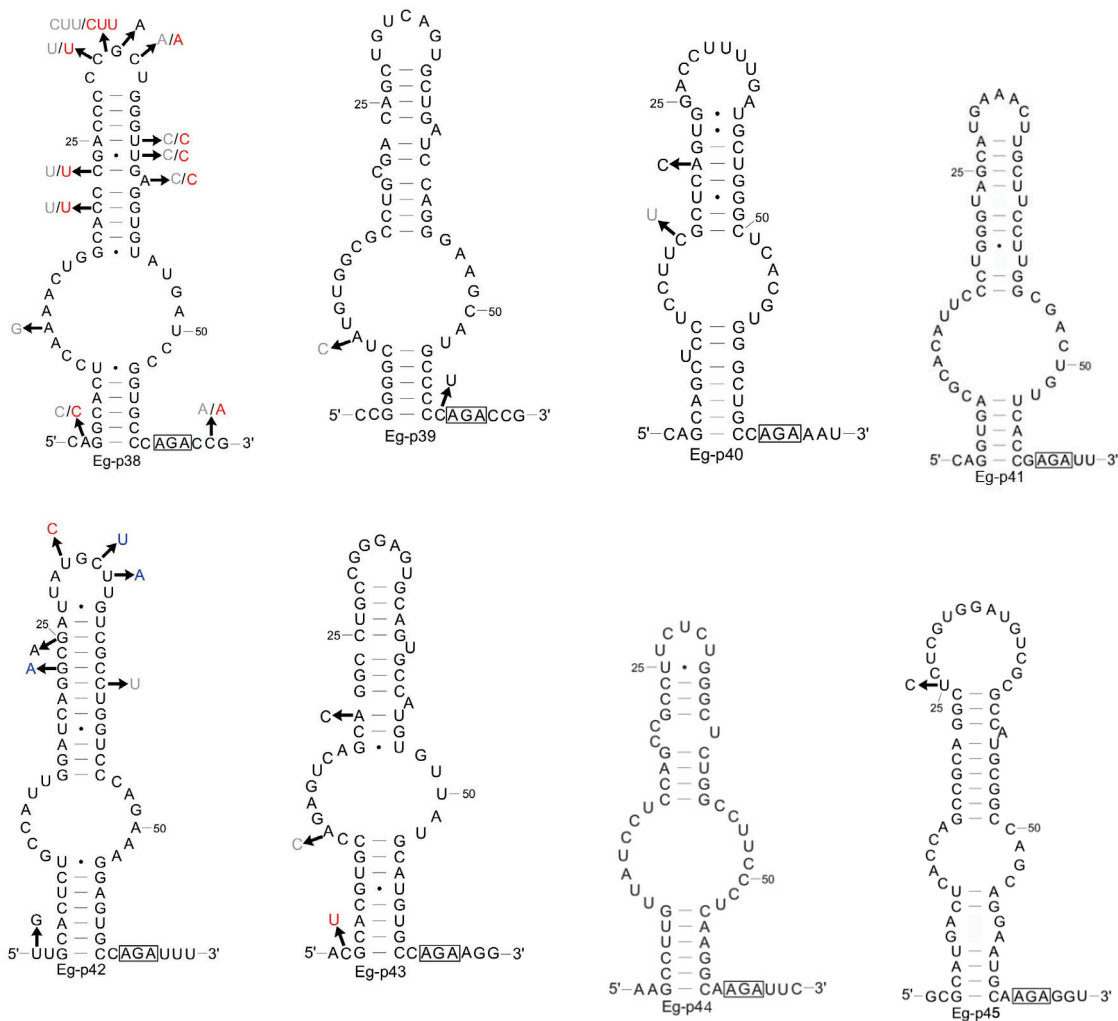
```



**Figure A.3.5. 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 box elements are highlighted as in Figure A.3.4. Predicted secondary structures for orphan AGAUGN snoRNAs formatted the same as in Figure A.3.6.







**Figure A.3.6. Predicted secondary structures of identified *E. gracilis* box AGAUGN modification guide snoRNAs.** In the structures, within AGAUGN box elements the highly conserved AGA sequence is highlighted 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

**Met (CAT)**

gcgagcttggcgcagtcggtagcgcgtaggtctcataatcctaaggctcgtgagttcgatcctcacagctcgca

**His (GTG)**

gggaagatagtttaggggcagaacatcacgttgtggccgtggatacccggttcgattcccggctcttccctcca

**Glu (CTC)**

tccccgatagtatagtggctagaacatcggcttctcacgccgaaaacccgggtcaattcccggtcggggagcc

**Glu (TTC)**

tccctgatagtatagtgggtagaacacctggcttccaccaggagacccgggtcaattcccggtcaggagcca

**Ser (AGA)**

gcactcatacccaagcgggttacgggggtggactagaaatccactgtgatcttcacgcgcaggttcgagtcctgctgagtgcg

**Gln (TTG)**

gtcccatagtgtagtggtcagcaccaggacttgaatcctgtgaccccggttcgagtcgggtggagtgcc

**Gln (CTG)**

ggctctatagtgtagcggtagcacatgggattctgattcccataaccccggttcgaatccgggtaggacctcca

**Cys (GCA)**

gggttcatagctcagtggttagagcattcggctgcaatccgagtggtccctggtcaaacccgggtgggccct

**Pro (TGG)**

ggccgttgggtctaggggcatgattctcgcttgggtgcgagaggtcccggttcgattcccggaacggccc

**Pro (CGG)**

ggccgcttgggtctcggggcatgattctcgctccgggtgcgagaggtcccggttcgattcccggagcggccc

**Ala (AGC)**

cggcgtgtagctcagtggttagagcgtcgtcagcatgcgagaggtcctgggatcgatcccagctcgcca

**Ala (CGC)**

gggcgtgtagctcagtggttagagcggcgcttcgcatgcgggaagtccagggttcgatcccctgttcgtcca

**Ala (TGC)**

tgggcgttagctcagtggttagagcggcgcttgcgcatgcgggatgtcgtgggttcaaaccacatcgtccag

**Leu (CAG)**

tggtcagatgggtcgagcgggtccaagacggcagcttcaggtgctgttctcctccggggcgtgggttcaaactcctgatcag  
c

**Figure A.3.7. Newly identified tRNA sequences from *E. gracilis*.** DNA sequences of newly identified tRNAs with anti-codon sequence in parentheses.

