2014

Spliceosomal intron and spliceosome evolution in Giardia lamblia and other diplomonads

Hudson, Andrew J.

Lethbridge, Alta. : University of Lethbridge, Dept. of Biological Sciences

SPLICEOSOMAL INTRON AND SPLICEOSOME EVOLUTION IN *GIARDIA LAMBLIA* AND OTHER DIPLOMONADS

ANDREW JOHN HUDSON
B.Sc. University of Lethbridge, 2009

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

DOCTOR OF PHILOSOPHY

IN

BIOMOLECULAR SCIENCE

Biological Sciences
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

PREPARATION OF THESIS

ANDREW JOHN HUDSON

Date of Defence: December 8, 2014

Dr. Anthony (Tony) Russell Co-supervisor	Assistant Professor	Ph.D.
Dr. Cameron Goater Co-supervisor	Associate Professor	Ph.D.
Dr. Roy Golsteyn Thesis Examination Committee Member	Associate Professor	Ph.D.
Dr. Ute Wieden-Kothe Thesis Examination Committee Member	Associate Professor	Ph.D.
Dr. Nehalkumar Thakor Internal Examiner	Assistant Professor	Ph.D.
Dr. Staffan Svärd External Examiner Uppsala University Uppsala, Sweden	Professor	Ph.D.
Dr. James Thomas Chair, Thesis Examination Committee	Professor	Ph.D.

Dedication

This thesis is dedicated to Becky, Audrey and my parents Ann and Rob…

…for all your support and unconditional love.

Abstract

Spliceosomal introns interrupt protein coding genes in all characterized eukaryotic nuclear genomes and are removed by a large RNA-protein complex termed the spliceosome. Diplomonads are diverse unicellular eukaryotes that display compact genomes with few spliceosomal introns. My thesis objectives were to explore spliceosomal intron and spliceosome diversity as well as RNA processing mechanisms in the diplomonads *Giardia lamblia* and *Spironucleus spp*. Surprisingly, *G. lamblia* was found to contain a proportionally large number of fragmented spliceosomal introns that are spliced *in trans* from separate pre-mRNA molecules. Next, both evolutionarily divergent and conventional spliceosomal small nuclear RNAs were identified in *G. lamblia* and *Spironucleus spp.* and an RNA 3′ end motif was determined to be involved in processing of both non-coding RNAs and *trans*-introns in *G. lamblia*. These findings shed light on spliceosome and spliceosomal intron evolution in eukaryotes undergoing severe genomic reduction and potentially complete loss of their spliceosomal introns.

Acknowledgements

Firstly, I would like to express my most sincere gratitude to my supervisor and mentor Dr. Tony Russell for the opportunity to perform studies in his laboratory and allowing me independence in choosing my research directions. Whatever success I have had, I link closely with the guidance and encouragement I have received from Dr. Russell over the years. Truly, I will never forget all of the stimulating 'what if' conversations we have had regarding RNA-protein evolution and I look forward to more of these in the future.

I would also like to thank the other members of my Ph. D. supervisory committee: Drs. Ute Kothe, Roy Golsteyn and Cam Goater who have so patiently provided me with critical advice and guidance through the years of my graduate studies. I am also extremely grateful (and honored) that one of world's most prolific *Giardia* researchers, Professor Staffan Svärd, was willing to travel from Uppsala, Sweden to serve as my external examiner – particularly on such short notice! Indeed, much of my research project has relied heavily on diplomonad genomic and transcriptomic databases which were generated in a large part by Professor Svärd's group and for that, I cannot thank you enough. I also thank Dr. Nehal Thakor for accepting to be my internal-external examiner and Dr. Jim Thomas for serving as my Ph. D. defence chair.

I have left my lab mates Ashley and David late in this list, however, I cannot express how thankful (and lucky) I feel to have been able to work with you over the years. In addition to your useful suggestions, you have made all of the challenges associated with pursuing a graduate degree more bearable and the successes even more joyous! I will forever cherish our 'Lab-opoly' nights and the countless (and ridiculous) puns spouted off (admittedly mostly by yours' truly) in the Russell Lab. However, I also go forward with a

sense of excitement for both of you: David, I see a promising young research scientist and Ashley, you will soon be an instructor at RDC!

In addition to my 'core' lab mates, I also want to thank the many Russell Lab undergraduate students who have contributed to my success (and many who have become my friends): Dave Elniski (the '*E*-pun *coli*' and banjo master), Ashlee Matkin (Lab-opoly extraordinaire and future MD), Kenzie Visser (Witty pun-ner and future MD), John Stimson (You're good at EVERYTHING and future MD), Colleen Chen (a not so 'underwhelming' DVM), Shayne Rybchinski, Ben Vuong (future DVM), Peter Van Herk, Thomas Scott and Ryan Taylor (future MD). I also want to acknowledge members of the Wieden-Kothe and Wieden Labs (particularly Raja, Laura, Evan, Fan, Luke, Harland and Dylan) who have been helpful in providing advice and sharing equipment over the years.

I would also like to acknowledge the sources of financial support provided to me during my graduate studies: Natural Sciences and Engineering Research Council of Canada (NSERC) and the University of Lethbridge, School of Graduate Studies (SGS). I hope you know that your financial support has made a tremendous contribution to my success.

Lastly, I would like to thank my family and friends for supporting me in this long (and selfish) endeavor of pursuing a doctoral degree and giving me nothing but love, support and encouragement throughout my whole life. Mom and Dad, I hope you see this as the 'end' of my studentship and beginning of my career in science! Becky, this work is as much your accomplishment as it is mine. Joey, thank you for everything! Chad and Sharon, thanks for all your encouragement and inspiration. Christin and Ryan, thank you for your friendship and support over the years. There are also so many others who I would like to thank (e.g. previous mentors: Dr. Kevin Floate and Paul Coglin) but don't have the space and thus I can only say – thank you all for your support!

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Ψ	Pseudouridine
AS	Alternative splicing
ATP	Adenosine 5′ triphosphate
ATPase	Adenosine triphosphatase
BLAST	Basic logical alignment search tool
bp	Base pair
BP	Branch point
cDNA	Complementary DNA
Ci	Curie
CIP	Calf intestinal phosphatase
CM	Co-variation model
Da	Dalton
DB	Database
DHC	Dynein Heavy Chain
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dNTP	Deoxynucleotide 5′ triphosphate
DTT	Dithiothreitol
E	Expect value
EBS	Exon-binding sequence
EDTA	Ethylenediaminetetraacetic acid
EST	Expressed sequence tag
Hsp90	Heat shock protein 90
HEPES	N-2-Hydroxyethylpiperazine-N'-2-Ethanesulfonic acid
IBS	Intron-binding sequence
IEP	Intron-encoded protein
IL	Internal loop
ISL	Intramolecular stem-loop
k-turn	Kink-turn
LECA	Last eukaryotic common ancestor
LSm	Sm-like
mRNA	Messenger RNA
MRP	Mitochondrial RNA processing
ncRNA	Non-coding RNA
NEB	New England Biolabs
nt	Nucleotide
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PAP	Poly (A) polymerase
PCR	Polymerase chain reaction
PNK	Polynucleotide kinase
pol	polymerase (RNA)
pre-RNA	Precursor RNA
RACE	Rapid amplification of cDNA ends
RLM	RNA linker mediated

List of Abbreviations (Continued)

RNA	Ribonucleic acid
RNase	Ribonuclease
RNP	Ribonucleoprotein particle
rRNA	Ribosomal RNA
RP	Ribosomal protein
RT	Reverse transcriptase
SDS	Sodium dodecyl sulphate
SL	Spliced leader OR Stem-loop
snRNA	Small nuclear RNA
snRNP	Small nuclear ribonucleoprotien particle
snoRNA	Small nucleolar RNA
SS	Splice site
SSC	Saline-sodium citrate
TAP	Tobacco acid pyrophosphatase
Taq	*Thermus aquaticus*
TBE	Tris/borate/EDTA
TMG	2,2,7-trimethylguanosine
Tris	Tris(hydroxymethyl)aminomethane
tRNA	Transfer RNA
U-snRNA	Uridine-rich snRNA
U2AS	U2 snRNP auxiliary factor
UTR	Untranslated region

Chapter 1: Introduction

1.0 Genes in Pieces

Genes are the fundamental units of inheritance, forming the building blocks of genomes that define whole organisms. In most cases genes consist of stretches of deoxyribonucleic acid (DNA), however, some viruses break this rule and have ribonucleic acid (RNA) genes and genomes. By the mid-20th century, a 'central dogma of molecular biology' dictated the flow of hereditary information in two steps: 1) the synthesis of an RNA copy of a portion of a (DNA) gene, referred to as transcription and 2) the conversion of information within the RNA molecule into a specific polypeptide sequence in a process called translation (Crick 1970). While still seeming to hold true in many cases, this dogma would soon be contradicted by the finding that many genes do not encode protein and yield so-called non-coding (nc)RNA products. In the late 1970s, this paradigm was further 'altered' by the finding that some viral protein coding genes contain additional sequences that are transcribed in precursor (pre-)RNAs but then are mysteriously 'spliced' to form final mature messenger (m)RNAs (Berget *et al.* 1977, Chow *et al.* 1977). Although initially thought to be rare occurrences limited to certain viral genes, innumerable cases of 'split genes' (and spliceosomal introns) were soon discovered in eukaryotic protein coding genes. Regions of genes (and corresponding RNA transcripts) present in final mature RNAs became known as 'exons' whereas gene portions which are transcribed but later removed were termed 'introns' (Gilbert 1978).

The finding of 'gene in pieces' was most unexpected. Though at first seeming metabolically wasteful, it was soon discovered that introns may be spliced from pre-mRNA transcripts in different ways (alternative splicing), contributing to increased proteomic diversity from a limited set of nuclear genes (Maki *et al.* 1981). However, pre-mRNA

splicing was decidedly a eukaryote-specific process as no single prokaryote appeared to possess equivalent fragmented protein coding genes or the splicing machinery (spliceosome) to remove them. Since their discovery nearly four decades ago, numerous hypotheses have been posed to explain the enigmatic origins of introns and their potential roles in the evolution of modern gene architectures. Although advances in understanding of intron splicing mechanisms, intron structures and their phylogenetic distributions have shed light on these mysteries, many fundamental questions regarding intron and spliceosome structure, function and evolution remain unresolved.

Part of the key to finding answers to these questions lies in the characterization of spliceosomal introns from phylogenetically diverse organisms. Protists represent most of the eukaryotic diversity and thus are important model organisms for elucidating spliceosomal intron/spliceosome structure-function relationships and reconstructing evolutionary pathways to explain the origins of spliceosomal introns and the spliceosome. In this study, I focused on the study of spliceosomal introns and spliceosomal components from a group of mainly parasitic protists known as diplomonads.

1.1 Introns and their Mechanisms of Splicing

Thus far, at least four main types of widely-dispersed introns have been identified: transfer (t)RNA introns, group I introns, group II introns and spliceosomal introns. A fifth 'nonconventional' intron type is also present in some members of the phylum *Euglenozoa*.

1.1.1 Group II introns

Group II introns are mobile genetic elements found in bacterial genomes as well as plastid and mitochondrial genomes of plants, fungi and some protists (Toro *et al.* 2007, Lambowitz and Zimmerly 2011). Although rare, group II introns have been found in archaea (Rest and Mindell 2003); however, group II introns have not been identified in any eukaryotic nuclear genome to date.

Group II introns are ribozymes (catalytic RNAs) capable of self-splicing from RNA precursors by catalyzing two sequential transesterification reactions (Lambowitz and Zimmerly 2011) (Figure 1.1). In the first step of splicing, the 2′ hydroxyl group of an internal adenosine nucleotide acts as a nucleophile to attack the intron 5′ splice site (SS) (Figure 1.1A, step 1). This breaks the phosphodiester bond at the 5′ SS junction and results in 'branching' of the intron via a 2′-5′ phosphodiester bond between the catalytic adenosine and intron 5′ terminal nucleotide. In the second step, the free 3′ hydroxyl of the upstream exon attacks the scissile bond at the intron 3′ SS, resulting in ligation of the flanking exons and release of an excised lariat intron (Figure 1.1A, step 2 and products).

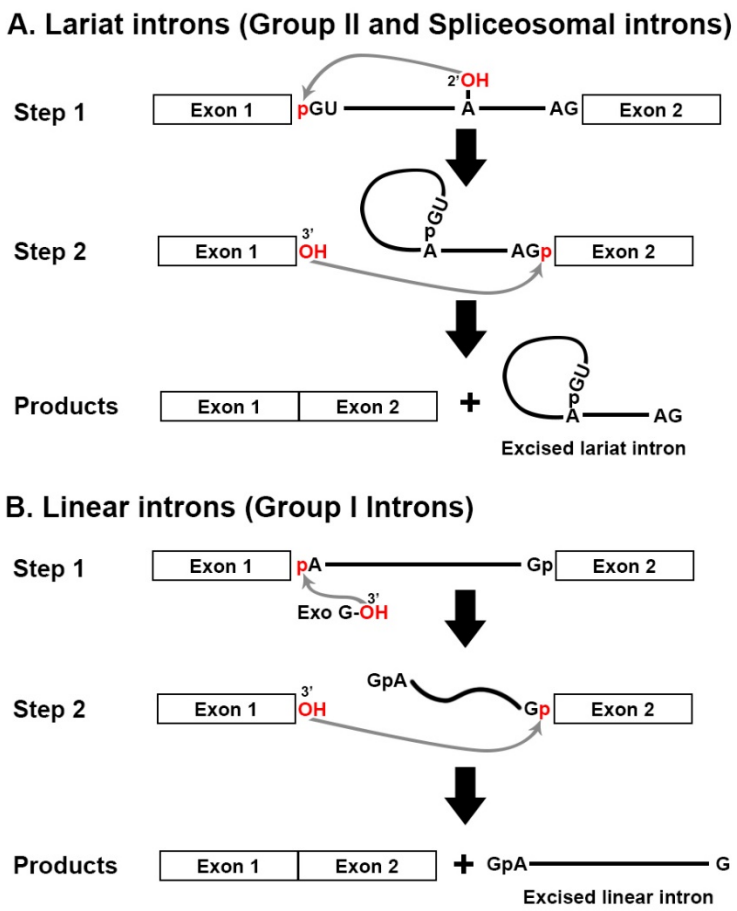


Figure 1.1. Mechanisms of splicing for lariat-forming and linear introns
Schematics for lariat (**A**) and linear (**B**) excised intron splicing mechanisms are shown with chemical groups involved in transesterification reactions in red and grey arrows indicating nucleophilic attack. In part (**A**), dinucleotide boundaries are shown as they occur in most spliceosomal introns. Descriptions of intron splicing mechanisms are provided in the text.

Group II intron RNA ribozymes are 400-800 nt in length and consist of six conserved secondary structural domains (DI – DVI) emanating from a central single-stranded 'wheel' (Lambowitz and Zimmerly 2011) (Figure 1.2). Distant regions of group II intron domains interact to form a conserved tertiary fold, creating the splicing active site. Domains DV and DVI comprise the catalytic core of the ribozyme and DV contains a conserved 'AGC' triad and 'AY' dinucleotide involved in binding two catalytic Mg^{2+} ions (Seetharaman *et al.* 2006, Toor *et al.* 2008) (Figure 1.2). DVI contains the intron branch site, a bulged nucleotide (typically an adenosine, 'A') involved in the first step of splicing.

Domains DI-DIII play important structural roles to position intron elements for catalysis. Notably, DI loop elements EBS1 and EBS2 make long-range tertiary contacts with the 5′ upstream exon (IBS1 and IBS2), whereas the DI loops δ and ε′ interact with the 3′ exon (δ′) and conserved 5′ intron splice site (ε) sequence 'GUGYG', respectively (Figure 1.2). Interestingly, the group II intron 5′ SS sequence is similar to the conserved 'GU' 5′ SS dinucleotide of spliceosomal introns and catalytic domains DV, DVI and regions of DI show similarities to spliceosomal small nuclear (sn)RNAs (Figure 1.2, and see below).