**Table A.3.1. Oligonucleotides used during synthesis of the *E. gracilis* small/capped RNA library**

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

**Table A.3.2. Blocker oligonucleotides used during PCR amplification of cDNA synthesized from *E. gracilis* small RNA or cap-enriched RNA. Oligos used to prevent amplification of *E. gracilis* specific LSU rRNA fragments are listed with blocker and linker nucleotides indicated in capital and lowercase letters respectively.**

Oligonucleotide	LSU fragment targeted	Sequence
<b>oAM264</b>	14	5' GGCTTTGATGAAAtgtccgatccgtt/3SpC3/ 3'
<b>oAM284</b>	11	5' GGCTTTGATGAAAggagcatcgaggc/3SpC3/ 3'
<b>oAM286</b>	13	5' GGCTTTGATGAAAccaacaccccgcc/3SpC3/ 3'
<b>oAM288</b>	12	5' GGCTTTGATGAAAcggagtgttgc/3SpC3/ 3'
<b>oAM290</b>	4	5' GGCTTTGATGAAAtgaccaagcgtct/3SpC3/ 3'
<b>oAM292</b>	2	5' GGCTTTGATGAAAgtagctggcgca/3SpC3/ 3'
<b>oAM294</b>	1	5' GGCTTTGATGAAAcctgtgtgttg/3SpC3/ 3'
<b>oAM296</b>	10	5' GGCTTTGATGAAAtgggcgtgacaat/3SpC3/ 3'
<b>oAM298</b>	7	5' GGCTTTGATGAAAcgggcaggaatgg/3SpC3/ 3'
<b>oAM300</b>	3	5' GGCTTTGATGAAActcgatcggtt/3SpC3/ 3'
<b>oAM301</b>	6	5' GGCTTTGATGAAAcgttgaacaatgg/3SpC3/ 3'
<b>oAM302</b>	9	5' GGCTTTGATGAAAggtgcgagcctgc/3SpC3/ 3'
<b>oAM303</b>	5	5' GGCTTTGATGAAAtcgtgtgtgttg/3SpC3/ 3'
<b>oAM304</b>	8	5' GGCTTTGATGAAAaagtgagcagtcac/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 A.3.3. Reads per million (RPM) for RNAs found in single end reads for size-selected and TMG-capped small RNA libraries.** RNAs identified in previous studies are displayed on a gray background. \*indicates these RNAs were found in a separate library file not used for RPM calculations.

Reads per million (RPM)			Reads per million (RPM)		
RNA name	Size-selected Library	TMG-capped library	RNA name	Size-selected Library	TMG-capped library
<b>Methylation Guide snoRNAs</b>			Eg-m130	2.621776608	2.459562739
Eg-m1	19.00788041	11.47795945	Eg-m131	7.209885671	2.459562739
Eg-m2	2.621776608	17.21693917	Eg-m132	11.14255058	25.41548164
Eg-m3	126.5007213	5.738979724	Eg-m133	6.554441519	0
Eg-m4	137.6432719	4.099271232	Eg-m134	7.865329823	32.79416985
Eg-m5	62.26719443	14.75737643	Eg-m135	5.243553215	4.919125478
Eg-m6	98.97206694	0.819854246	Eg-m136	19.66332456	9.838250956
Eg-m7	8.520773975	0	Eg-m137	9.176218127	0
Eg-m8	62.26719443	14.75737643	Eg-m138	2.621776608	0
Eg-m9	30.15043099	27.05519013	Eg-m139	2.621776608	0
Eg-m10	59.64541782	0.819854246	Eg-m140	1.310888304	1.639708493
Eg-m11	8.520773975	2.459562739	Eg-m141	1.310888304	0
Eg-m12	36.04942835	0.819854246	Eg-m142	15.07521549	4.099271232
Eg-m13	28.18409853	0	Eg-m143	1.310888304	0.819854246
Eg-m14	64.88897104	74.60673641	Eg-m144	410.3080391	227.0996262
Eg-m15	84.55229559	127.8972624	Eg-m145	17.6969921	19.67650191
Eg-m16	260.2113283	4.099271232	Eg-m146	0.655444152	0
Eg-m17	487.650449	0	Eg-m147	1.966332456	0

Eg-m18	31.46131929	22.9559189	Eg-m148	11.79799473	0.81985424 6
Eg-m19	211.7084611	98.38250956	Eg-m149	5.243553215	1.63970849 3
Eg-m20	11.79799473	0	Eg-m150	4.588109063	8.19854246 3
Eg-m21	20.31876871	6.55883397	Eg-m151	81.93051899	114.779594 5
Eg-m22	2.621776608	0	Eg-m152	11.79799473	6.55883397
Eg-m23	58.33452952	64.76848546	Eg-m153	56.36819706	2.45956273 9
Eg-m24	49.81375554	11.47795945	Eg-m154	1.310888304	9.83825095 6
Eg-m25	21.62965701	0.819854246	Eg-m155	0.655444152	0
Eg-m26	150.0967108	20.49635616	Eg-m156	43.91475818	74.6067364 1
Eg-m27	58.98997367	25.41548164	Eg-m157	5.243553215	0
Eg-m28	45.22564648	2.459562739	Eg-m158	106.1819526	0.81985424 6
Eg-m29	13.10888304	1.639708493	Eg-m159	2.621776608	0.81985424 6
Eg-m30	110.7700617	77.06629915	Eg-m160	14.41977134	0.81985424 6
Eg-m31	21.62965701	2.459562739	Eg-m161	14.41977134	3.27941698 5
Eg-m32	1.310888304	6.55883397	Eg-m162	20.31876871	0
Eg-m33	210.3975728	0	Eg-m163	5.243553215	0
Eg-m34	5.898997367	1.639708493	Eg-m164	1.310888304	1.63970849 3
Eg-m35	5.898997367	0.819854246	Eg-m165	58.98997367	10.6581052
Eg-m36	27.52865438	0	Eg-m166	0	0.81985424 6
Eg-m37	70.7879684	0	Eg-m167	2.621776608	0
Eg-m38	205.1540195	24.59562739	Eg-m168	54.40186461	244.316565 4
Eg-m39	68.1661918	20.49635616	Eg-m169	41.94842572	2.45956273 9
Eg-m40	13.76432719	1.639708493	Eg-m170	3.277220759	0

Eg-m41	1.310888304	0	Eg-m171	9.831662278	1.63970849 3
Eg-m42	64.88897104	0	Eg-m172	3.932664911	0
Eg-m43	35.3939842	1.639708493	Eg-m173	13.10888304	6.55883397
Eg-m44	3.932664911	0.819854246	Eg-m174	26.87321023	3.27941698 5
Eg-m45	5.243553215	0.819854246	Eg-m175	138.2987161	2.45956273 9
Eg-m46	15.73065965	2.459562739	Eg-m176	11.79799473	9.83825095 6
Eg-m47	40.63753742	2.459562739	Eg-m177	32.11676344	4.09927123 2
Eg-m48	12.45343889	0.819854246	Eg-m178	57.02364122	7.37868821 7
Eg-m49	2.621776608	1.639708493	Eg-m179	1.310888304	7.37868821 7
Eg-m50	26.87321023	4.919125478	Eg-m180	17.04154795	19.6765019 1
Eg-m51	0	0.819854246	Eg-m181	184.8352508	4.91912547 8
Eg-m52	49.15831139	0	Eg-m182	61.61175028	18.0367934 2
Eg-m53	98.97206694	17.21693917	<b>Pseudouridylation Guide snoRNAs</b>		
Eg-m54	13.10888304	122.9781369	Eg-p1	74.72063332	1.63970849 3
Eg-m55	56.36819706	3.279416985	Eg-p2	83.24140729	0.81985424 6
Eg-m56	254.9677751	73.78688217	Eg-p3	244.4806687	1.63970849 3
Eg-m57	9.831662278	18.85664767	Eg-p4	15.73065965	0.81985424 6
Eg-m58	1.966332456	0	Eg-p5	4.588109063	1.63970849 3
Eg-m59	59.64541782	11.47795945	Eg-p6	20.97421286	0
Eg-m60	3.277220759	0.819854246	Eg-p7	505.3474411	23.7757731 4
Eg-m61	79.30874238	0.819854246	Eg-p8	0.655444152	0.81985424 6
Eg-m62	2.621776608	0	Eg-p9	19.00788041	10.6581052

Eg-m63	72.75430086	46.73169204	Eg-p10	0.655444152	0
Eg-m64	135.6769394	4.919125478	Eg-p11	59.64541782	0
Eg-m65	1.966332456	0	Eg-p12	0.655444152	0
Eg-m66	17.6969921	0	Eg-p14	9.831662278	0
Eg-m67	4.588109063	0	Eg-p15	10.48710643	2.45956273 9
Eg-m68	1.966332456	1.639708493	Eg-p16	76.03152162	0
Eg-m69	1045.433422	30.33460711	Eg-p17	344.1081797	4.09927123 2
Eg-m70	49.15831139	13.11766794	Eg-p18	5.898997367	0
Eg-m71	134.3660511	11.47795945	Eg-p19	1.310888304	0
Eg-m72	92.41762542	64.76848546	Eg-p20	5.898997367	0
Eg-m73	0	0	Eg-p21	7.865329823	0