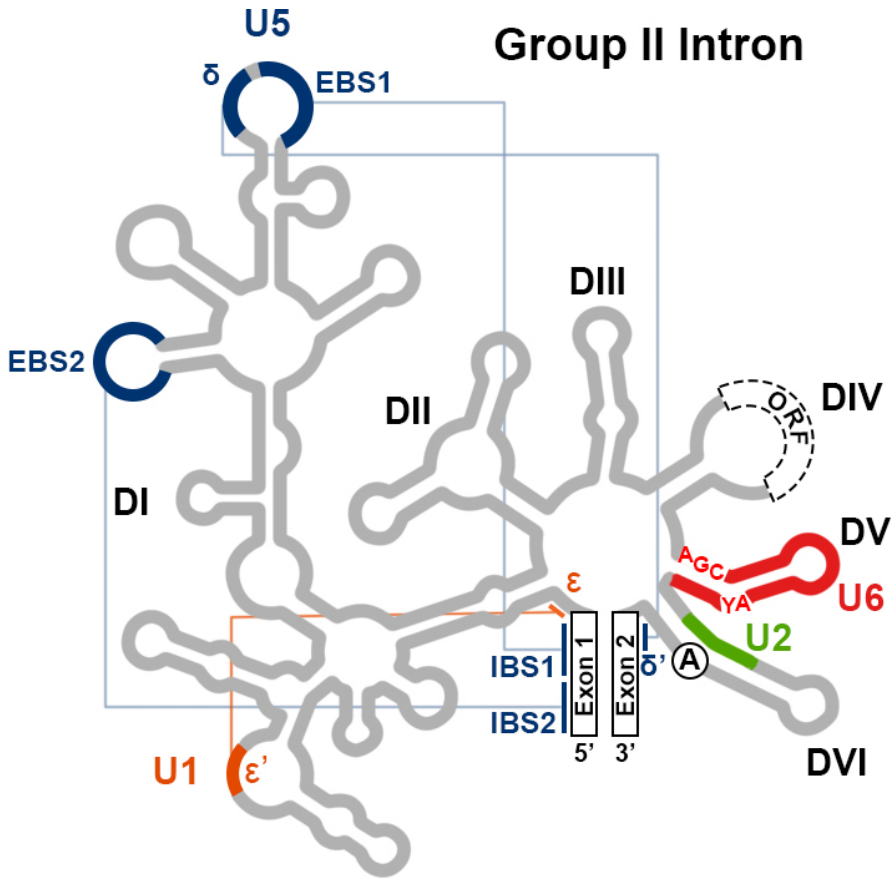


Figure 1.2. Conserved sequences and structures of group II introns.
The diagram depicts conserved domains (DI – DVI) of group II introns. Coloured regions show intron elements with functional similarity to spliceosomal snRNAs (see discussion below). The branch point nucleotide involved in the first step of splicing is circled in domain DVI and a thatched box in DIV denotes the location of intron encoded protein open reading frames (ORFs). Intron regions making tertiary contacts during splicing are indicated by connecting lines.

In vitro, group II introns are capable of catalysis in the absence of proteins, however, intron splicing is generally slow and requires non-physiological conditions (Jarrell *et al.* 1988). Consequently, efficient splicing of group II introns *in vivo* requires protein splicing factors encoded in DIV of the intron (Figure 1.2). Such 'intron encoded proteins' (IEPs) bind a 5′ portion of DIV of their host intron (Watanabe and Lambowitz 2004) and promote group II intron splicing by stabilizing the ribozyme active site (Matsuura *et al.* 1997). The best characterized IEP is LtrA which is encoded by the *Lactococcus lactis* Ll.LtrB intron. LtrA contains four domains: i) reverse transcriptase/fingers palm (RT); ii) X/thumb; iii) DNA binding (DNA); and iv) Endonuclease (En) domains (Blocker *et al.* 2005).

1.1.2 Spliceosomal introns

Spliceosomal introns are unique to eukaryotic nuclear genomes and are absent from all known prokaryotic and eukaryotic organellar genomes. Unlike group II introns, spliceosomal introns are incapable of self-splicing and require a large ribonucleoprotein (RNP) complex termed the 'spliceosome' to remove them from pre-mRNAs (Wahl *et al.* 2009). Nonetheless, spliceosomal introns are excised by the same two step mechanism as group II introns, utilizing an internal branch point (BP) nucleotide during the first step of splicing and yielding an excised intron lariat upon exon ligation (Figure 1.1A).

Spliceosomal introns are usually located within protein coding genes at a position within the open reading frame (ORF); although in some instances, they may be found in either 5′ or 3′ untranslated region (UTR) or in genes specifying ncRNAs (Cabili *et al.* 2011). When found in ORFs, spliceosomal introns may be inserted into any of the three nucleotide positions within a codon. Introns inserted between codons are called 'phase 0' introns, whereas introns inserted between the first and second, or second and third nucleotide positions are termed 'phase 1' and 'phase 2' introns, respectively.

With exception of one nucleomorph genome (Lane *et al.* 2007) and possibly a microsporidian (Keeling *et al.* 2010), all fully-sequenced eukaryotic nuclear genomes have been found to contain spliceosomal introns. However, total intron numbers per genome may vary by several orders of magnitude – from only a few dozen (or less) in some intron-poor genomes (Vanacova *et al.* 2005, Lee *et al.* 2010) to over 200,000 in some intron-rich species (Venter *et al.* 2001). Intron numbers may also vary widely between genes from the same organism. For example, humans genes contain on average ~9 spliceosomal introns, yet some human genes lack any introns while others contain more than ten times this number (Sakharkar *et al.* 2004). Intron length may be equally variable as evidenced by comparing the miniscule 18-21 nt introns from the nucleomorph genome of *Bigelowiella natans* (Gilson *et al.* 2006) to some vertebrate genes which may reach hundreds of kilo-bases in length (Sakharkar *et al.* 2004). Indeed, in mammals, average intron length (~4000 nt) dwarfs that of flanking exons (~150 nt) and human intron sequences account for more than a quarter of total genomic DNA sequences (Lander *et al.* 2001, Gelfman *et al.* 2012).

Spliceosomal introns do not typically show significant secondary structural conservation but they may be identified by short sequence motifs at their 5′ and 3′ termini, as well as an internal sequence known as the branch point (BP) sequence (Figure 1.3). The general eukaryotic consensus is a '/GT' (the slash indicates the exon-intron boundary) at the 5′ splice site (SS) and 'AG/' at the intron 3′ SS (Figure 1.3). The BP sequence contains the catalytic adenosine involved in the first step of splicing and although the position of this element may vary in some eukaryotes, it is usually located 20-40 nt upstream of the intron 3′ SS (Kol *et al.* 2005). Many spliceosomal introns also contain a poly-pyrimidine tract sequence between their BP and 3′ SS elements (Figure 1.3). However this element is missing in spliceosomal introns from some eukaryotes (Bon *et al.* 2003).

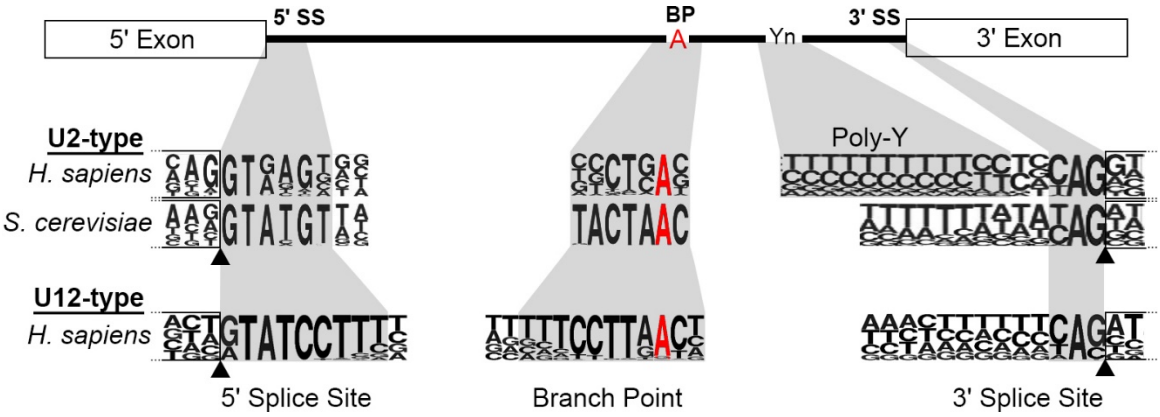


Figure 1.3. Conserved sequence motifs of U2- and U12-type spliceosomal introns. Conserved spliceosomal intron sequence motifs are shown in grey boxes. Exonic sequences flanking introns are in open white boxes and triangles denote exon-intron junctions. Heights of letters in sequence plots indicate their frequency of occurrence. Branch point adenosine nucleotides are in red and the poly-pyrimidine tract in human U2-type introns is indicated (Yn, Poly-Y). Sequence plots for *H. sapiens* and *S. cerevisiae* U2-type introns were adapted from (Lim and Burge 2001) and *H. sapiens* U12-type intron consensus generated using intron sequences retrieved from U12DB (Alioto 2007) and WebLogo software (Crooks *et al.* 2004).

Beyond the prototypical 'GT-AG' splice boundaries, the conservation of extended intron 5′ SS and BP sequences may vary substantially in different eukaryotes (Figure 1.3). For instance, most human spliceosomal introns display the 5′ SS consensus 'GTRAGT' (R is a purine); however, a significant subset of human introns do not conform to this consensus – more than 2500 different 5′ SS sequences have been identified (Roca and Krainer 2009). In contrast, spliceosomal introns from the yeast *Saccharomyces cerevisiae* display a highly conserved 'GTATGT' 5′ SS sequence, with individual introns showing very little deviation from this hexamer (Lopez and Seraphin 2000) (Figure 1.3, cf. *H. sapiens* and *S. cerevisiae* U2-type). The 3′ SS consensus is shorter, and is typically limited to 'YAG' (Y is a pyrimidine) in most eukaryotes.

1.1.3 U2-type and U12-type spliceosomal introns

By the time of early eukaryotic genome sequencing projects, it became evident that two distinct types of spliceosomal introns co-exist in eukaryotes. Although most spliceosomal introns possess ‘GT-AG’ terminal dinucleotides, it was noted that a small subset of human introns bear unusual ‘AT-AC’ intron boundaries and were dubbed ‘ATAC’ introns. Closer examination revealed that most ATAC introns also lacked poly-pyrimidine tracts and contained extended 5′ SS sequences (ATATCCTTT) and BP (TCCTTAACT) sequences which differed from the majority of ‘GT-AG’ spliceosomal introns (Jackson 1991) (Figure 1.3, cf. *H. sapiens* U2-type and U12-type). Moreover, the extended 5′ and BP sequences of ATAC introns showed complementarity to two coincidentally identified small nuclear (sn)RNAs of unknown function: U11 and U12 snRNAs (Montzka and Steitz 1988). It was later confirmed that U11 and U12 snRNAs were two components of a rare ‘minor’ spliceosome required for the splicing of ATAC spliceosomal introns (Hall and Padgett 1994, Tarn and Steitz 1996). The more common spliceosomal introns which are removed by the major/U2 snRNA-dependent spliceosome became known as U2-type introns while introns removed by the minor/U12 snRNA-dependent spliceosome were named U12-type introns. Paradoxically, it was later discovered that terminal ‘GT-AG’ dinucleotides are more common for U12-type introns than ‘AT-AC’ dinucleotides (Dietrich *et al.* 1997) (Figure 1.3). However, the ATAC terminology still applies to some spliceosomal components involved in U12-type intron removal (see section 1.2).

U2-type introns have been identified in all eukaryotic nuclear genomes; however, U12-type introns are considerably rarer and are only found in a smaller subset of eukaryotes. Where they are found, U12-type introns are also significantly less abundant in

a genome than are U2-type introns and account for < 0.5% of all spliceosomal introns in these organisms (Alioto 2007). Nonetheless, U12-type introns and/or U12-dependent spliceosomal components have been identified in representatives from four out of the five major eukaryotic groups including: animals and some fungi (supergroup Opisthokonta), higher plants (Archaeplastida), amoebas (Amoebozoa), and *Phytophthora spp.* (Stramenopiles) (Russell *et al.* 2006, Lopez *et al.* 2008, Bartschat and Samuelsson 2010). These findings indicate U12-type introns and the U12-dependent spliceosome were established early in eukaryotic evolution and may have been present in the last eukaryotic common ancestor (LECA) (Russell *et al.* 2006, Lopez *et al.* 2008). However, some fungi such as ascomycetes (e.g. *S. cerevisiae*) and the nematode *Caenorhabditis elegans* lack U12-type introns, and the absence of U12-type introns in many other eukaryotes indicate U12-type introns have been lost entirely numerous times during eukaryotic evolution (Bartschat and Samuelsson 2010).

Besides showing differences in splice site sequences, U12-type introns display some other interesting properties with respect to U2-type introns. For instance, U2-type introns display more variable distance between their BP and 3′ SS sequences. Conversely, U12-type introns show strict conservation in the distance between these elements (~10-15 nt) and artificial introduction of insertions or deletions to perturb this distance results in significantly lowered splicing efficiency for U12-type introns (Dietrich *et al.* 2001, Dietrich *et al.* 2005). In humans, splicing of U12-type introns is approximately two fold slower than for U2-type introns and their removal is likely rate-limiting for maturation of pre-mRNAs containing these sequences (Patel *et al.* 2002). However, U12-type intron splicing may be enhanced when U12-type introns are located near U2-type introns, indicating cross-talk

occurs between U2- and U12-dependent spliceosomal components (Lewandowska *et al.* 2004).

1.1.4 Intron structures of 'intron-rich' versus 'intron-poor' eukaryotes

There is a strong inverse correlation between intron number per genome and conservation of intron splicing sequences. Intron-rich eukaryotes (e.g. humans and higher plants) tend to show more degenerate intron 5′ SS and BP sequence motifs in their U2-type introns, whereas, intron-poor species (e.g. *S. cerevisiae* containing ~250 U2-type introns) display extended and more strictly-conserved intron 5′ SS and BP sequences (Irimia *et al.* 2007, Irimia and Roy 2008). The lower information content of U2-type intron splice site and BP sequences in humans is estimated to account for only half of the information required to accurately define intron boundaries (Lim and Burge 2001). Instead, intron-rich eukaryotes contain additional *cis*-sequence elements located within introns or flanking exon sequences which provide the additional information necessary for delineating intron splicing boundaries and play important roles for the selection of splice sites for alternative splicing (see section 1.1.7).

Intron-rich eukaryotes tend to show a more uniform distribution of intron insertion positions in their genes, however, intron-poor eukaryotes show a marked bias for intron insertions towards the 5′ ends of genes (Vanacova *et al.* 2005). Interestingly, numerous distantly-related intron-poor eukaryotes have been identified with total intron numbers approaching zero. It is currently unclear what evolutionary forces ensure the maintenance of a small number of spliceosomal introns, although the retention of specific spliceosomal introns in evolution suggests these introns may have beneficial functions in gene expression. Consistent with this idea, in severely intron-reduced species (< 0.2 introns per gene), the few retained spliceosomal introns are usually found within genes involved in

core metabolic functions (such as ribosomal proteins), suggesting they play important roles in gene regulation (Vanacova *et al.* 2005, Juneau *et al.* 2006, Lee *et al.* 2010).

Finally, intron-poor eukaryotes also show more strict conservation of the distance between intron BP sequence and 3′ SS elements and generally lack poly-pyrimidine tracts in their U2-type introns which are also features of rare U12-type introns.

1.1.5 Other intron types

Group I introns are 250-500 nt self-splicing RNAs found in nuclear, plastid and mitochondrial genes in eukaryotes as well as in the genomes of some bacteria and bacteriophages, but have so far not been identified in archaea (Haugen *et al.* 2005). Group I introns are spliced by two sequential transesterification reactions with several differences in their splicing mechanism as compared to group II and spliceosomal introns including: i) the requirement for an exogenous guanosine nucleotide during the first step of splicing and ii) generation of linear excised intron products (Vicens and Cech 2006) (Figure 1.1B).

tRNA introns are small, 10-150 nt sequences found in archaeal tRNA as well as eukaryotic nuclear tRNA genes (Yoshihisa 2014). tRNA introns are unique with respect to other intron types in their requirement for protein-only enzymes to catalyze their removal (Phizicky and Hopper 2010).

Euglenids are a diverse group of mainly free-living protists belonging to the eukaryotic supergroup Excavata (Adl *et al.* 2012). Although, complete nuclear genomic sequence from any euglenid is currently unavailable, examination of a small collection nuclear genes from members of the genus *Euglena* has revealed the presence of three distinct types of introns: 1) conventional (spliceosomal type) introns, 2) nonconventional introns and 3) intermediate introns. Nonconventional introns do not bear the conserved GT-AG (nor AT-AC) terminal dinucleotides or extended splice sites found in U2- or U12-type

spliceosomal type introns but instead have the ability to form extended terminal base pairs which bring the 5′ and 3′ splice sites into proximity (Tessier *et al.* 1992). Currently, nothing is known about the mechanism(s) involved in nonconventional intron removal. Several *Euglena* introns have been classified as 'intermediate' because they show properties of both conventional and nonconventional introns (e.g. shows a 'GT' 5′ SS dinucleotide and terminal base pairing) (Canaday *et al.* 2001, Russell *et al.* 2005). Whether intermediate introns are removed by the spliceosome or by some nonconventional splicing mechanism (or both) remains to be determined.

1.1.6 Alternative splicing

After the discovery of introns, the finding that some multi-intron containing genes may undergo differential or 'alternative' splicing of their exons was also surprising (Maki *et al.* 1981). As opposed to invariant, 'constitutive' splicing, alternative splicing (AS) refers to the alternate selection of splice sites during pre-mRNA splicing such that different portions of precursor transcript are included or excluded from the final mRNA product (Nilsen and Graveley 2010). By varying intron splice sites, it is possible to increase total proteomic output without necessitating an increased number of genes. For instance, the 24 exons of the *Drosophila Dscam* gene may be alternatively spliced to generate potentially >38,000 different mature mRNA isoforms (Schmucker *et al.* 2000).