Eg-m74	7.865329823	14.75737643	Eg-p22	3.277220759	0
Eg-m75	15.73065965	1.639708493	Eg-p23	14.41977134	0
Eg-m76	4.588109063	2.459562739	Eg-p24	9.176218127	19.6765019 1
Eg-m77	39.98209327	4.919125478	Eg-p25	0.655444152	0.81985424 6
Eg-m78	48.50286724	6.55883397	Eg-p26	98.31662278	0.81985424 6
Eg-m79	33.42765175	27.87504437	Eg-p27	0.655444152	0
Eg-m80	72.75430086	1.639708493	Eg-p28*	0	0
Eg-m81	193.3560248	4.919125478	Eg-p29	25.56232192	1.63970849 3
Eg-m82	81.27507484	1.639708493	Eg-p30	164.5164821	4.09927123 2
Eg-m83	0.655444152	2.459562739	Eg-p31	5.243553215	0
Eg-m84	151.4075991	1.639708493	Eg-p32	239.2371154	3.27941698 5
Eg-m85	43.25931403	214.8018125	Eg-p33	151.4075991	0
Eg-m86	170.4154795	19.67650191	Eg-p34	0.655444152	0
Eg-m87	30.80587514	9.018396709	Eg-p35	63.57808273	17.2169391 7
Eg-m88	25.56232192	50.01110903	Eg-p36	228.750009	13.1176679 4

Eg-m89	2.621776608	0.819854246	Eg-p37*	0	0
Eg-m90	165.1719263	6.55883397	Eg-p38	172.3818119	1.63970849 3
Eg-m91	12.45343889	4.919125478	Eg-p39	50.4691997	0
Eg-m92	0.655444152	0	Eg-p40	106.1819526	3.27941698 5
Eg-m93	0	0	Eg-p41	15.07521549	0
Eg-m94	26.21776608	27.87504437	Eg-p42	47.19197894	0
Eg-m95	541.3968695	29.51475287	Eg-p43	139.6096044	38.5331495 8
Eg-m96	150.7521549	4.099271232	Eg-p44	7.209885671	0
Eg-m97	1.966332456	0	Eg-p45	90.45129296	2.45956273 9
Eg-m98	148.1303783	9.018396709	<b>Orphan snoRNAs</b>		
Eg-m99	9.831662278	1.639708493	Eg-m-Orphan1	7.209885671	0
Eg-m100	47.19197894	27.05519013	Eg-m-Orphan2	1.966332456	0
Eg-m101	12.45343889	3.279416985	Eg-m-Orphan3	40.63753742	22.9559189
Eg-m102	49.81375554	12.29781369	Eg-m-Orphan4	68.1661918	0.81985424 6
Eg-m103	12.45343889	7.378688217	Eg-m-Orphan5	81.27507484	1.63970849 3
Eg-m104	1.310888304	0	Eg-m-Orphan6	19.66332456	0
Eg-m105	7.209885671	0	Eg-m-Orphan7	1.310888304	1.63970849 3
Eg-m106	57.67908537	12.29781369	Eg-m-Orphan8	11.79799473	0
Eg-m107	76.03152162	1.639708493	Eg-m-Orphan9	0.655444152	0
Eg-m108	31.46131929	55.75008875	Eg-m-Orphan10	11.79799473	6.55883397
Eg-m109	903.8574855	13.11766794	Eg-m-Orphan11	1.310888304	0
Eg-m110	16.3861038	5.738979724	Eg-m-Orphan12	3.277220759	0
Eg-m111	41.94842572	1.639708493	Eg-m-Orphan13*	0	0
Eg-m112	79.30874238	14.75737643	Eg-m-Orphan14	0.655444152	0

Eg-m113	210.3975728	20.49635616	Eg-m-Orphan15	0.655444152	0
Eg-m114	33.42765175	36.89344108	Eg-m-Orphan16*	0	0
Eg-m115	47.84742309	12.29781369	Eg-m-Orphan17*	0	0
Eg-m116	32.77220759	2.459562739	Eg-p-Orphan1	0.655444152	0
Eg-m117	3.277220759	0	Eg-p-Orphan2	10.48710643	0
Eg-m118	7.865329823	0	Eg-p-Orphan3	5.243553215	0.81985424 6
Eg-m119	55.05730876	0	Eg-p-Orphan4	0.655444152	0
Eg-m120	9.831662278	1.639708493	Eg-p-Orphan5	0.655444152	0
Eg-m121	23.59598947	2.459562739	Eg-p-Orphan6	19.00788041	0
Eg-m122	0.655444152	0	Eg-p-Orphan7	399.8209327	4.09927123 2
Eg-m123	1.966332456	0.819854246	Eg-p-Orphan8	178.9362535	4.09927123 2
Eg-m124	53.0909763	109.860469	<b>Additional RNAs</b>		
Eg-m125	3.932664911	0.819854246	U2 snRNA	530.909763	15220.5940 8
Eg-m126	4.588109063	4.919125478	U3 snoRNA	35.3939842	43934.3493 5