Selection of splice sites during AS is highly dynamic and depends on both core (i.e. 5′/3′ SS and BP sequence) and additional *cis* elements that are present in either intronic or exonic regions of an RNA primary transcript (Black 2003, McManus and Graveley 2011). *Cis* elements are bound by *trans*-acting protein splicing factors that may act antagonistically to promote or repress usage of a particular splicing donor/acceptor. More recently, it has become apparent that a 'splicing code' may exist in which certain

combinations of *cis* elements are correlated with particular alternative splicing patterns (Wang and Burge 2008).

AS is prevalent in intron-rich eukaryotes which contain on average several spliceosomal introns per gene. Most strikingly, >95% of human multi-exon genes are predicted to undergo some form of AS in a tissue or developmentally-controlled manner (Pan *et al.* 2008, Wang *et al.* 2008). However, more intron-poor eukaryotes rarely contain more than a single spliceosomal intron per gene and thus, the prevalence of AS is likely low or absent in these organisms (Bon *et al.* 2003).

1.1.7 RNA *trans*-splicing

Thus far, only introns which are spliced from a single contiguous RNA molecule have been discussed. However, some introns (and their flanking exons) are transcribed on multiple precursor RNA molecules which must be joined via *trans*-splicing to yield contiguous spliced RNA products. While the phenomenon of *trans*-splicing is rare, naturally occurring examples of *trans*-splicing have been documented for group I (Burger *et al.* 2009, Nadimi *et al.* 2012), group II (Goldschmidt-clermont *et al.* 1991) and tRNA introns (Randau *et al.* 2005). In cases of group I and group II intron *trans*-splicing, intron functional domains are transcribed on separate precursor RNA molecules which associate post-transcriptionally to reconstitute intron tertiary structures competent for splicing. Most interestingly, *in vitro* studies have revealed that splicing of group II introns lacking important functional domains/subdomains (DIc, DId3 and DV) may be restored by supplying *trans*-acting RNAs containing these elements (Goldschmidt-clermont *et al.* 1991, Suchy and Schmelzer 1991, Hetzer *et al.* 1997).

Eukaryotic nuclear pre-mRNA *trans*-splicing is especially rare, although, several cases have been described in humans (Li *et al.* 1999, Takahara *et al.* 2000), insects (Dorn

et al. 2001, Robertson *et al.* 2007) and a nematode (Fischer *et al.* 2008). Mechanisms of pre-mRNA *trans*-splicing are not well understood, although in at least some cases, base pairing complementarity between separate pre-mRNAs is predicted to mediate association of pre-mRNAs prior to spliceosome-mediated exon ligation (Fischer *et al.* 2008). Despite the rarity of pre-mRNA *trans*-splicing, in trypanosomes, euglenids, dinoflagellates and the nematode *Caenorhabditis elegans*, a large number of pre-mRNAs undergo a specialized form of *trans*-splicing called spliced leader (SL) *trans*-splicing (Liang *et al.* 2003) (Figure 1.4). SL *trans*-splicing involves the spliceosome-mediated addition of a common 'spliced leader' RNA sequence to the 5′ ends of protein coding cistrons (ORFs) embedded within a polycistronic precursor transcript (Figure 1.4). The SL RNA consists of a 5′ exon (leader) moiety containing a nucleotide cap and an intron-like moiety containing a canonical spliceosomal intron 5′ SS sequence. During SL *trans*-splicing, the 5′ SS from the SL RNA is used in conjunction with BP and 3′ SS sequences upstream of protein coding cistrons to fuse leader RNAs to the 5′ ends of each cistron (Liang *et al.* 2003). This simultaneously liberates cistrons and caps their 5′ ends. *Trans*-spliced mRNAs are subsequently polyadenlylated at their 3′ ends, yielding mature mRNAs and Y-shaped spliced intron products (Figure 1.4).

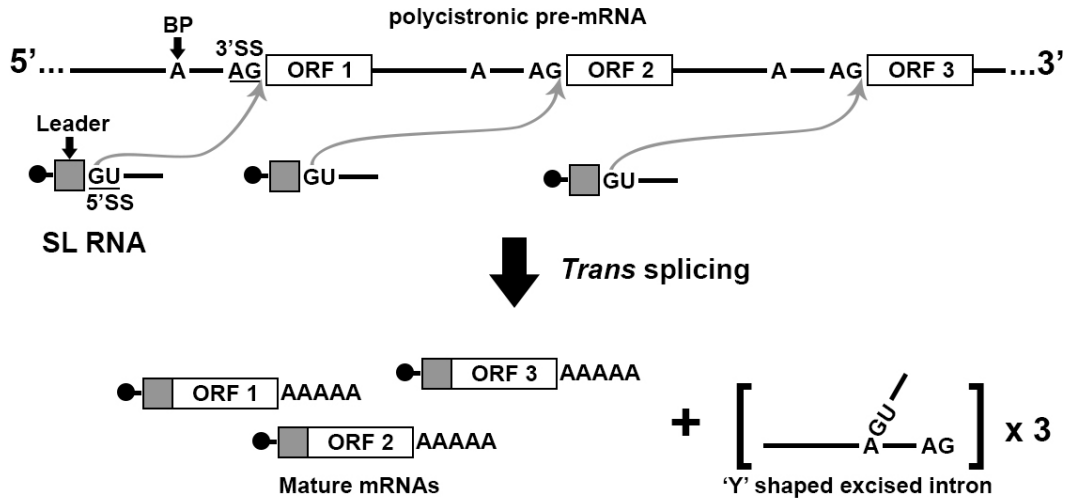


Figure 1.4. Spliced leader (SL) *trans*-splicing.
A polycistronic pre-mRNA containing three separate protein coding cistrons (ORF #) is depicted, showing locations of branch point (BP) and 3′ splice site (SS) sequences. Exon/leader moieties of SL RNAs are represented by grey boxes with SL RNA 5′ nucleotide caps shown as circles. Arrows denote the splicing of SL RNA exons to 5′ ends of pre-mRNA cistrons (ORFs). Polyadenylation signals in the pre-mRNA are not shown.

1.2 The Spliceosome

Excision of spliceosomal introns from eukaryotic pre-mRNA transcripts is catalyzed by spliceosomes – large ribonucleoprotein (RNP) complexes consisting of five uridine-rich snRNAs and dozens to hundreds of stable or transiently associated spliceosomal proteins (Jurica and Moore 2003). With the combined molecular weight of snRNPs approaching 5 MDa in humans (Muller *et al.* 1998), spliceosomes are amongst the largest and most dynamicmacromolecular machines in eukaryotic cells.

1.2.1 Spliceosomal snRNAs

U2-type and U12-type introns are removed by two distinct spliceosomes: the U2-dependent (major) and U12-dependent (minor) spliceosomes, respectively (Patel and Steitz 2003). U2-dependent spliceosomes contain U1, U2, U4, U5 and U6 snRNAs (Figure 1.5, upper panel), whereas, the U12-dependent spliceosome contains a set of distinct but functionally analogous snRNAs: U11, U12, U4atac and U6atac as well as the U5 snRNA found in U2-dependent spliceosomes (Figure 1.5, lower panel). U2-dependent and U12-dependent spliceosomal snRNAs show striking similarities in their secondary structures and in humans, equivalent snRNAs from either spliceosome share 40-50% sequence identity over their lengths (Figure 1.5). The catalytic core of the two spliceosomes also show striking similarities (Figure 1.6), and experiments replacing U6 intramolecular stem-loop (ISL) involved in catalytic metal ion binding (Huppler *et al.* 2002) with the equivalent ISL from U6atac showed splicing activity *in vivo* (Shukla and Padgett 2001). Most interestingly, substitution of U6atac ISL with DV from group II introns is also active in splicing (Shukla and Padgett 2002), further suggesting similarities in splicing mechanism for group II introns and the spliceosome.

There are important differences in snRNA structures from the two spliceosomes (Figures 1.5 and 1.6). For example, U2 snRNAs contain stem-loops (SL) III and SL IV in their 3′ portions while U12 snRNAs possesses an extended SL III but lack SL IV (Figure 1.5) (Sikand and Shukla 2011). Differences in snRNA structures may be important for correct incorporation of snRNPs into their respective spliceosomes and a chimeric U6 snRNA containing the U6atac-specific 3′ SL directs it to the U12-dependent spliceosome (Dietrich *et al.* 2009).

In vivo, snRNAs form intricate RNA-RNA intermolecular base pairing interactions with their intron substrates as well as other snRNAs throughout the splicing reaction. For instance, U1 and U11 snRNAs base pair with U2-type and U12-type intron 5′ splice sites using sequences at their 5′ ends whereas, U2 and U12 snRNAs bind their intron branch point sequences using an internal antisense sequence (Seraphin *et al.* 1988, Frilander and Steitz 1999) (Figures 1.5 and 1.6, red sequences). During early spliceosome assembly, U6 and U4 snRNAs are base-paired extensively and form two extended intermolecular stems I and II (Vankan *et al.* 1992) (Figure 1.5). Upon spliceosome activation, U6 and U4 snRNA base pairs are unwound and U6 snRNA forms new intermolecular base pairs with the intron 5′ SS and U2 snRNA (Burke *et al.* 2012) (Figure 1.6). The equivalent interaction also occurs for U6atac, U4atac and U12 snRNAs from the U12-dependent spliceosome, except that the 5′ end of U12 snRNA is truncated relative to U2 and thus, intermolecular helix II is absent in the U6atac-U12 snRNA complex (Sikand and Shukla 2011) (Figure 1.6).

Despite differences in U2- and U12-dependent spliceosomal snRNAs, class-specific RNA primary and secondary structures are remarkably well conserved across widely diverged eukaryotes (Lopez *et al.* 2008). U6 snRNA show the highest primary sequence conservation and displays ~60% identity between humans and *S. cerevisiae* and 85% between humans and *Arabidopsis thaliana* U6 snRNAs (Bon *et al.* 2003). Other snRNAs show differing levels of primary sequence conservation in different eukaryotes. However, regions of snRNAs involved in intermolecular interactions show especially high sequence conservation and sequence changes usually occur in single stranded regions or else show compensatory changes maintaining conserved snRNA structures (Shukla and Padgett 1999).

The spliceosome has long been suspected to be a ribozyme. This prediction is based on similarities of snRNA structures with functional domains of group II intron ribozymes (Lambowitz and Zimmerly 2011) and a shared two metal ion splicing mechanism. More recently, additional support for this supposition has come from phosphorothioate metal rescue experiments which revealed that U6 snRNA coordinates the two catalytic metal ions required for splicing using nucleotides which have functional equivalents in group II introns (Fica *et al.* 2013).

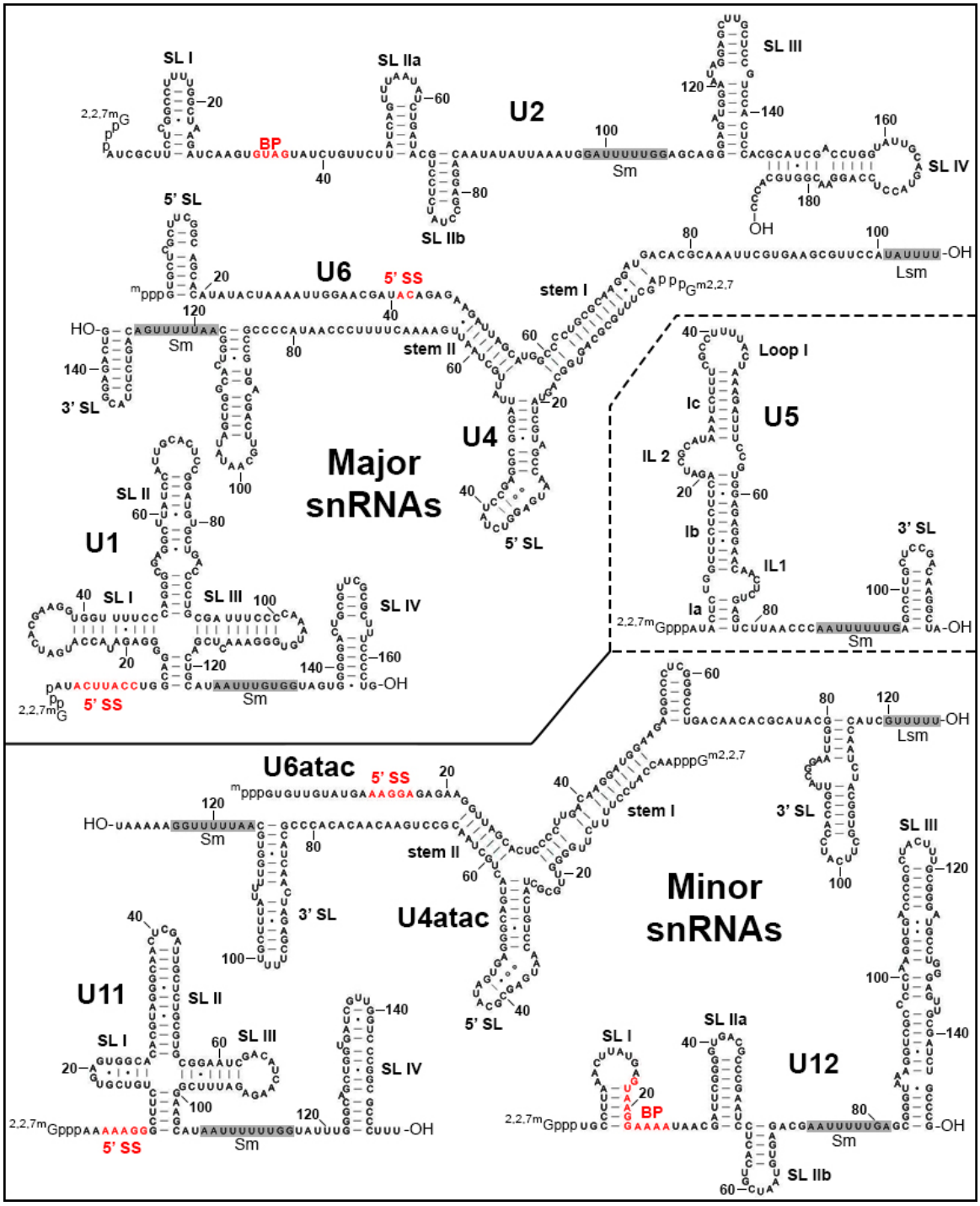


Figure 1.5. Major and minor spliceosomal snRNAs.
Predicted secondary structures for human major (U2-dependent) and minor (U12-dependent) spliceosomal snRNAs are depicted based on previous works (Patel and Steitz 2003, Sikand and Shukla 2011). U5 snRNA is present in both spliceosomes. Regions of snRNAs which interact with intron sequences during splicing are in red and binding sites for Sm and Lsm protein complexes are highlighted in grey. Cap structures present at snRNA 5′ ends are indicated, however, other nucleotide modifications (at internal locations) occurring in natural snRNAs are not shown.

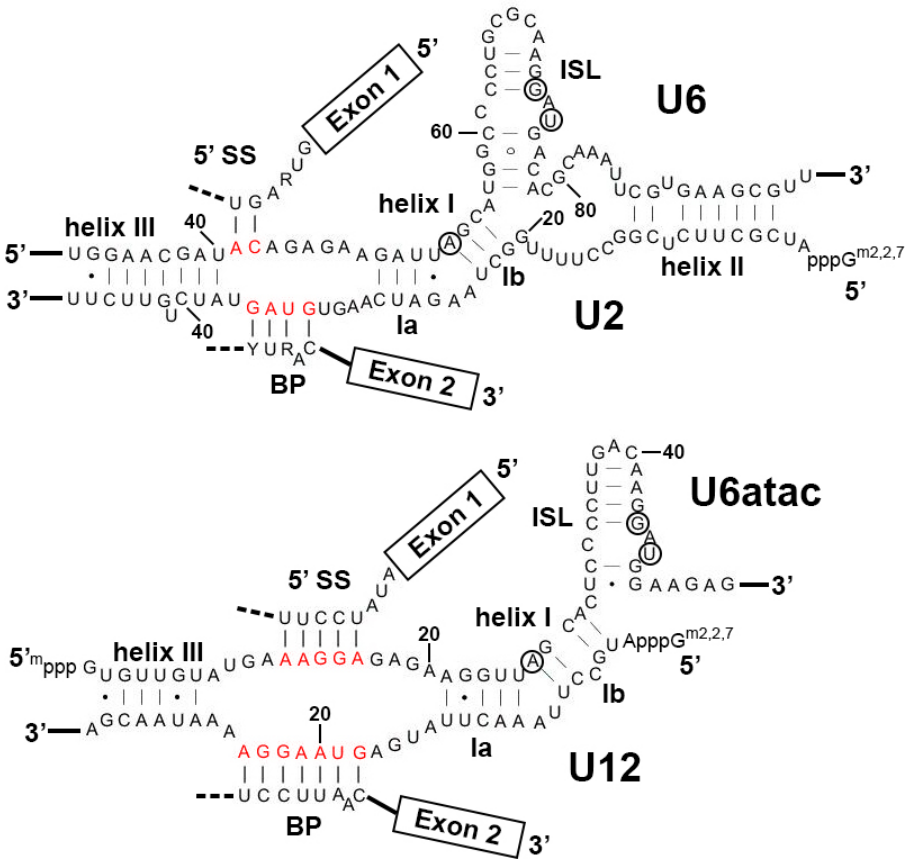


Figure 1.6. RNA-RNA interactions at the catalytic core of U2 and U12 spliceosomes. Human U2- and U12-dependent spliceosomal snRNAs are shown base pairing to U2-type and U12-type spliceosomal intron 5′ splice sites (5′ SS) and branch point (BP) sequences. Regions of snRNAs involved in binding to intron sequences are in red (as in Figure 1.5). Names of conserved snRNA-snRNA intermolecular helices are indicated. Nucleotides involved in binding catalytic Mg^{2+} ions in U6 (Fica *et al.* 2013) and their equivalent nucleotides in U6atac are circled. ISL = Intramolecular stem-loop. Figure was adapted from Patel and Steiz (2003); and Turunen *et al.* (2012).