Eg-m127	22.94054532	2.459562739	All rRNAs	93965.7845	126064.068 3
Eg-m128	20.31876871	27.05519013			
Eg-m129	203.8431312	22.9559189			

**Table A.3.4. Characteristics of box AGAUGN snoRNAs identified in *Euglena gracilis***

Box AGAUG N RNA	Size (nt)	Target site	Target match*	Distance to Ψ pocket (nt)	Basal Stem (nt)	Apical Stem (nt)	Apical loop (nt)
Eg-p5	66	LSU34 51	4/0 + 4/0	15	6	10; 0+1 nt bulge	8
Eg-p6	64	SSU10 5	5/0 + 8/1	15	5	11; 1 m.m.	7
Eg-p7	70	LSU33 32	5/0 + 6/0	18	8	9; 1 m.m.	10-12
Eg-p8	60	LSU13 65	4/0 + 6/0	14	6	11; 0+1 nt bulge	6
Eg-p9	69	LSU35 03	5/0 + 5/0	15	6	15; two 0+1 nt bulges	7
Eg-p10	68	LSU28 42	4/0 + 6/0	15	9; 1 nt bulge	11; 0+1 nt bulge	8
Eg-p11	72	SSU54 4	4/0 + 9/0	16	6	14; 1 m.m. 0+1 nt bulge	5
Eg-p12	66	LSU29 04	4/0 + 4/0	13	7	10; 0+1 nt bulge	12
Eg-p13	67	LSU15 68	4/0 + 4/0	15	9; 1 nt bulge	13; 0+2 nt bulge	7
Eg-p14	71	LSU32 35	5/0 + 5/0	15	7	14-16; 2 m.m. 0+1 nt bulge	4 - 6
Eg-p15	72	LSU27 52	5/0 + 5/0	15	8	12; 0+1 nt bulge & 0+4 nt bulge	4
Eg-p16	67	SSU15 92	4/0 + 7/0	15	6	11; 2+3 nt bulge	10
Eg-p17	66	LSU37 01	5/0 + 5/0	16	6	12; 1 m.m.	6
Eg-p18	67	SSU13 93	5/0 + 6/0	15	7	11	8
Eg-p19	63	LSU39 53	4/0 + 8/0	14	6	12; 1+0 nt bulge	6
Eg-p20	70	LSU29 15	4/0 + 8/0	15	6	14; 1+0 nt bulge	8
Eg-p21	67	LSU29 14	6/0 + 4/0	15	7	12; 1+0 nt bulge	7
Eg-p22	65	LSU10 23	3/0 + 8/0	16	6	13; 1+2 nt bulge	4
Eg-p23	62	LSU12 66	6/0 + 6/0	15	6	13; 1+2 nt bulge	3
Eg-p24	64	LSU28 0	4/0 + 9/0	16	6	10; 2+2 nt bulge	6
Eg-p25	64	LSU12 35	4/0 + 9/0	15	7	11; 0+1 nt bulge	5
Eg-p26	66	LSU36 97	6/0 + 6/1	15	6	15; 1 m.m. 0+2 nt bulge	4
Eg-p27	66	LSU35 68	7/0 + 4/1	15	7	11	10
Eg-p28	65	LSU30 42	4/0 + 8/2	16	8; 0+1 nt bulge	8; 1 m.m.	12
Eg-p29	66	SU166 0	3/0 + 7/0	14	6	10; 2 m.m.	12
Eg-p30	68	SSU20 65	4/0 + 5/0	17	7; 0+1 nt bulge	14; 0+1 nt bulge, 1 m.m.	5
Eg-p31	67	LSU21 19	4/0 + 7/0	14	5	10; 1 m.m.	12

Eg-p32	69	LSU35 91	5/0 + 5/1	16	6	13; 1+0 nt bulge	8
Eg-p33	70	LSU29 43	5/0 + 5/0	15	8	10	14
Eg-p34	64	SSU21 16	4/0 + 6/0	16	6	13; 0+1 nt bulge	4
Eg-p35	67	LSU35 42	8/1 + 7/0	14	4	11	10
Eg-p36	68	SSU10 68	3/0 + 5/0	15	6	7; 0+1 nt bulge	17
Eg-p37	65	LSU48 0	7/2 + 6/0	15	6	12; 0+1 nt bulge	5
Eg-p38	65	SSU23 05	7/0 + 4/0	15	6	12; 0+1 nt bulge	5
Eg-p39	64	LSU14 07	7/0 + 6/0	14	5	12; 1+0 & 0+1 nt bulges	7
Eg-p40	62	LSU56 7	4/0 + 6/1	15	8; 1+0 nt bulge	8	10
Eg-p41	63	LSU19 26	7/1 + 7/0	14	5	11; 2 1+1 nt bulges	7
Eg-p42	66	SSU89	4/0 + 6/0	15	7	12	7
Eg-p43	67	LSU27 46	5/0 + 7/0	15	8	12; 2 0+1 nt bulges	7
Eg-p44	60	SSU16 24	3/0 + 4/0	16	6	10; 2+1 nt bulge	5
Eg-p45	68	LSU28 37	4/0 + 5/0	14	8; 2 m.m.	10; 0+1 nt bulge	15

\* 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 S3 for further clarification.  
m.m. = mismatch

**Table A.3.5. Identity of the first nucleotide upstream of the basal stem of AGAUGN snoRNAs in *Euglena gracilis*.** Invariant nucleotides have this identity in all isoforms of that RNA species at this position. Variable nucleotides indicate that isoforms of a single RNA species have one of two nucleotides at this position. Does not include orphan RNAs.