1.2.2 Spliceosomal proteomes

The spliceosome may be a ribozyme; however, snRNAs function in the context of snRNP (pronounced 'snurp') complexes containing both stable and transiently bound proteins. Proteomic analysis of purified human U2-dependent spliceosomes has revealed more than 300 interacting spliceosomal proteins (Jurica and Moore 2003). However, more intron-poor eukaryotes appear to have substantially fewer spliceosomal proteins and for *S. cerevisiae* this number is ~100 protein factors (Fabrizio *et al.* 2009). The large discrepancy

in numbers of spliceosomal proteins is likely due to an expanded repertoire of alternative splicing regulators in higher eukaryotes whereas, with exception of a few rare cases (Juneau *et al.* 2009, Meyer *et al.* 2011), alternative splicing is essentially absent in *S. cerevisiae*. At least some of the spliceosomal proteins shared between humans and yeast likely represent 'core' U2-dependent spliceosomal proteins. Although, excluding some trypanosomes (Luz Ambrosio *et al.* 2009), biochemical examination of spliceosomes from the majority of eukaryotic diversity (i.e. protists) are lacking and it is unclear to what extent certain spliceosomal proteins are conserved across the eukaryotic tree.

Bioinformatic surveys have helped in defining the basal spliceosomal proteome by identifying spliceosomal protein gene orthologs in diverse eukaryotes (Nixon *et al.* 2002, Collins and Penny 2005, Simoes-Barbosa *et al.* 2008). These studies have highlighted a subset of U2-dependent spliceosomal proteins which are not only likely to be critical to spliceosome function, but also present in the last eukaryotic common ancestor (LECA) (Collins and Penny 2005). The most highly conserved spliceosomal protein is the U5 snRNP protein Prp8p (220-kDa protein in humans) which shows ~62% amino acid identity between *S. cerevisiae* and humans in a pairwise BLASTP alignment. Recently, a crystal structure of Prp8p lacking only the amino-terminal portion of the protein revealed a domain organization strikingly similar to IEPs from group II introns, suggesting the two protein families have common ancestry and may function similarly to stabilize splicing active sites (Galej *et al.* 2013) (Figure 1.8B). Another group of universal core spliceosomal proteins are the Sm and Sm-Like (LSm) protein families. Sm proteins form heteroheptameric rings and stably associate with uridine-rich motifs (Sm binding sites) found in all U2-dependent and U12-dependent snRNAs with the exception of U6 and U6atac snRNAs (Scofield and Lynch 2008) (Figure 1.5). Similarly, LSm proteins form heptameric rings and associate

with the uridine-rich sequence at the 3′ ends of U6 and U6atac snRNAs. However, snRNA-LSm complex binding is dynamic and LSm complex association and dissociation is required at different steps during spliceosome assembly (Scofield and Lynch 2008).

Considerably less is known about the proteome of U12-dependent spliceosomes. This is due in part to the low abundance of U12-dependent spliceosomal components, which are ~100-fold less abundant than those of the U2-dependent spliceosome (Tarn and Steitz 1996). Nonetheless, biochemical purification of the U11/U12 di-snRNP from humans identified both shared U1 and U2 snRNP spliceosomal proteins as well as several proteins unique to the U11/U12 di-snRNP (Will *et al.* 1999). In another study, purified minor spliceosomal U6atac/U4atac•U5 tri-snRNP complexes were found to contain apparently identical protein constituents as major spliceosomal U6/U4•U5 complexes (Schneider *et al.* 2002).

1.2.3 Spliceosomal snRNP biogenesis

Details of spliceosomal snRNPs biogenesis are described more completely in other reviews (Will and Luhrmann 2001, Kiss 2004); only limited aspects of snRNP biogenesis will be discussed here and where they have been best studied, in humans. With exception of U6 and U6atac snRNAs, all major and minor snRNAs are transcribed by RNA polymerase II and initially receive a standard 7-methylguanosine cap, but unlike pre-mRNAs, they are not polyadenylated. U6 and U6atac snRNAs are transcribed by RNA polymerase III and their 3′ ends terminate in poly-uridine tracts as many other RNA pol III transcripts. Unlike the other snRNAs, U6 and U6atac do not have nucleotide cap structures and instead bear γ-monomethyl-triphosphates at their 5′ ends (Figure 1.5).

Precursor snRNA transcripts are exported to the cytoplasm (U6 and U6atac remain in the nucleus) (Ohno *et al.* 2000) where they are recognized by the survival of motor

neurons (SMN) complex which facilitates binding of Sm protein complexes to snRNA Sm binding sites (Friesen and Dreyfuss 2000) (Figure 1.5). Next, snRNA 3′ ends are trimmed and nucleotide caps are hyper-methylated to 2,2,7-trimethylguanosine (TMG) caps (Mouaikel *et al.* 2002). Sm complexes and TMG caps act as nuclear localization signals and snRNPs are reimported into the nucleus. Nascent snRNPs then transit to Cajal bodies where they associate with snRNP-specific proteins and snRNAs undergo numerous post-transcriptional modifications, mostly consisting of 2′-O-methylation of certain ribose moieties and uridine to pseudouridine (Ψ) isomerization (Karijolich and Yu 2010). Mature snRNPs are then assembled into functional spliceosomes and recruited for splicing.

1.2.4 The spliceosome cycle

Spliceosomes are dynamic machines which require coordinated assembly (and disassembly) of spliceosomal components to achieve intron splicing. These steps are described in greater detail in several excellent reviews (Matlin and Moore 2007, Will and Luhrmann 2011). Here, I provide an overview of the spliceosome cycle, comparing both major and minor splicing pathways.

U2-dependent spliceosome assembly begins with the binding of U1 snRNP to a pre-mRNA by base pairing interactions between nucleotides at the 5′ end of U1 snRNA and the intron 5′ SS sequence (Seraphin *et al.* 1988, Siliciano and Guthrie 1988). Independently, U2 auxiliary factor (U2AS) binds the poly-pyrimidine tract (if present) and intron 3′ SS (Zamore and Green 1989). Together, this forms the spliceosomal ‘E’ (early) complex (Figure 1.7, E). U2 snRNP then associates and binds to the intron BP sequence (‘UACUAAC’, in humans) using its BP interacting sequence (‘GUAGUA’) (Zhuang *et al.* 1989). This interaction bulges the intron BP adenosine and forms the spliceosomal ‘A’ complex (Figures 1.6 and 1.7, A). Next, U4, U5 and U6 snRNPs join as a pre-formed

U6/U4•U5 tri-snRNP complex forming the ‘B’ complex (Figure 1.7, B). Within this complex, U6 and U4 snRNAs are engaged in extensive intermolecular base pairing (Vankan *et al.* 1992) (also see Figure 1.5) and U5 snRNP is associated via protein-protein interactions with U4 snRNP (Liu *et al.* 2006). A series of structural rearrangements culminates in the unwinding of U6/U4 snRNA base pairs (Laggerbauer *et al.* 1998) and U1 snRNP is displaced at the 5′ SS by U6 snRNP (Sawa and Abelson 1992). At this time, new intermolecular base pairs are formed between U6 and U2 snRNAs (Madhani and Guthrie 1992) (Figure 1.6) and U1 and U4 snRNPs dissociate. U5 snRNP binds the flanking exons (Newman and Norman 1992) and the spliceosome is poised for the first step of splicing in the B* complex (not shown). The first step of splicing then occurs by attack of the BP adenosine on the 5′ SS (Figure 1.7, C1). Subsequently, additional rearrangements position the intron 3′ SS for attack by the 5′ exon (Figure 1.7, C2) and results in the ligation of the flanking exons. Spliceosomal components then dissociate to begin subsequent rounds of splicing, releasing the spliced mRNA and excised intron lariat.

The overall splicing cycle of the U12-dependent spliceosome is thought to occur similarly to that of the U2-dependent spliceosome; however, involving the minor spliceosome-specific U11, U12, U4atac and U6atac snRNPs and common U5 snRNP (Patel and Steitz 2003) (Figure 1.7, right panel). However, one major difference is that U11 and U12 snRNPs bind to U12-type introns as a pre-formed di-snRNP (Frilander and Steitz 1999). Thus, there is no equivalent of the spliceosomal ‘E’ complex in the U12-dependent spliceosome cycle.

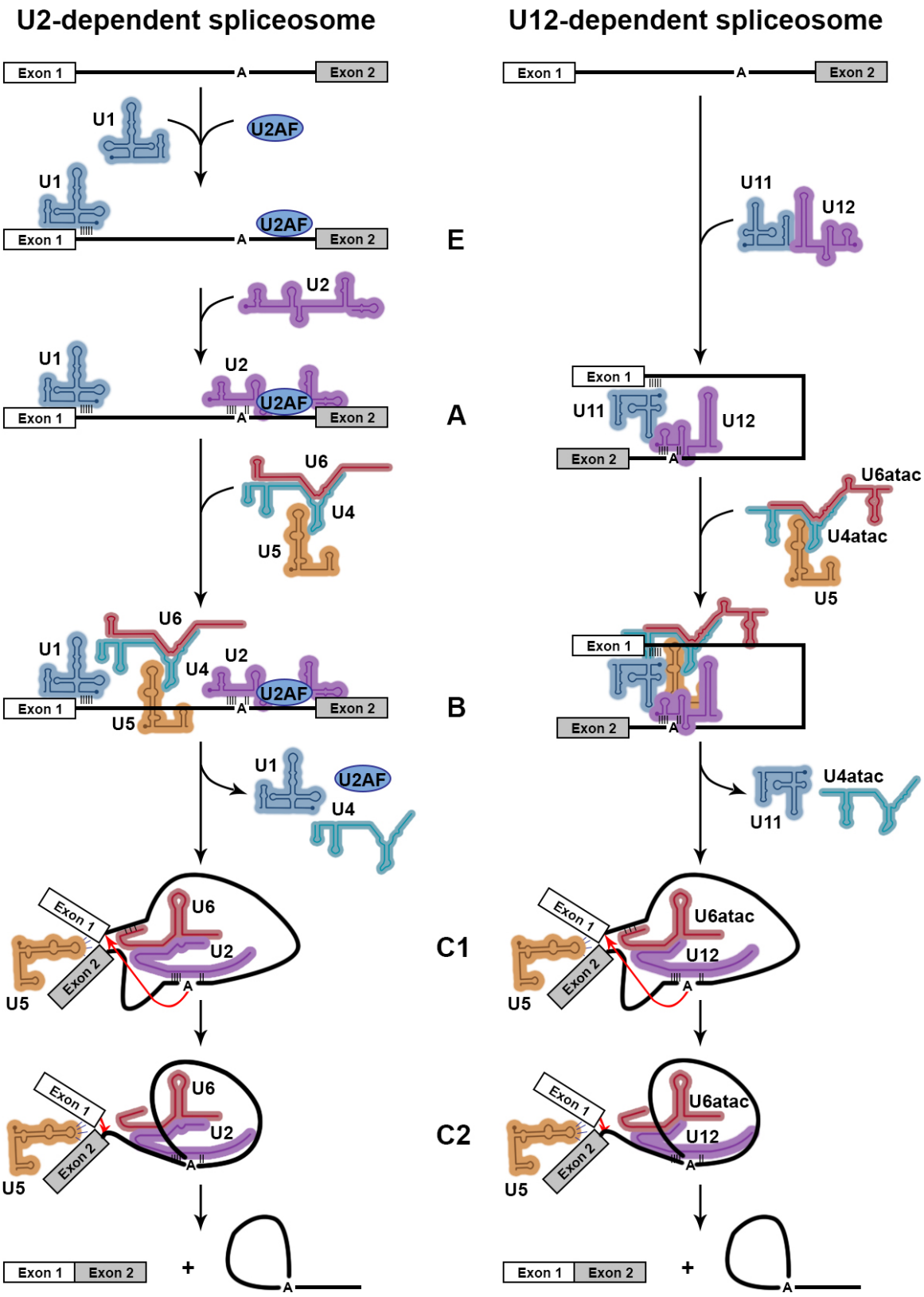


Figure 1.7. U2-dependent and U12-dependent spliceosome cycles.
See text for descriptions for each step of splicing.

1.3 Spliceosomal Intron and Spliceosome Evolution

Since their discovery almost 40 years ago, questions surrounding the evolution, proliferation and ancient functions of spliceosomal introns have been topics of extensive debate. However, in this time, fundamental questions regarding spliceosomal intron evolution remain unanswered. What evolutionary forces and processes led to the creation of spliceosomal introns and the spliceosome? When did spliceosomal introns arise? How are introns lost and gained? While complete answers to these questions are still elusive, recent advances in genomics and large-scale examination of introns in diverse eukaryotes have provided some clues about these mysteries.

1.3.1 'Introns early' versus 'introns late'

Traditionally, it is appropriate to begin discussion of intron evolution under the framework of two classical competing hypotheses: the so-called 'introns early' versus 'introns late' debate (Doolittle 1987). 'Introns early' posits that introns were present at the very earliest stages of cellular life and played an important role in the generation of longer polypeptides via intron-mediated recombination of smaller protein-coding modules (exons). The related 'exon theory of genes' builds on this hypothesis by suggesting that differential recombination of exon coding modules, called 'exon shuffling' may have been important for generating new protein domain topologies with emergent functions (Gilbert 1987). 'Introns late' counters that introns evolved later, at some point during eukaryotic evolution and that introns accumulated in eukaryotic genomes over time. Initial support for introns early came from the observation of a higher than expected number of phase 0 intron insertions occuring between protein functional domains (Fedorov *et al.* 1992, Long *et al.* 1995). Indeed, this observation was consistent with the exon theory of genes and provides a rationale for the appearance of multidomain proteins in both eukaryotes and prokaryotes

(de Souza *et al.* 1996). On the other hand, introns early is undermined by the (near) absence of introns in any prokaryotic genome and evidence for more recent intron gain events during eukaryotic evolution (Logsdon 1998). However, advocates of 'introns early' insist the absence of introns in prokaryotes may be explained by genomic streamlining which purged ancestral introns from prokaryotic genomes (Long *et al.* 1995).

1.3.2 Reconstruction of ancestral exon-intron structures

Availability of ever-increasing genomic sequence data from diverse eukaryotes combined with computational strategies to reconstruct ancestral exon-intron structures has provided additional fuel to the introns early/late debate. Most reconstructions have been based on the analysis of intron insertional positions within orthologous genes in extant eukaryotes (Fedorov *et al.* 2002, Rogozin *et al.* 2003). The logic goes that if an intron interrupts the same position in orthologous genes from two or more eukaryotes, it is reasonable to assume that the intron was present in the gene from the last common ancestor of those eukaryotes. That is, if a particular intron insertion position shows conservation in orthologous genes from eukaryotes separated by large evolutionary distances, the intron is assumed to have been acquired earlier than an intron whose insertion position shows a more limited phylogenetic distribution. Thus, by examining conservation of intron insertion positions in hundreds of orthologous genes from diverse eukaryotes, it is possible to infer exon-intron structures in ancient eukaryotic ancestors.

While this maximum parsimony analysis (or argument) has some utility for inferring the ancestral exon-intron structure for a given gene (i.e. intron insertion or non-insertion), other variables make this analysis less than straightforward. One major caveat is that there may be preferential identical sites for independent intron insertion events. Exonic sequences flanking intron insertion positions show sequence bias with the consensus:

(C/A)AG//G, where '//' indicates the 5′ and 3′ intron splice boundaries (Rogozin *et al.* 2012). These so-called 'proto-splice sites' have been suggested to be preferrential sites for the insertion of new spliceosomal introns (Dibb and Newman 1989, Dibb 1991). Thus, when attempting to infer ancestral exon-intron structures, one must consider the possibility of parallel independent intron gains at proto-splice sites.

The latest generation of methods for reconstructing ancestral exon-intron structures utilize more sophisticated probabilistic models which may account for parallel intron gains at proto-splice sites (as well as other variables). A recent reconstruction of ancestral exon-intron structure using 245 orthologous genes from 99 complete eukaryotic genomes, representing three of the five eukaryotic supergroups, inferred an intron-rich last eukaryotic common ancestor (LECA), with a predicted intron density of 54-74% that of modern humans (4.3 introns/kb gene) (Csuros *et al.* 2011). While other studies predict slightly differing values, all seem to echo that LECA was probably intron rich (Rogozin *et al.* 2003, Csuros *et al.* 2008) and contained a complex spliceosome capable of alternative splicing (Collins and Penny 2005). Most reconstructions also indicate that eukaryotic evolution has been dominated by gradual intron loss punctuated by sporadic, massive intron gain events which coincide with the emergence of new eukaryotic groups (Csuros *et al.* 2011).

1.3.3 Origins of spliceosomal introns and the spliceosome

Early eukaryotes likely possesed intron-rich genomes, although it is unclear how (and when) spliceosomal introns and the spliceosome arose in evolution. However, there is strong support that spliceosomal introns and the spliceosome share common ancestry with group II introns (Rogozin *et al.* 2012). This prediction is founded on multiple lines of evidence: 1) an identical two-step mechanism of splicing, resulting in an excised intron lariat, 2) structurally and functionally analogous regions of group II introns and snRNAs

involved in splicing (i.e. group II intron domains DV and DVI are analogous to U6 and U2 snRNAs) with similar metal ion binding centres involved in catalysis, and 3) the most conserved spliceosomal protein Prp8p shares significant structural similarity to group II intron IEPs (Figure 1.8). Several hypotheses have been put forth to explain how ancestral group II intron-like elements were possibly transferred to the nuclear genome of a eukaryotic ancestor (Lynch and Richardson 2002, Koonin 2006). One widely accepted view is that spliceosomal introns arose following engulfment of an α-proteobacterium (which contained group II introns) by a eukaryotic progenitor (likely an archaeon) (Koonin 2006). Group II introns would have then have been transferred to the host chromosome and subsequently proliferated until much of the genome consisted of group II intron-derived elements. Most protein coding genes would then have contained one or more group II introns. While their ability to self-splice from pre-mRNAs would have had a (near) neutral effect on gene function, the establishment of a common set of *trans*-acting RNA cofactors (the snRNAs) would have aleviated mutational pressure to maintain group II intron structures. Indeed, examples of *trans*-complementation by group II intron fragments exist naturally for the splicing of degenerate group II introns from eukaryotic organellar genomes (Jarrell *et al.* 1988, Goldschmidt-clermont *et al.* 1991). Simultaneously, the resulting 'proto-spliceosomal' snRNA components would have recruited pre-existing protein factors involved in other cellular processes which (presumably serendipidously) aided in the splicing reaction or facilitated beneficial alternative splicing of pre-mRNAs. In a recent study, expression of a group II intron-containing pre-mRNA in *S. cerevisiae* resulted in cytoplasmic localization of both spliced and non-spliced mRNAs as well as suppression of translation of both mRNA forms (Qu *et al.* 2014). Thus it appears that mechanisms to

control expression of group II intron-containing transcripts now exist in modern eukaryotes and this may explain the absence of group II introns in extant eukaryotic nuclear genomes.

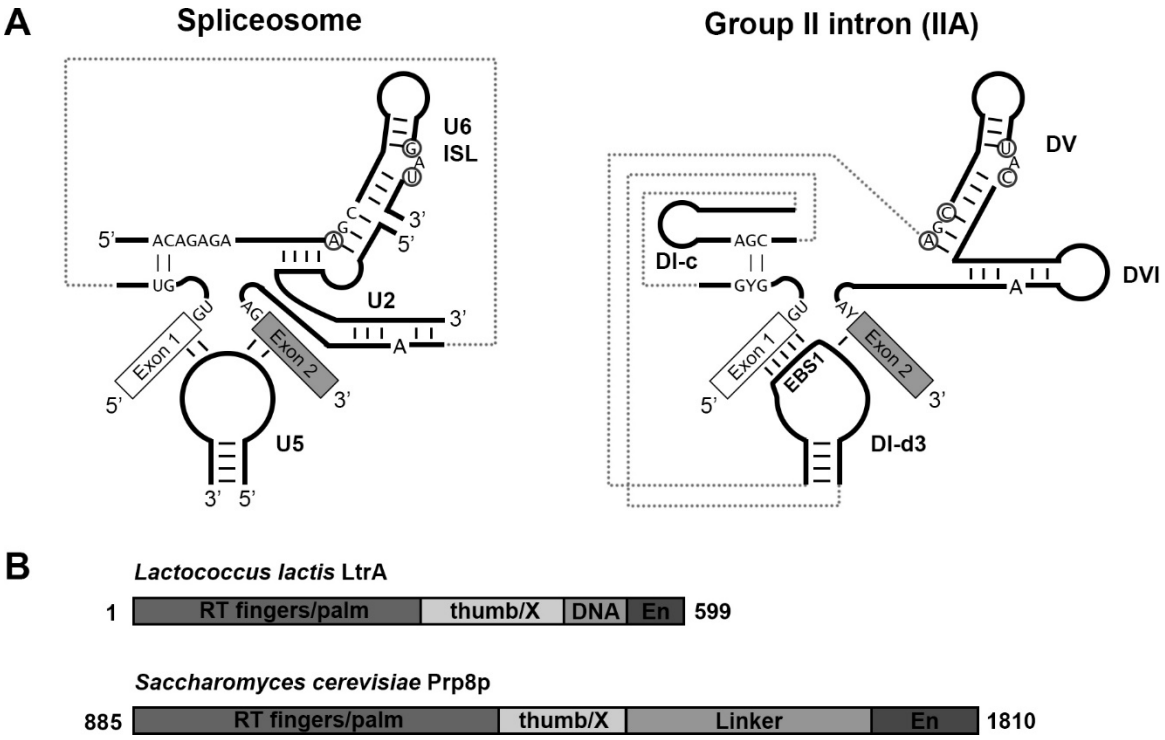


Figure 1.8. Similarities between the spliceosome and group II introns.
(**A**) Diagrams of the RNA catalytic core of the spliceosome and group IIA introns are shown side by side with structurally analogous regions indicated. Conserved nucleotides involved in forming important inter- and intramolecular base pairings are shown and catalytic metal binding nucleotides are circled. ISL = Intramolecular stem-loop. (**B**) Domain organizations for *Lactococcus lactis* LtrA intron encoded protein (Blocker *et al.* 2005) and the spliceosomal protein Prp8p from *Saccharomyces cerevisiae* (Galej et al. 2013).

1.3.4 Emergence of U2- and U12-dependent spliceosomes

Even more mysterious than the appearance of one spliceosome in eukaryotes, is the appearance of a second spliceosome. Whilst U12-type introns and U12-dependent spliceosomal components are only found in a few eukaryotic genomes, their phylogenetic distribution indicates they were both present in LECA (Russell *et al.* 2006, Lopez *et al.* 2008). The striking similiarities in snRNA structures and protein components strongly suggest the two spliceosomes have common evolutionary origins. While speculative, three

separate models for the evolution of the U2- and U12-dependent spliceosomes have been proposed: i) parasitic invasion, ii) co-divergence and iii) fission-fusion models (Burge *et al.* 1998). The parasitic invasion model postulates that U12-type type introns originated independently via the parasitic invasion of group II introns into a eukaryotic genome already containing U2-type introns and U2-dependent spliceosomal system. U12-dependent spliceosomal snRNAs would then have been derived from group II intron fragments with spliceosomal proteins being recruited from the existing U2-dependent spliceosome. The co-divergence and fission-fusion models presume U2-type and U12-type introns and their respective spliceosomes are homologous. In the co-divergence model, a eukaryotic ancestor possessing U2-type introns is proposed to have undergone a partial or complete genome duplication. Following this, some U2-type introns would have then co-diverged their splice site and branch point sequences with the duplicated snRNA gene copies, giving rise to U12-type introns and U12-dependent spliceosomal snRNAs. Finally, the fission-fusion model invokes a scenario where a speciation event produced two separate lineages whose spliceosomal introns and spliceosomes subsequently diverged. Next, genetic material from the two lineages would have combined into a common nuclear genome upon some later cell-cell fusion event, creating a new species which then harbored both intron types and the spliceosomes to remove them.

1.4 Diplomonads as Model Organisms

Diplomonads are genetically-diverse anaerobic and micro-aerophilic protists belonging to the eukaryotic supergroup Excavata and phylum Metamonada (Adl *et al.* 2012) (Figure 1.9). Members of this group are marked by several unusual cellular features including highly reduced mitochondrial remnants (Tovar *et al.* 2003, Jerlstrom-Hultqvist

et al. 2013) and the presence of two transcriptionally-active diploid nuclei containing (apparently) identical genomic contents (Kabnick and Peattie 1990, Yu *et al.* 2002). Many diplomonads are parasitic although some are also free-living (e.g. *Trepomonas spp.*) or commensals (*Enteromonas spp.* and *Spironucleus barkhanus*). The most-studied diplomonad is *Giardia lamblia* (syn. *G. intestinalis*, *G. duodenalis*), an important waterborne intestinal parasite of mammals and causative agent of giardiasis (Lane and Lloyd 2002). Certain members of the diplomonad genus *Spironucleus* have also received attention as important veterinary pathogens affecting freshwater and marine fish causing significant economic losses of both ornamental and farmed fish (Williams *et al.* 2011).

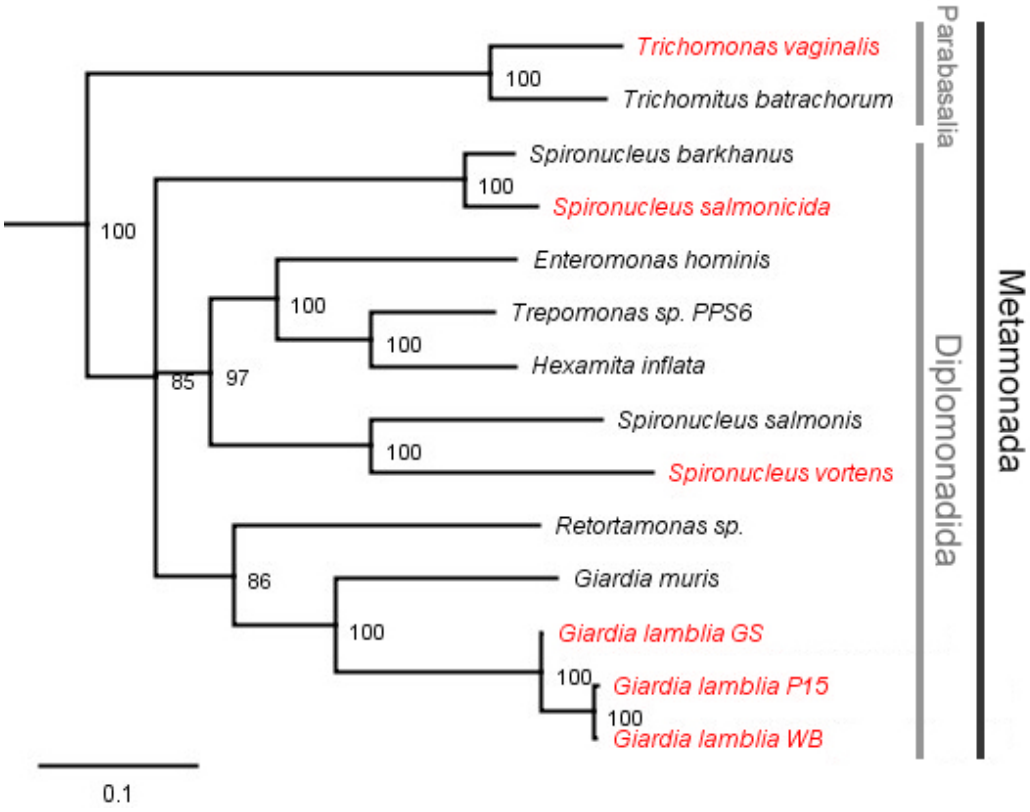


Figure 1.9. Phylogeny of Metamonada.
Neighbor-joining phylogenetic tree of diplomonad, parabasalid and *Retortamonas sp.* based on small subunit rRNA gene sequences. Bootstrap values from 100 replicates are indicated at nodes and organisms examined in the thesis are highlighted in red text.

G. lamblia (and other diplomonads) were originally believed to be the deepest branching extant eukaryotes owing to their apparent lack of mitochondria and early phylogenies based on rRNA sequence (Sogin *et al.* 1989) and other conserved molecular markers (Hirt *et al.* 1999). However, this notion has been undermined by the finding of reduced mitochondrial relics (mitosomes and hydrogenosomes) in diplomonads (Tovar *et al.* 2003, Jerlstrom-Hultqvist *et al.* 2013) and evidence of substantial lateral gene transfer from bacteria to diplomonad nuclear genomes (Andersson *et al.* 2003, Andersson *et al.* 2007). Consequently, many of the unusual cellular and molecular features of diplomonads may be more recently derived in this lineage and *a priori* should not be considered characteristics of ancestral eukaryotes.

Recent sequencing of several diplomonad nuclear genomes has revealed a surprising level of genetic diversity in members of this group (Figure 1.9). Of the seven *G. lamblia* species complex assemblages (A-G), draft genome sequences are now available from assemblages A (WB isolate), B (GS) and E (P15) (Morrison *et al.* 2007, Franzen *et al.* 2009, Jerlstrom-Hultqvist *et al.* 2010). Interestingly, *G. lamblia* assemblages A and B display only ~77% nucleotide identity in protein coding regions, suggesting they represent distinct *Giardia* species (Franzen *et al.* 2009). Moreover, examination of expressed sequence tags (ESTs) from the morphologically-identical diplomonads *Spironucleus barkhanus* and *Spironucleus salmonicida* have revealed an expanded repertoire of genes not found in *G. lamblia* and differences in codon usage and frequency of allelic variants between the two *Spironucleus* species (Roxstrom-Lindquist *et al.* 2010). Very recently, a complete draft genome (280X coverage for Illumina sequence reads) for *S. salmonicida* has become publically accessible (Xu *et al.* 2014) and partial genomic sequences from *Spironucleus vortens* are available for download (Joint Genome Institute, unpublished

data). The large volume of sequence data available from diverse diplomonads provide the opportunity to perform powerful comparative genomic studies to evaluate both conserved and unique molecular characters in different diplomonads.

Diplomonads offer an interesting perspective in the study of spliceosomal intron and spliceosome evolution and function. Most characterized diplomonads show extreme genomic minimalism and generally more simplified molecular machinery with numerous components involved in DNA replication, transcription and RNA processing apparently missing in *G. lamblia* (Morrison *et al.* 2007). Diplomonads also retain very few spliceosomal introns (Russell *et al.* 2005, Xu *et al.* 2014). Consequently, *G. lamblia* and other diplomonads make useful models for the study of spliceosomal intron and spliceosome structures in extremely intron-poor eukaryotes which are possibly undergoing complete loss of their U2-type (or U12-type) spliceosomal introns.

Only ~30 genes encoding spliceosomal proteins could be confidently predicted in various bioinformatics studies for *G. lamblia* (Nixon et al. 2002, Collins and Penny 2005, Korneta et al. 2012) and numerous other 'core' spliceosomal proteins appear to be absent or sufficiently diverged to escape detection using bioinformatic methods. Thus, similar to many other cellular processes in *G. lamblia*, it is anticipated that pre-mRNA splicing is also highly simplified in this organism. As opposed to human spliceosomes which may contain hundreds of spliceosomal protein factors, the characterization of reduced spliceosomal systems from *G. lamblia* and other diplomonads will help to elucidate the minimal core of spliceosomal proteins and snRNA structures required to facilitate spliceosomal intron excision. Moreover, study of divergent splicing systems will highlight the natural flexibility of form and function of spliceosomes and perhaps provide insights into evolutionary

transitions giving rise to U2- and U12-dependent splicing systems and ultimately, the origins of spliceosomal introns and the spliceosome.

1.5 Objectives

The overall objectives of my studies were to use molecular and bioinformatic techniques to characterize pre-mRNA splicing and RNA processing strategies in intron-poor diplomonads.

Very few spliceosomal introns have been identified in any diplomonad to date and only specifically in *G. lamblia* and *S. salmonicida*. Consequently, my initial objective was to use computational and molecular techniques to predict additional spliceosomal introns in *G. lamblia* and *S. vortens* genomic DNA sequences and assess the conservation of intron features in this group (Chapters 2 and 4). Mechanisms of ncRNA processing are largely unexplored in diplomonads. Therefore, as a second objective, I sought to use bioinformatic techniques to identify ncRNA processing signals in diplomonads and employ experimental methods (RT-PCR and RNA-end mapping) to verify modes of ncRNA expression and processing (Chapter 3). Finally, a collection of spliceosomal snRNAs were predicted bioinformatically in *G. lamblia* (Chen et al. 2008), although as of yet no single spliceosomal snRNA has been experimentally-confirmed in any diplomonad. Thus, my third objective was to use molecular (primer-extension, northern blot, RNA-end mapping) and bioinformatic (pattern searching and covariation model) strategies to identify additional ncRNAs from *G. lamblia* and *Spironucleus spp.* (Chapters 3 and 4).