Invariant	Number of Sequences
A	16
C	15
G	5
U	5

Variable	
A or C	2
C or U	1



**Table A.3.6. PCR conditions used for amplification of cDNA during library construction.**

Cycle	Temperature	Duration	Reaction Components
1 cycle	98°C	2 min	1 X Phusion HF buffer (Thermo Scientific), 200 µM dNTPs, 0.5 M each oligonucleotide, 5 µL (of 20 µL reaction) of cDNA, 1 U of Phusion Taq Polymerase (Thermo Scientific)
35 cycles	98°C	10 s	
	60°C	15 s	
	72°C	30 s	
1 cycle	72°C	7 min	

## Appendix 4 – Table of oligonucleotides used in this thesis:

**Table A.4.1. Oligonucleotide primers used in experimental Chapters 2-4.**

Oligo Name	Sequence (5' to 3')	Description
DM1	GCTGTAATACGACTCACTATAGGCA AGGTGCGTGCTCCCTCGGTG	Forward primer for <i>G. lamblia</i> U4 5' K-turn mutant for GA to GC in NC stem <i>in vitro</i> transcription
DM2	CAAGGCAGCACCGAGGGA	Reverse primer for <i>G. lamblia</i> U4 5' K-turn mutant for GA to GC in NC stem <i>in vitro</i> transcription
DM3	GGCGCATATGTCTGCCCCAAACCCA AAGGCTTTC	Forward primer for cloning of Yeast Snu13p CDS into pET28a expression vector
DM4	GGACGAATTCTTAAATTAATAATGT TTCAATCTTGTC	Reverse primer for cloning of Yeast Snu13p CDS into pET28a expression vector
DM7	GCTGTAATACGACTCACTATAGGGA GCCTCTGTGCTTATGCC	Forward primer for <i>in vitro</i> transcription of K-turn containing stem of <i>G. lamblia</i> RNase P
DM8	GAGCACTCTCAGTCGGGCAAGG	Reverse primer for <i>in vitro</i> transcription of K-turn containing stem of <i>G. lamblia</i> RNase P
DM9	GCTGTAATACGACTCACTATAGGCA AGGTGCGTGACCCTCGGG	Forward primer for <i>in vitro</i> transcription of <i>G. lamblia</i> U4 5' K-turn lacking UU pair mutant. (Does not have 5' most C)
DM10	CAAGGCATCCCGAGGGTCACGC	Reverse primer for <i>in vitro</i> transcription of <i>G. lamblia</i> U4 5' K-turn lacking UU pair mutant. (Does not have 3' most A)
DM15	CGGACCTCGAGTTACTCGTCCTCGA GCCAG	Reverse primer for cloning <i>G. lamblia</i> Rsa1p/NUFIP homolog into expression vector
DM20	CGGACGGTACCCTCGTCCTCGAGCC AGTGGC	Reverse primer for cloning <i>G. lamblia</i> Rsa1p/NUFIP homolog into pETDuet expression vector
DM47	GGCGGGATCCATGGGTACAGACTAT CGAAACAGCGGGCGC	Forward primer for cloning of <i>G. lamblia</i> fibrillarin CDS into pAN/pAC <i>G. lamblia</i> expression vectors
DM49	GGCGGGATCCATGGGTACAGACTAT CGAAACAGCGGGCGC	Forward primer for cloning of <i>G. lamblia</i> $\alpha$ -Snu13p CDS into pAN/pAC <i>G. lamblia</i> expression vector
DM50	GGCGGGATCCATGCAAATTGACCCC AGAGCCATTCCG	Reverse primer for cloning of <i>G. lamblia</i> $\alpha$ -Snu13p CDS into pAC <i>G. lamblia</i> expression vector
DM51	GGCGGGATCCATGCCAGATGCACGC GCTGTTC	Forward primer for cloning of <i>G. lamblia</i> $\beta$ -Snu13p CDS into pAN/pAC <i>G. lamblia</i> expression vector
DM52	CATAGCGGCCGCTGGTGTCTAGCT CCGTAAGGAATATGTTAGC	Reverse primer for cloning of <i>G. lamblia</i> $\beta$ -Snu13p CDS into pAC <i>G. lamblia</i> expression vector

Oligo Name	Sequence (5' to 3')	Description
DM53	GGCGGAATTCTGCGCTGCCTTGACA CGGAATCTCCCCAC	Reverse primer for cloning of <i>G. lamblia</i> fibrillarin CDS into pAN <i>G. lamblia</i> expression vector (Paired with DM47 forward primer)
DM54	GGCGGAATTCAATACCAGTGCCTCC ACTTTCCCTAC	Reverse primer for cloning of <i>G. lamblia</i> $\alpha$ -Snul3p CDS into pAN <i>G. lamblia</i> expression vector (Paired with DM49 forward primer)
DM55	GGCGGAATTCTGGTGTCTAGCTCC GTAAGGAATATGTTTAGC	Reverse primer for cloning of <i>G. lamblia</i> $\beta$ -Snul3p CDS into pAN <i>G. lamblia</i> expression vector (Paired with DM51 forward primer)
DM61	GCTGTAATACGACTCACTATAGGAA TACGCATATCAGTGAGGATTCGTCC GAG	Forward primer for <i>in vitro</i> transcription of <i>S. cerevisiae</i> U4 5' K-turn
DM62	AAACACAATCTCGGACGAATCC	Reverse primer for <i>in vitro</i> transcription of <i>S. cerevisiae</i> U4 5' K-turn