Chapter 2: Numerous Fragmented Spliceosomal Introns, AT–AC Splicing, and an Unusual Dynein Gene Expression Pathway in *Giardia lamblia*

Reprinted with permission from:

"Numerous Fragmented Spliceosomal Introns, AT–AC Splicing, and an Unusual Dynein Gene Expression Pathway in *Giardia lamblia*"

Molecular Biology and Evolution **2012**. *29*, 43-49

Scott W. Roy*, Andrew J. Hudson*, Joella Joseph, Janet Yee and Anthony G. Russell

* These authors contributed equally

Contributions:

SWR performed the initial search for introns which identified the single *cis*-intron (26S proteasome non-ATPase regulatory subunit 4) and *trans*-introns for *HSP90* and DHC β genes and performed data analysis and co-wrote the manuscript. AJH identified the DHC γ *trans*-intron, wrote the first draft of the paper, generated figures and performed all experiments. JJ and JY cultivated *Giardia* cells, generated nucleic acid samples for the study and provided useful discussion for the writing of the manuscript. AGR designed experiments and wrote the final draft of the paper. All authors read and approved the final manuscript.

Changes Incorporated:

Some wording has been changed to connect the results from this chapter to those discovered subsequently (in Chapters 3 and 4). Details for experimental methods have been described in greater detail.

2.1 Introduction

Spliceosomal introns are quasi-random sequences that interrupt nuclear-coding genes and are removed from RNA transcripts by the spliceosome (Jurica and Moore 2003). Intron–exon structures vary dramatically, with orders of magnitude differences in number of introns and median intron length across species (Logsdon 1998). The evolution and functional significance of spliceosomal introns and the evolutionary forces underlying

these striking interspecific differences remain matters of much debate (Roy and Gilbert 2006).

Trans-splicing is an RNA splicing event in which two or more separate RNA primary transcripts are ligated to yield a single mature RNA. In spliced leader (SL) *trans*-splicing in some protists and metazoans, one or more common noncoding leader sequences are spliced onto 5′ ends of pre-mRNA transcripts of various genes (Bonen 1993). *Trans*-splicing of independently transcribed nuclear-coding mRNAs is considerably rarer with few reported examples, such as *mod* (*mdg4*) in *Drosophila*, the bursicon gene in mosquitoes, and a few human mRNAs (Li *et al.* 1999, Takahara *et al.* 2000, Dorn *et al.* 2001, Robertson *et al.* 2007). Very recently, a single case of *trans*-splicing was reported in *Giardia lamblia* (Nageshan *et al.* 2011).

The extremely intron-poor intestinal parasite *G. lamblia* occupies a unique position in the study of spliceosomal intron evolution. Consistent with its lack of mitochondria, *G. lamblia* was initially (and is sometimes still) thought to be an "early-branching" eukaryote, representing a primitive lineage within eukaryotes (Sogin 1991, Morrison *et al.* 2007). *Giardia lamblia* was also originally thought to be potentially intronless. Together, these possibilities suggested that early eukaryotes were intronless or extremely intron poor and that spliceosomal introns became abundant later, for instance, by spread of type II (Group II) self-splicing introns transferred from an early eukaryotic organelle (Cavalier-Smith 1991). However, evidence for a mitochondrial ancestry of *G. lamblia* (Hashimoto *et al.* 1998, Roger *et al.* 1998) and the repositioning of *G. lamblia* on more recent phylogenetic trees has greatly altered our understanding of *G. lamblia*'s evolutionary history. Meanwhile, the finding of introns and splicing machinery in *G. lamblia* and other potentially early-diverging lineages (Fast *et al.* 1998, Nixon *et al.* 2002, Russell *et al.* 2005,

Vanacova *et al.* 2005, Hudson *et al.* 2012) as well as the finding that both splicing machinery and intron positions are shared across widely diverged eukaryotes (Archibald *et al.* 2002, Fedorov *et al.* 2002, Rogozin *et al.* 2003, Collins and Penny 2005) indicate that early eukaryotic ancestors already had a well-developed spliceosomal system, with recognizably modern splicing machineries and intron complements.

We report bioinformatic and molecular studies of splicing and spliceosomal introns in *G. lamblia*. We find a variety of surprising phenomena, including a high frequency of *trans*-spliced introns, division of a single ancestral gene into four separate pieces, and utilization of atypical AT–AC splicing boundaries. Extensive base pairing between intron halves, and similar RNA secondary structures in long *cis*-spliced introns, suggests an evolutionary pathway for transition from *cis*- to *trans*-splicing of coding introns. These results reveal remarkable complexity of gene expression in a species often thought to be highly "simplified," and point to an unappreciated diversity of spliceosomal structures in eukaryotes.

2.2 Materials and Methods

2.2.1 Identification and comparative genomics of *G. lamblia* introns

To identify *cis*-spliced introns, we performed bioinformatic searches of the *Giardia* genome to identify all instances of consensus sequences from previously known *Giardia* introns: [G/C]TATGTN$_{0–500}$CT[A/G]ACN$_{3–5}$AG, where N$_{n–m}$ indicates a block of between n and m nucleotides. We then identified instances in which removal of the sequence would extend an open reading frame (ORF) by at least 50 codons. Next, we performed BLASTN searches of these extended ORFs against *Giardia* WB isolate whole-genome shotgun contigs on the GiardiaDB web site (http://GiardiaDB.org/GiardiaDB/), using default

parameters unless otherwise specified, (Expect '*E*' threshold = 10, word size = 28, match/mismatch scores = 1/-2) to look for evidence of interruption of a conserved gene. We then identified all matches to the canonical 3′ splice site CT[A/G]AC<u>ACA</u>CAG, allowing one mismatch within the underlined three positions. We also searched for potential upstream 5′ splice sites with similarity to [G/C]TATGT but did not find clear candidates. Sequence searches of the downstream sequence led us to the two dynein heavy chain *β* (*DHC β*) split introns.

Stimulated by the finding of *trans*-spliced introns in one dynein gene, we decided to study other dynein genes. To catalog the various dynein heavy chain (DHC) isoforms present in the *G. lamblia* genome, we first analyzed the 16 annotated ORFs designated as DHCs at the GiardiaDB web site. To verify the designation of each isoform, we performed BLASTP searches using default parameters using translated *G. lamblia* DHC coding sequences against non-redundant protein sequences in GenBank (http://blast.ncbi.nlm.nih.gov/) and analyzed the 20 lowest *E* value hits to dynein sequences of other organisms. In each case, this clearly delineated the isoform designation. Importantly, we found that all expected and essential heavy chain isoforms were present in the genome either as contiguous or fragmented genes and were single copy. We also used isoform sequences of *DHC β* and *γ* from other organisms as queries for BLASTN searches against the *Giardia* genome database to confirm that there were no additional contiguous gene copies. Our searches identified two *G. lamblia* ORFs encoding DHC γ proteins whose lengths, in sum, were similar to the sum of the *DHC β* fragments and therefore suggestive of a *trans*-splicing mechanism to unite the two *DHC γ* transcript halves.

Protein structural domains encoded within the *Giardia DHC β* and *γ* gene fragments were predicted using the 3D-JIGSAW program on the BioMolecular Modelling

Cancer Research UK web site (bmm.cancerresearchuk.org/~3djigsaw/) in order to pinpoint the regions of structural discontinuity relative to other dynein orthologs.

To compare sequences between different *Giardia* isolates, genome assemblies for the GS and P15 isolates were downloaded from GiardiaDB (GiardiaDB.org). For each *cis*- and *trans*-spliced intron, WB sequences were BLASTed against GS and P15. Sequences were aligned and analyzed using ClustalW2 software (Larkin *et al.* 2007) and by eye.

2.2.2 PCR and RT-PCR mediated confirmation of *trans*-splicing and *G. lamblia* genome annotation

Axenic *G. lamblia* trophozoites (strain WB clone 6; ATCC 30957) were cultured in modified TYI-S-33 medium until mid-late log phase (~10^5 cells/mL) using methods described elsewhere (Davids and Gillin 2011). Total RNA was extracted by use of TRIZOL Reagent (Invitrogen), and genomic DNA was isolated by the use of DNeasy kits (Qiagen).

Polymerase chain reaction (PCR) and reverse transcription (RT)-PCR reactions on *Giardia* nucleic acids were performed employing oligonucleotide primers listed in Appendix IV. For RT-PCR, first strand cDNAs were synthesized by mixing 1 μg of total *Giardia* RNA with 2 pmol reverse oligonucleotide primer in 12 μL nuclease free water and then incubated at 65°C for 5 minutes (template denaturation) and then 47°C for 10 minutes (primer annealing). The remaining RT reaction components were then added to create a 20 μL final reaction volume containing 1 X First Strand Buffer (50 mM Tris-HCl, pH 8.3, 75 mM KCl, 3 mM $MgCl_2$), 10 mM dithiothreitol (DTT), 0.5 mM of each deoxynucleotide triphosphate (dNTP) and 200 U SuperScript™ II reverse transcriptase (Life Technologies) and incubated at 47°C for 60 minutes. PCR reactions were then performed in 50 μL reactions containing 1 X ThermoPol® buffer (20 mM Tris-HCl, pH 8.8, 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2 mM $MgSO_4$, 0.1% Triton® X-100), 0.4 mM each dNTP, 20

pmol each forward and reverse oligonucleotide primer, 5 U *Taq* DNA Polymerase (New England Biolabs, NEB) and either 5 μL cDNA from RT experiments or 50 ng total *Giardia* DNA. PCR reactions typically consisted of an initial denaturation step at 94°C for 5 minutes, followed by 35 cycles of 94°C for 30 seconds, 55°C for 30 seconds and 68°C for 1 minute extension time per kilo base pair (kbp) fragment amplified, followed by a 7 minute final extension at 68°C. Product bands were excised from agarose gels and purified using QIAquick (Qiagen) or Omega Biotek gel extraction kits employing the manufacturer's protocols. PCR products were then cloned into the pCR2.1-TOPO (Invitrogen) or pJET1.2 vectors (Thermo Scientific) according the manufacturer's protocols and plasmids were subjected to Sanger dideoxy chain terminator DNA sequencing (Applied Biosystems ABI3730XL sequencer, MacroGen USA Corporation) using vector-specific sequencing primers (Appendix IV).

2.3 Results and Discussion

2.3.1 *Cis* and *trans*-spliced introns in *Giardia*

Previously identified *G. lamblia* introns exhibit atypical extended conserved 5′ and 3′ sequences (Nixon *et al.* 2002, Russell *et al.* 2005, Morrison *et al.* 2007), with the four known introns beginning [G/C]TATGTT and ending CT[A/G]AC[A/C]ACAG (Figure A1.1.1A). We searched the *G. lamblia* genome (http://GiardiaDB.org/GiardiaDB/) for intron-like sequences containing motifs with similarity to 5′ and 3′ splice site consensus sequences and filtered these candidates for additional indications of splicing (interruption of extended or conserved ORFs, evidence in expressed sequence tag [EST] databases, etc.). We identified two predicted "conventional" *G. lamblia cis*-introns, one in the gene for 26S proteasome non-ATPase regulatory subunit 4 (*PSMD4*), at the same position as introns in

various orthologs (Figure 2.1), and another since reported elsewhere (Morrison *et al.* 2007). We could not find an EST to support splicing of the *PSMD4* intron, however, inclusion of this intron is predicted to result in a frame shift and a truncated 26S proteasome non-ATPase regulatory subunit 4 protein (Figure 2.1A).

Figure 2.1. A putative *cis*-intron in 26S proteasome non-ATPase regulatory subunit 4 (**A**) Translated *G. lamblia* WB isolate *PSMD4* gene sequences either containing (w/ intron) or lacking (spliced) introns were aligned with *PSMD4* orthologs from: *Plasmodium vivax* (GenBank Accession XP_001613490), *Homo sapiens* (NP_002801) and *Arabidopsis thaliana* (AEE86956). Translated *PSMD4* amino acids conserved between all four species are in black highlighting while residues shared between at least one organism and *G. lamblia* are in grey highlighting. Intron-containing and intronless versions of the *G. lamblia PSMD4* gene are shown below and above alignments, respectively. Grey boxes represent coding exons and an arrow head indicates the splice junction. Intron sequences and their corresponding translated amino acids are in red text. (**B**) ClustalW alignment of *PSMD4* orthologs from *G. lamblia* WB, P15 and GS isolates are shown with translated amino acid sequences shown above alignments.

We also found instances of perfect matches to the 3′ consensus sequence without a nearby *Giardia* 5′ splice site consensus sequence. We noted two cases in which translation of downstream sequences revealed extensive amino acid sequence similarity to the conserved eukaryotic proteins: DHC β outer arm [nomenclature as in (Hook and Vallee 2006)] and heat shock protein 90 (Hsp90) (Figure 2.2A and B and Figures A.1.2 and A.1.3). In both cases, amino acid sequence similarities end abruptly at the 3′ splice site consensus

sequence match and consequently both translated sequences lack significant conserved amino-terminal regions (Figure 2.2A and B). Surprisingly, BLASTN searches of the upstream contig using orthologs of the missing conserved upstream coding sequences as queries, revealed no hits to upstream sequences. However, BLASTN searches against the genome revealed clear sequence similarity to internal regions of other long *G. lamblia* genomic contigs (in each case >20 kb from either end of the contig). In both cases, sequence conservation ends abruptly at a canonical *Giardia* 5′ splice site sequence (GTATGT). Splicing at the predicted splice junctions via a *trans*-splicing mechanism would generate contiguous mature mRNAs encoding a portion of the DHC β outer-arm protein and all of the Hsp90 protein (also since reported elsewhere by Nageshan *et al.* 2011), showing clear ungapped alignments to homologs (Figure 2.2A and B and Figures A.1.2 and A.1.3).

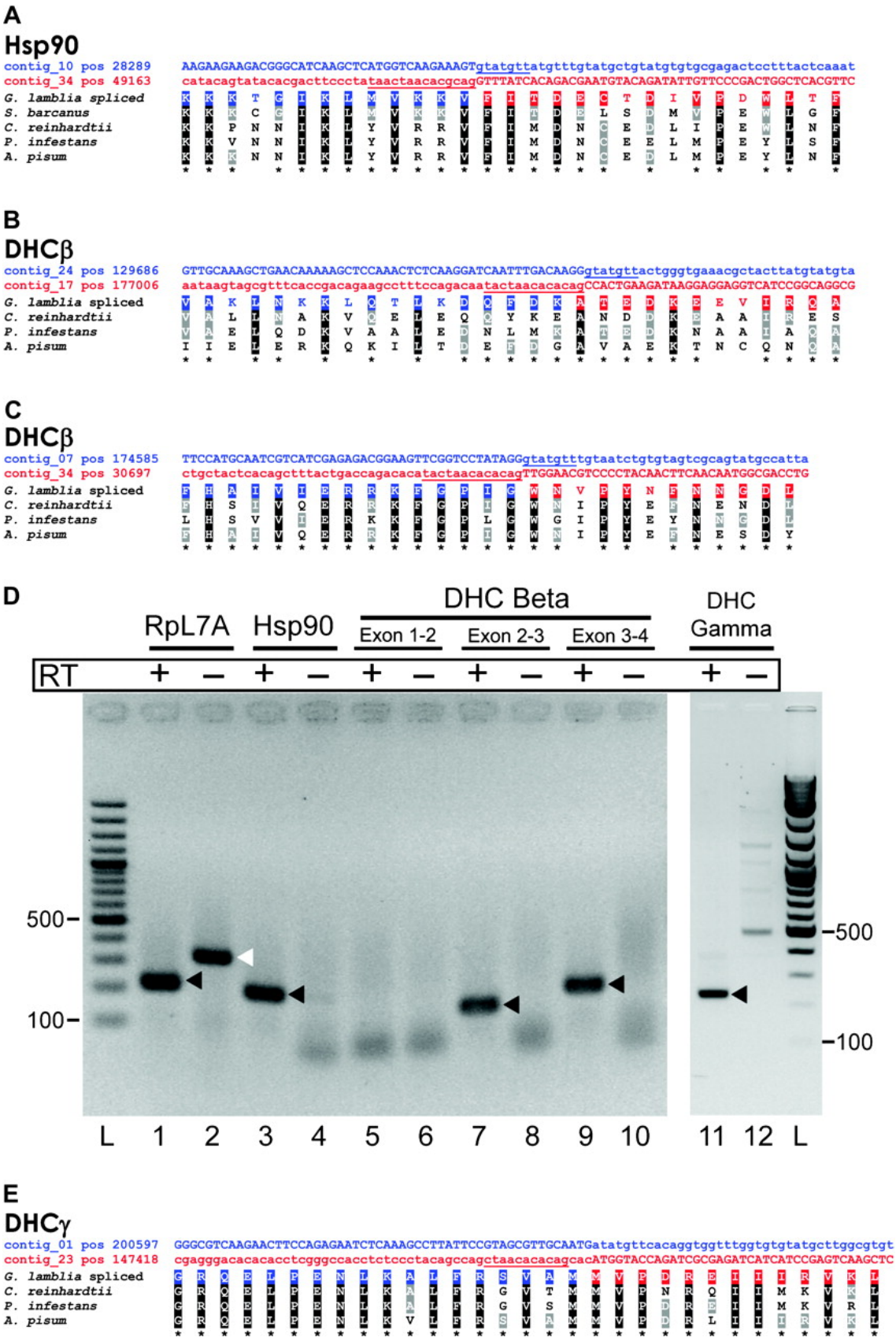


Figure 2.2. *Trans*-splicing in *Giardia lamblia*.