DM63	GCTGTAATACGACTCACTATAGGTC AAATGATGTAATAACATATTTGCTA C	Forward primer for <i>in vitro</i> transcription of <i>S. cerevisiae</i> snR24 C/D snoRNA
DM64	TTCATCAGAGATCTTGGTGATAATT GG	Reverse primer for <i>in vitro</i> transcription of <i>S. cerevisiae</i> snR24 C/D snoRNA
DM95	GGCGGGATCCATGCCAATCATCGTT AAGGGTGC	Forward primer for cloning of <i>G. lamblia</i> Ucp1 into pAN <i>G. lamblia</i> expression vector
DM96	CACGGAATTCTGCCTTGCCTCTTCG AGTGTAC	Reverse primer for cloning of <i>G. lamblia</i> Ucp1 CDS into pAN <i>G. lamblia</i> expression vector
DM98	GGAGCATATGCCAATCATCGTTAAG GGTGC	Forward primer for cloning of <i>G. lamblia</i> Ucp1 CDS into pET28a expression vector
DM99	GGACGAATTCTCACCTTGCCTCTTC GAGTGTAC	Reverse primer for cloning of <i>G. lamblia</i> Ucp1 CDS into pET28a expression vector
DM100	GGCGGGATCCATGCCGATCATTATT AAGGGTGTGAAGC	Forward primer for cloning of <i>G. lamblia</i> Ucp2 CDS into pAN <i>G. lamblia</i> expression vector
DM101	CACGGAATTCTGGTTCACATCCGAC ACCGTAGTGC	Reverse primer for cloning of <i>G. lamblia</i> Ucp2 CDS into pAN <i>G. lamblia</i> expression vector
DM103	GGAGCATATGCCGATCATTATTAAG GGTGTGAAGC	Forward primer for cloning of <i>G. lamblia</i> Ucp2 CDS into pET28a expression vector
DM104	GGACGAATTCTCAGTTCACATCCGA CACCGTAGTGC	Reverse primer for cloning of <i>G. lamblia</i> Ucp2 CDS into pET28a expression vector
DM105	GGCGGGATCCGATGCCAATCATCGT TAAGGGTGC	Forward primer for cloning of <i>G. lamblia</i> Ucp1 CDS into multiple cloning site 1 of pETDuet vector
DM106	CACAGCGGCCGCTCACCTTGCCTCT TCGAGTGTAC	Reverse primer for cloning of <i>G. lamblia</i> Ucp1 CDS into multiple cloning site 1 of pETDuet vector

Oligo Name	Sequence (5' to 3')	Description
DM107	CACGCTCGAGTCAGTTCACATCCGA CACCGTAGTGC	Reverse primer for cloning of <i>G. lamblia</i> Ucp2 CDS into multiple cloning site of pETDuet vector for use with Ucp2 pET28a forward primer (DM103)
DM108	GGCGGGATCCGATGCCGATCATTAT TAAGGGTGTGAAGC	Forward primer for cloning <i>G. lamblia</i> Ucp2 CDS into multiple cloning site 1 of pETDuet
DM109	CACGGCGGCCGCTCAGTTCACATCC GACACCGTAGTGC	Reverse primer for cloning <i>G. lamblia</i> Ucp2 CDS into multiple cloning site 1 of pETDuet
DM110	CACGCTCGAGTCACCTTGCTCTTC GAGTGTCAC	Reverse primer for cloning <i>G. lamblia</i> Ucp1 CDS into multiple cloning site 2 of pETDuet. For use with Ucp1 pET28a forward primer (DM98)
DM129	TTCAGTTCGATGCGCCCAGG	Forward primer for PCR in <i>G. lamblia</i> RNase P RNA 3' RACE (used with oP94 reverse primer)
oP94	AATAAAGCGGCCGCGGATCCAATTT TTTTTTTTTTTTT(A/C/G)	Reverse primer used for RT and PCR steps of RNase P RNA 3' RACE
oAH21	GGCCCGATTAATATGCAAATTGACC CCAGAGCC	Forward primer for cloning of <i>G. lamblia</i> $\alpha$ -Snu13p CDS into pET28a expression vector (Used with oAR314 reverse primer)
oAR314	GGCGAAGCTTTTCATACCAGTGCCTC CACTTTCCC	Reverse primer for cloning <i>G. lamblia</i> $\alpha$ -Snu13p CDS into pET28a expression vector (Used in oAH21)
oAH24	GGCGCATATGGGTACAGACTATCGA AACAGCGGG	Forward primer for cloning of <i>G. lamblia</i> fibrillarin CDS into pET28a expression vector
oAH25	GGACGAATTCTCACGCTGCCTTGAC ACGGAATCTCC	Reverse primer for cloning of <i>G. lamblia</i> fibrillarin CDS into pET28a expression vector
oAH30	GCGCATATGTCTAGCGAGCTCATAC CGGC	Forward primer for cloning of <i>G. lamblia</i> putative Nufip domain containing protein into pET28a expression vector
oAH31	CGGACGAATTCTTACTCGTCCTCGA GCCAGTTGC	Reverse primer for cloning of <i>G. lamblia</i> putative Nufip domain containing protein into pET28a expression vector
oAH40	GCTGTAATACGACTCACTATAGGAG CTGTGATGACAGGTTCTTG	Forward primer for <i>in vitro</i> transcription of <i>G. lamblia</i> box C/D snoRNA GlsR5.
oAH41	GAGCTCAGCCGGTTGGCTTGG	Reverse primer for <i>in vitro</i> transcription of <i>G. lamblia</i> box C/D snoRNA GlsR5.
oAH46	GCTGTAATACGACTCACTATAGGTA GCAACCCGTGATTTGCAACG	Forward primer for <i>in vitro</i> transcription of <i>G. lamblia</i> box C/D snoRNA GlsR9.
oAH47	AACCTCATGTGTGCTTTCAC	Reverse primer for <i>in vitro</i> transcription of <i>G. lamblia</i> box C/D snoRNA GlsR9.
oAH206	GCTGTAATACGACTCACTATAGGCA AGGTGCGTGATCCCTCGGTG	Forward primer for <i>in vitro</i> transcription of <i>G. lamblia</i> U4 snRNA 5' K-turn.
oAH207	CAAGGCATCACCGAGGGA	Reverse primer for <i>in vitro</i> transcription of <i>G. lamblia</i> U4 snRNA 5' K-turn.

Oligo Name	Sequence (5' to 3')	Description
GM01	CATAGGATCCATGACACAGAACTAC CGCATAAAGCAGCTG	Forward primer for cloning <i>G. lamblia</i> Ucp3 CDS into pAN (Used with GM02)
GM02	CATAGAATTCAAGTACTTCTTGACG GTTTCAGAGCTCCCTGA	Reverse primer for cloning <i>G. lamblia</i> Ucp3 CDS into pAN (Used with GM01)
DM83	GGCGGGATCCATGTACCTGTTATAT GAGGCCGCG	Forward primer for cloning <i>G. lamblia</i> Nop58 CDS into pAN (Used with DM84)
DM84	CACGGAATTCTGCTTGAGTGCCTC TTGTTCTTCTTATCG	Reverse primer for cloning <i>G. lamblia</i> Nop58 CDS into pAN (Used with DM83)
IllMulti	AATGATACGGCGACCACCGAGATCT ACACGTTTCAGAGTTCTACAGTCCGA CGATC	Reverse primer used for construction of all Illumina RNA-seq libraries
IlluBar1	CAAGCAGAAGACGGCATAACGAGAT CGTGATGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCT	Forward primer containing barcode for multiplexing for Illumina $\beta$ -Snu13p input RNA library replicate 1
IlluBar2	CAAGCAGAAGACGGCATAACGAGAT ACATCGGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCT	Forward primer containing barcode for multiplexing for Illumina $\beta$ -Snu13p co-precipitated RNA library replicate 1
IlluBar3	CAAGCAGAAGACGGCATAACGAGAT GCCTAAGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCT	Forward primer containing barcode for multiplexing for Illumina $\beta$ -Snu13p input RNA library replicate 2
IlluBar4	CAAGCAGAAGACGGCATAACGAGAT TGGTCAGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCT	Forward primer containing barcode for multiplexing for Illumina $\beta$ -Snu13p co-precipitated RNA library replicate 2
IlluBar5	CAAGCAGAAGACGGCATAACGAGAT CACTGTGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCT	Forward primer containing barcode for multiplexing for Illumina $\beta$ -Snu13p input RNA library replicate 3
IlluBar6	CAAGCAGAAGACGGCATAACGAGAT ATTGGCGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCT	Forward primer containing barcode for multiplexing for Illumina $\beta$ -Snu13p co-precipitated RNA library replicate 3
IlluBar7	CAAGCAGAAGACGGCATAACGAGAT GATCTGGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCT	Forward primer containing barcode for multiplexing for Illumina $\alpha$ -Snu13p input RNA library replicate 1
IlluBar8	CAAGCAGAAGACGGCATAACGAGAT TCAAGTGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCT	Forward primer containing barcode for multiplexing for Illumina $\alpha$ -Snu13p co-precipitated library replicate 1
IlluBar9	CAAGCAGAAGACGGCATAACGAGAT CTGATCGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCT	Forward primer containing barcode for multiplexing for Illumina $\alpha$ -Snu13p input RNA library replicate 2
IlluBar10	CAAGCAGAAGACGGCATAACGAGAT AAGCTAGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCT	Forward primer containing barcode for multiplexing for Illumina $\alpha$ -Snu13p co-precipitated library replicate 2
IlluBar11	CAAGCAGAAGACGGCATAACGAGAT GTAGCCGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCT	Forward primer containing barcode for multiplexing for Illumina $\alpha$ -Snu13p input RNA library replicate 3
IlluBar12	CAAGCAGAAGACGGCATAACGAGAT TACAAGGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCT	Forward primer containing barcode for multiplexing for Illumina $\alpha$ -Snu13p co-precipitated library replicate 3

Oligo Name	Sequence (5' to 3')	Description
R1R DNA	/5Phos/GATCGTCGGACTGTAGAACTC TGAACGTGTAG/3SpC3/	Adapter for library construction using TGIRT™ -III
R2R DNA	GTGACTGGAGTTCAGACGTGTGCTC TTCCGATCTTN	Adapter added during template switching for library construction using TGIRT™-III