Figure 2.2. *Trans*-splicing in *Giardia lamblia* (continued).
Parts (**A–C**) and (**E**) show DNA and translated amino acid sequences for *Giardia trans*-splicing pairs, with ClustalW generated alignments of translated protein sequences from orthologous genes. Exonic/intronic sequence is shown in capital/lowercase. *Hsp90* (**A**), *DHC β* outer-arm exons 2–3 (**B**) and exons 3–4 (**C**), and *DHC γ* outer arm (**E**) are shown. Canonical *Giardia* intron boundaries are underlined. In each case, "contig_X" indicates the contig with name AACB020000X. GenBank accession numbers and extended protein sequence alignments are given in Figure A.1.2 to A.1.4 (**D**) Confirmation of *trans*-splicing. RT-PCR products generated from *Giardia* total RNA were resolved by 2% agarose gel electrophoresis and visualized by ethidium bromide staining (gel image inverted for better visualization). Filled arrowheads indicate spliced cDNA products generated from *cis*-spliced *Rpl7a* or *trans*-spliced *Hsp90* or DHC isoform mRNAs. The open arrowhead indicates a genomic product generated by PCR amplification of the intron-containing *Rpl7a* gene. RT indicates whether reverse transcriptase was added (+) or omitted (−) during the cDNA synthesis step. L lanes contain 100 bp DNA size ladders. Lanes 3 and 4 use primers (oAH 1 + 4) to detect fusion of exons 1 and 2 in *Hsp90*. For *DHC β*, lanes 5 and 6 assess exon 1–2 splicing (oAH 5 + 8), lanes 7 and 8 assess exon 2–3 splicing (oAH 9 + 12), and lanes 9 and 10 assess exon 3–4 splicing (oAH 13 + 16). Lanes 11 and 12 assay splicing of the two *DHC γ* exons (oAH 32 + 35). See Figure A.1.5 and Appendix IV for complete primer descriptions.

2.3.2 Extensive fragmentation of the DHC *β* gene

Further analysis of genomic regions encoding the *DHC β* outer-arm protein was even more surprising. The conserved gene is remarkably fragmented, with four putative exons encoded on internal regions of four separate genomic contigs. Regions of disjunction between exons 2 and 3 and between exons 3 and 4 exhibit obvious candidate splicing consensus sequences, and the putative *trans*-spliced product gives clear ungapped alignments to homologs (Figure 2.2B and C and Figure A.1.3). In contrast, no candidate splice sites for the junction between putative exons 1 and 2 were observed. PCR-mediated amplification of *Giardia* genomic DNA confirmed the fragmented organization of both *DHC β* and *Hsp90*: Primer combinations flanking each predicted exon–intron boundary generated expected products (Figures 2.3 and A.1.5), but primers flanking each split intron did not.

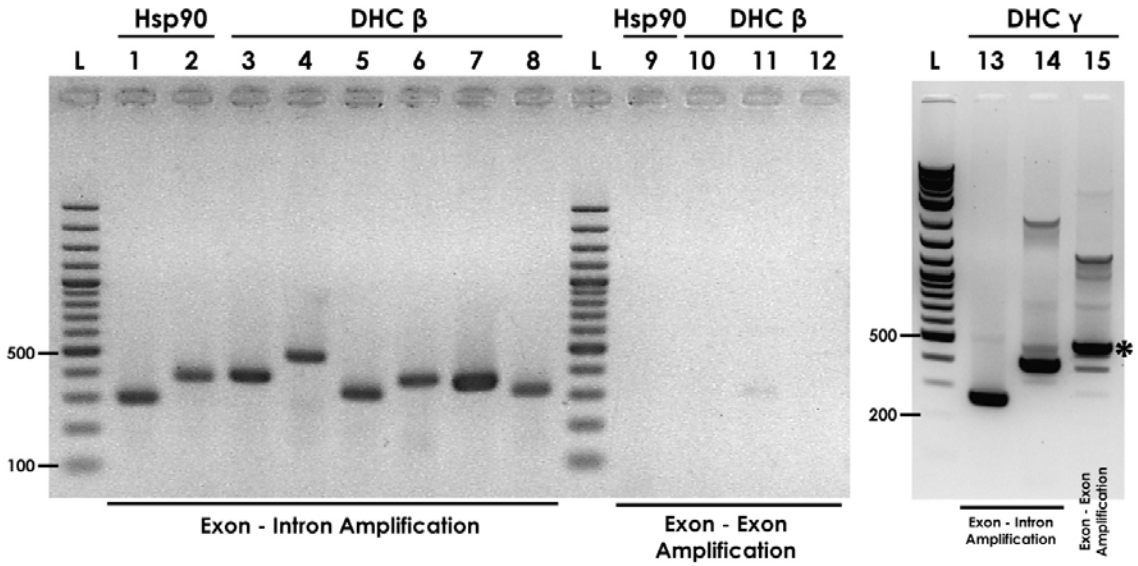


Figure 2.3. Verification of the genomic organization of Hsp90 and dynein gene fragments by polymerase chain reaction.
PCR products generated from *Giardia* genomic DNA were separated by 2% agarose gel electrophoresis and visualized by ethidium bromide staining. Lanes 1-8 and 13-14 are genomic products produced when employing oligonucleotide primers flanking exon-intron boundaries (i.e., configuration in the assembled genomic sequence) as diagrammed in Figure A.1.5. Lanes 9-12 and 15 are genomic DNA amplifications performed when using a primer targeted to an upstream exon, relative to mature mRNA structure, combined with a primer targeted to the next adjacent downstream exon (i.e., spliced configuration). Primer combinations used in each lane are: **1**. oligonucleotides oAH1 + oAH2 (*Hsp90* Exon 1-Intron 1 amplification), **2:** oAH3 + oAH4 (Intron 1-Exon 2 amplification), **3**. oAH5 + oAH6, **4**. oAH7 + oAH8, **5**. oAH9 + oAH10, **6**. oAH11 + oAH12, **7**. oAH13 + oAH14, **8**. oAH15 + oAH16, **9**. oAH1 + oAH3, **10**. oAH5 + oAH8, **11**. oAH9 + oAH12, **12**. oAH13 + oAH16. **13.** oAH32 + oAH33 **14.** oAH34 + oAH35 **15.** oAH32 + oAH35. Lanes **L** contain a 100 bp DNA size ladder. Refer to Figures A.1.5 and Appendix IV for primer annealing positions and sequences. The band indicated by an asterisk (*) in lane 15 was sequenced and determined to result from non-specific annealing of primers and amplification of an unrelated *G. lamblia* non-dynein gene.

2.3.3 Confirmation of *trans*-splicing

To assess *trans*-splicing, we performed RT-PCR on *Giardia* RNA using primer pairs within conserved exon regions, spanning positions of discontinuity (Figures 2.2D and A.1.5). As predicted, we observed RT-PCR products (confirmed by cloning and sequencing) for *trans*-splicing of *Hsp90* and for *DHC β* exons 2/3 and 3/4 (Figure 2.2D, lanes 3, 7, and 9) but not for *DHC β* exons 1/2 (lane 5). Notably, when the cDNA synthesis step was omitted (RT-), a genomic product was generated for the *cis*-spliced *Rpl7a* intron

control (presumably reflecting some genomic DNA in the RNA sample, lane 2) but not for exons flanking *trans*-spliced introns (lanes 4, 8, and 10), as expected for distantly-located/unlinked exons. *Trans*-splicing reactions involving *DHC* exons 2, 3, and 4 were found to be specific: no exon 2–exon 4 or exon 3–exon 2 splicing products were observed when using primer combinations that would allow their amplification (data not shown). Evidence that *DHC β* "exon 1" is instead independently translated (rather than *trans*-spliced) comes from 11 corresponding polyadenylated ESTs (Table A.1.1), and a canonical *Giardia* poly(A) processing signal (AGT[A/G]AA[C/T]) (Franzen *et al.* 2013) following the genomic sequence for exon 1 (but not exons 2 or 3). Interestingly, poly(A) signals were not identified for *DHC β* exons 2 and 3, however, several ESTs confirm polyadenylation of exon 4 at a poly(A) signal downstream of the termination codon (data not shown).

2.3.4 Utilization of atypical splice boundaries in the DHC γ gene

The case of the *DHC β* outer-arm gene stimulated us to examine other dynein genes. We found that the *DHC γ* outer-arm gene is also fragmented on two distinct *G. lamblia* contigs. The position of discontinuity lies between the first two conserved AAA domains (ATPase), P1 and P2 (Figure 2.4), mirroring the exon 1/2 split in *DHC β*. RT-PCR (Figure 2.2D, lane 11) and sequencing of multiple cDNA clones revealed *trans*-splicing. Surprisingly, cloning and sequencing revealed that splicing consistently occurred at a pair of atypical splice sites (Figure 2.2E). First, the 5′ splice site is ATATGTT. Second, splicing consistently occurred not at the observed in frame consensus 3′ splice site (GCTAACACACAG) but at an "AC" dinucleotide sequence 3 nt downstream (GCTAACACACAGCAC). Again, the spliced sequence yields a clear ungapped alignment to homologous sequences (Figures 2.2E and A.1.4). Thus, the *DHC γ* outer-arm gene is interrupted by a *trans*-spliced AT–AC intron.

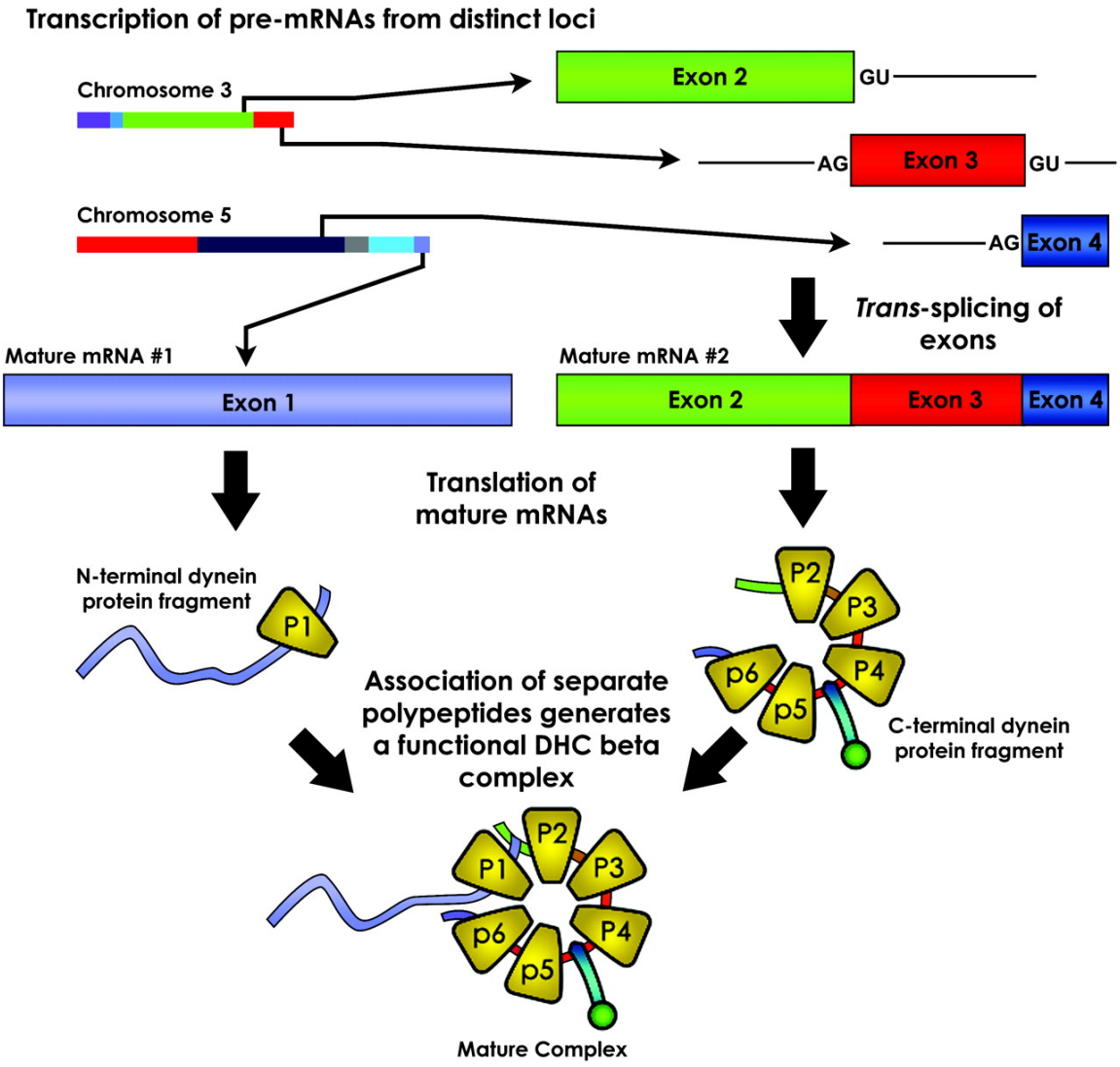


Figure 2.4. Model for the expression of the *Giardia lamblia DHC β* outer arm.
DHC β outer-arm gene fragments are independently transcribed from distant locations on different chromosomes in *G. lamblia*; scaffolds obtained by genome sequencing of the WB isolate are indicated by different colors (Upcroft *et al.* 2010). *Trans*-splicing events fuse the last three exons encoding the C-terminal half of the protein. Independent translation generates the N- and C-terminal polypeptides, which are then predicted to associate to produce an essential and a functional DHC. The six conserved AAA domains, P1–P6 form a conserved hexameric "wheel-like" structure, the so-called motor domain (Asai and Koonce 2001). The *Giardia DHC γ* outer-arm gene is also fragmented within the linker region between P1 and P2, but in this case, *trans*-splicing fuses the coding region to generate a continuous polypeptide.

It is tempting to speculate that *DHC β* also once contained a fragmented intron between P1 and P2. The fragmented intron could then have lost the ability to *trans*-splice,

possibly stimulated by the ability of the two protein halves to assemble posttranslationally to form a functional DHC, leading to the observed protein-level fragmentation. Split dynein genes are very unusual. The only known case is in the basidiomycete fungus *Ustilago maydis*, where a dynein is fragmented between the P4 and P5 domains (refer to Figure 2.4), and the two polypeptides were shown to form a complex *in vivo* (Straube *et al.* 2001). Notably, other non-dynein members of the superfamily of AAA-type (ATPase) domain-containing proteins typically contain only one or two AAA domains per polypeptide and can oligomerize to form homohexameric rings akin to the continuous P1–P6 ring structure seen in most dyneins (Asai and Koonce 2001).

The coupling of 5′ AT and 3′ AC splice boundaries is intriguing because AT–AC boundaries are frequently found in the rare U12-type/minor spliceosomal introns. Indeed, the spliceosomal snRNAs identified in *G. lamblia* also contain mixed features of both U2-dependent/major and U12-dependent/minor spliceosomal snRNAs (Hudson *et al.* 2012) (See Chapter 3). Based on these observations, we speculate that an ancestor of *G. lamblia* may have possessed both U2 and U12-dependent spliceosomal systems which converged to form a "hybrid" U2/U12 spliceosome in *G. lamblia* (Chapter 3). However, the fragmented *DHC* γ AT–AC intron is still predicted to be a major/U2-type spliceosomal intron because the extended splicing boundaries of this intron are more similar to U2-type/major consensus boundaries than to U12-type consensus sequences. In addition, sequence searches have failed to identify any U12-dependent spliceosomal protein components in *Giardia* (or any other intron-poor genome) (Russell *et al.* 2006, Lopez *et al.* 2008).

2.3.5 Base pairing and *trans*-splicing: evolutionary implications

We next asked how the various intron halves would associate *in vivo*. We discovered striking base-pairing potential between intron halves, with each stem stabilized by at least 17 Watson–Crick base pairs (Figure 2.5A–C), as observed for *Hsp90* (Nageshan *et al.* 2011). In each case, the 5′ boundary of intron sequence complementarity begins 10–13 nt downstream of the 5′ splice site and the 3′ boundary ends 25–34 nt upstream of the 3′ splice site, thereby leaving all the conserved splicing elements potentially exposed as "single-stranded" regions. Interestingly, the two longest previously identified *cis*-spliced *G. lamblia* introns also show extensive base-pairing potential with complementary regions in the same locations relative to splice sites (Figure 2.5D and E), suggesting that base pairing may be important for efficient splicing of *Giardia* introns (both *trans* and *cis*), perhaps by constraining the spatial distance between the splice sites, consistent with three of the known *Giardia cis*-introns being very short (29–35 nt). Study of gene sequences from other *Giardia* isolates (Figures A.1.6 and A.1.7) confirmed preferential conservation of the complementary regions: 1) complementary regions showed a higher rate of sequence conservation than flanking intronic sequences or than synonymous sites within flanking exons; 2) changes within predicted secondary structures were more common at unpaired sites (4 changes at 11 unpaired sites vs. 11 changes at 178 paired sites; $P = 0.0057$ by a Fisher Exact Test); and 3) all substitutions at paired sites preserved base pairing (Figure A.1.7).

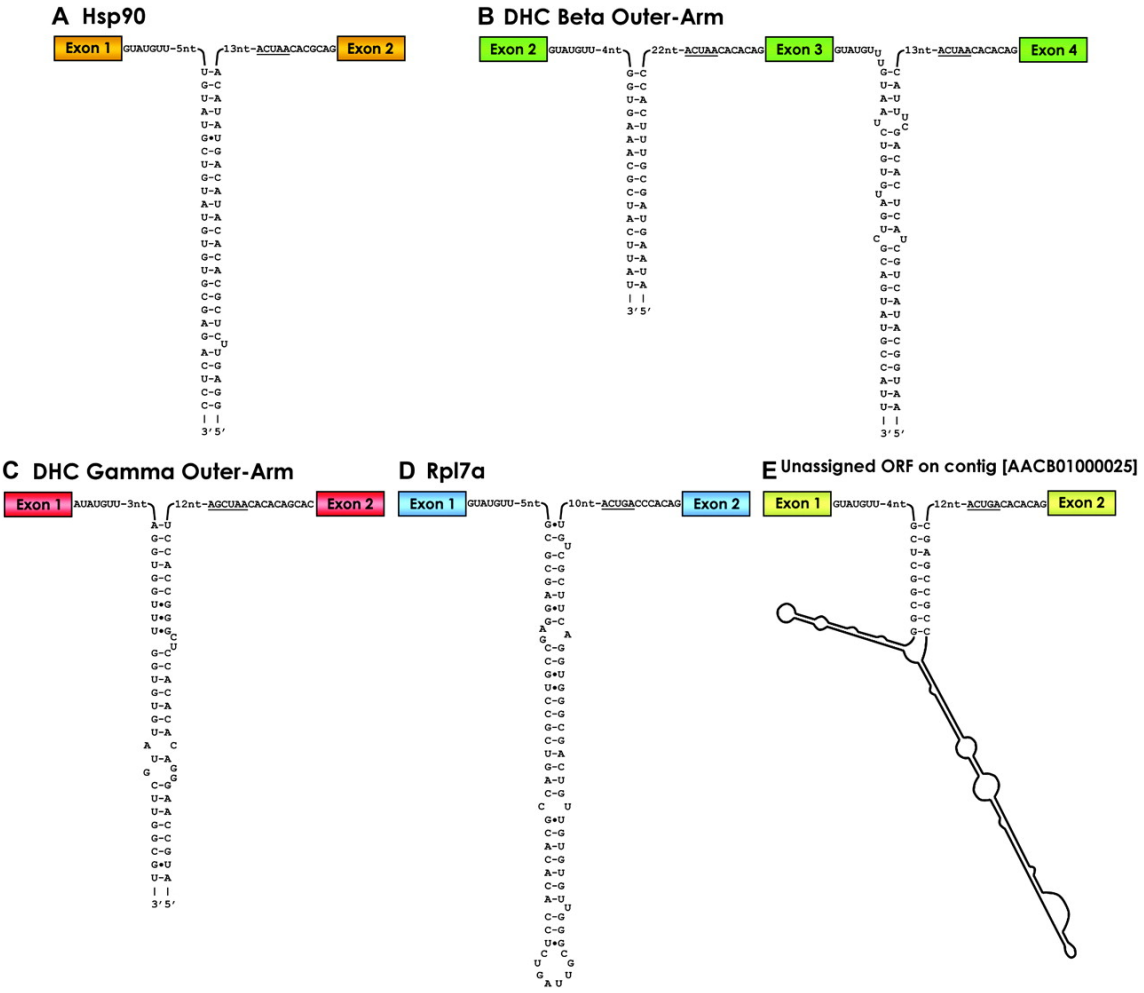


Figure 2.5. Extensive base-pairing potential in *Giardia* spliceosomal introns.
Predicted secondary structures illustrate the extensive base pairing predicted to mediate *in vivo* association of intron halves and *trans*-splicing of fragmented introns in Hsp90 and DHC β and γ outer-arm mRNAs (**A–C**). Predicted branch point sequences are underlined; the 3′ ends of intron 5′ halves and likewise, the 5′ ends of intron 3′ halves are indeterminate. Also illustrated are secondary structures predicted by MFOLD (Zuker 2003) for two previously identified *cis*-spliced *Giardia* introns in the *Rpl7a* gene and a gene encoding a protein of unknown function (**D** and **E**). Due to its larger size, only the sequences of the intron extremities are shown in (**E**).

Base-paired stems of long *cis*-spliced introns also suggest a model for recurrent intron fragmentation (four of nine introns identified so far). Pre-established extensive base pairing within contiguous (*cis*) introns would provide the means for immediate association of intron halves following exon dispersal. A second important feature may be *Giardia*'s strikingly degenerate and short promoters, comprising as little as 10–16 nt of AT-rich sequence (Yee *et al.* 2000, Yee *et al.* 2007). Such sequences might be expected to

frequently arise at random within intronic sequence upstream of a potential 3′ splice site. Such cryptic internal promoters could allow for immediate transcription of the downstream exon following genomic fragmentation.

2.4 Conclusions

We report the first case of multiple fragmented spliceosomal introns interrupting a single gene. We have identified the most fragmented nuclear protein–coding gene expression pathway in any organism characterized to date, with fragmentation of DHC β outer arm occurring at both the transcriptional and translational level. We propose that the common gene fragmentation position now found in both DHC outer arms β and γ was initially the result of both genes containing *cis*-introns at the same relative location (between P1 and P2) that later became fragmented and required *trans*-splicing (still seen in γ), the β isoform now having lost the intron completely. This represents a novel hypothetical pathway of spliceosomal intron loss that may occur in some eukaryotes undergoing genome size reduction. This also raises the tantalizing possibility that such a mechanism of gene fragmentation may have contributed to the evolution of multi-subunit protein complexes from simpler (monomeric) units. *Giardia* is now distinguished in having an unusually large fraction of split introns, evolutionarily-divergent spliceosomal small nuclear RNAs, and frequent occurrence of non-canonical 5′ intron terminal nucleotides (CT/AT), clearly a novel splicing system that warrants further characterization. The finding of unprecedented complexity of gene expression in such a radically reduced genome underscores the interplay between genomic “simplification” and intricacy in genome evolution.

Chapter 3: Evolutionarily Divergent Spliceosomal snRNAs and a Conserved Non-coding RNA Processing Motif in *Giardia lamblia*

Reprinted with permission from:

"Evolutionarily divergent spliceosomal snRNAs and a conserved non-coding RNA processing motif in *Giardia lamblia*"

Nucleic Acids Research **2012**. *40*, 10995-11008.

Andrew J. Hudson, Ashley N. Moore, David Elniski, Joella Joseph, Janet Yee and Anthony G. Russell

Contributions:

AJH co-designed the study, performed experiments and bioinformatic analyses and wrote the first manuscript draft. ANM and DE assisted with bioinformatic identification of ncRNAs as well as primer extension and northern blot experiments. JJ and JY cultivated *G. lamblia* and isolated nucleic acid samples for the study. AGR co-designed the study, made substantial intellectual contributions and completed the final manuscript draft. All authors read and approved the final manuscript.

Changes Incorporated:

Experiments performed after publication further characterizing the *G. lamblia* RNA motif processing pathway were added. Also, subsequent to publication, the GlsR28 ncRNA was found to contain features of telomerase RNA and this analysis has been incorporated. Some wording was changed to reflect findings in the other chapters.

3.1 Introduction

Eukaryotic precursor (pre-)RNA processing often requires ribonucleoprotein (RNP) complexes consisting of conserved and essential non-coding (nc)RNAs. Notable examples are the small nucleolar (sno) RNPs that participate in eukaryotic ribosome biogenesis through structural modification of specific nucleotides in ribosomal RNA (rRNA) and/or targeting cleavage of the pre-rRNA (Chow *et al.* 2007, Henras *et al.* 2008, Watkins and Bohnsack 2012). Another prevalent eukaryotic RNA processing event is mRNA splicing—the removal of intervening intron sequences from pre-mRNAs that is

catalysed by the dynamic RNP complex termed the spliceosome (Wahl *et al.* 2009). The vast majority of spliceosomal introns are classified as major-type (U2-type) and are removed by the major (U2-dependent) spliceosome. The U2-dependent spliceosome consists of five evolutionarily conserved small nuclear (sn) RNAs, U1, U2, U4, U5 and U6, and potentially hundreds of associated proteins (Jurica and Moore 2003, Valadkhan and Jaladat 2010). Spliceosome-mediated intron recognition and excision requires intricate base-pairing interactions between the snRNAs and conserved intron boundary and internal branch-point sequences and numerous snRNA–snRNA intermolecular base pairings, dynamically changing during the splicing cycle (Will and Luhrmann 2011).

Although nearly all examined eukaryotic genomes seem to contain major spliceosomal introns, a much smaller subset of eukaryotic organisms also possess a rare class of minor (U12-type) spliceosomal introns, which are excised by a distinct minor (U12-dependent) spliceosome (Patel and Steitz 2003, Will and Luhrmann 2005). The U12-dependent spliceosome contains a unique set of snRNAs, U11, U12, U4atac, U6atac that are functionally analogous to the U1, U2, U4 and U6 snRNAs, respectively, but shares the U5 snRNA also found in the U2-dependent spliceosome. Features shared between the U2-dependent and U12-dependent spliceosomes, including U5 and some common core protein constituents, and secondary structural similarities of the snRNAs, may indicate a common ancestral origin for both spliceosomes (Patel and Steitz 2003). The evolutionarily distant relationship of the limited number of species known to possess a U12-dependent minor spliceosome indicates its early origin in eukaryotes (Russell *et al.* 2006, Lopez *et al.* 2008, Bartschat and Samuelsson 2010). Based on detailed analysis of ancient intron insertion sites, Basu *et al.* 2008, proposed that major U2-type introns pre-dated the existence of U12-type minor spliceosomal introns, and the observation that all organisms containing a U12-

dependent spliceosome also have a U2-type major spliceosome is not inconsistent with this idea. Furthermore, it is hypothesized that spliceosomal introns and components of the spliceosomal machinery are derived from group II introns, based on observation of regions of similar snRNA and intron structure, and splicing reaction mechanism (Zhang and Doudna 2002, Sashital *et al.* 2004, Seetharaman *et al.* 2006, Michel *et al.* 2009). Identification of any extant organisms possessing splicing systems and introns with features characteristic of transition stages in such evolutionary pathways would help to further evaluate these models for intron evolution.

The diplomonad protist *Giardia lamblia* is a prevalent human enteric parasite that displays a highly reduced compact genome and somewhat limited metabolic capacity (Morrison *et al.* 2007). To date, only nine spliceosomal introns have been identified in *G. lamblia*, and they exhibit extended highly conserved 5′ splice sites, and atypical fused branch point and 3′ splice sites (Nixon *et al.* 2002, Russell *et al.* 2005, Morrison *et al.* 2007). Our group (Roy *et al.* 2012) and others (Kamikawa *et al.* 2011, Nageshan *et al.* 2011) identified several cases of *trans*-splicing of these *Giardia* spliceosomal introns (four of the nine characterized introns). In this *trans*-splicing pathway, exons dispersed to distant regions of the genome are expressed as distinct pre-mRNA transcripts that somehow associate to mediate exon ligation. Determining the mechanistic details of how this occurs will require identification and characterization of *Giardia* spliceosomal components and potentially other required RNA processing complexes. Association of individually transcribed exon–intron containing pre-mRNA precursors is predicted to occur through base-pairing potential evident in respective introns halves (Roy *et al.* 2012) somehow positioning intron splice sites for recognition by the *Giardia* spliceosome.

Identification of spliceosomal introns and putative core spliceosomal proteins in *G. lamblia* (Nixon *et al.* 2002) strongly argues for the existence of a functional spliceosome in this organism. The *Giardia* spliceosomal snRNAs have been elusive and previously, putative *G. lamblia* U1, U2, U4 and U6 spliceosomal snRNA candidates were predicted computationally by examining the *Giardia* WB strain, the only one for which extensive genomic DNA sequence information was available at that time (Chen *et al.* 2008). These candidates were structurally divergent, and our recent search for orthologues of these snRNA candidates in the genomic sequences now available from the two related *G. lamblia* isolates (see later in the text) (Franzen *et al.* 2009, Jerlstrom-Hultqvist *et al.* 2010) reveals extensive unexpected sequence variation, including nucleotide substitutions disrupting critical and strictly evolutionarily conserved sequence motifs and secondary structures that are fundamental to spliceosome function in other eukaryotes.

In our study, we have taken advantage of the genomic sequence information available for three different *Giardia* isolates (strains)(Morrison *et al.* 2007, Franzen *et al.* 2009, Jerlstrom-Hultqvist *et al.* 2010), non-coding RNA sequence information obtained from previously constructed cDNA libraries and our discovery of a conserved RNA processing motif, to identify and characterize new *Giardia* ncRNAs. This includes the identification of a new set of spliceosomal snRNAs that show strict conservation of functionally important sequence elements in all three isolates and compensatory mutations maintaining predicted secondary structures.

3.2 Materials and Methods

3.2.1 RNA motif identification and characterization

Genomic regions encoding biochemically isolated *G. lamblia* WB isolate ncRNAs (Yang *et al.* 2005, Luo *et al.* 2006, Chen *et al.* 2007, Chen *et al.* 2011) were identified by BLASTN searches using the GiardiaDB website (www.GiardiaDB.org). For each region, 300 nt of additional upstream and downstream flanking genomic sequence was then analysed by manual inspection for any conserved sequence elements evident when aligning the collection of genomic regions. This analysis revealed the presence of a conserved 12 nt motif residing adjacent to or overlapping genomic regions encoding documented (or predicted) mature ncRNA 3′ ends. Homologous ncRNA-encoding regions in the *Giardia* GS and P15 isolate genome sequences were identified using the WB sequences as BLASTN queries. After aligning all sequences using ClustalW2 (Larkin *et al.* 2007) to identify all motif sequences in ncRNA genes and those also found in *trans*-spliced intronic regions, a motif consensus frequency plot (n = 135 sequences) was generated using WebLogo software (Crooks *et al.* 2004). Secondary structures for ncRNAs plus their downstream motifs were predicted using MFOLD (Zuker 2003) or RNAalifold (Hofacker 2003).

3.2.2 Identification of new *Giardia* ncRNAs

Our search strategy for new ncRNAs exploited several emergent properties of *Giardia* ncRNA genes: (i) they are usually located in intergenic regions between open reading frames (ORFs); (ii) the conserved RNA sequence motif is located in their 3′ downstream flanking region; and (iii) many *Giardia* ncRNA sequences are preceded by A-T rich genomic sequence elements that are predicted initiation sites (Yang *et al.* 2005) based on similarity to the transcription initiation sites of *G. lamblia* protein-coding genes (Yee *et al.* 2000, Elmendorf *et al.* 2001). Initially, we performed simple BLASTN searches

using the more common individual RNA motif sequence variants as queries against the WB genome and searched for instances that also contained upstream A-T rich sequences. This analysis unexpectedly identified motif sequences in the downstream regions of all four previously characterized *Giardia trans*-spliced intron 5′ halves (the 3′ ends of which were previously unknown) and also motif sequences within four putative ORFs, but this method was not efficient for identifying new ncRNAs. Next, we utilized the pattern matching program 'Scan for Matches' (Dsouza *et al.* 1997). Whole genome sequences for *G. lamblia* WB, P15 and GS isolates (provided at GiardiaDB.org) were used as local databases in 'Scan for Matches' searches using the following parameters: AAAAAAAAAA (allowing five mismatches)… 1–500 nt … CCTTYNHTHAA, where 'Y' is a pyrimidine, 'H' is A, C or T nucleotide and 'N' is any nucleotide. This scan yielded ~400 matches in each *G. lamblia* isolate, which were further screened using pairwise BLASTN comparisons of each *G. lamblia* WB match against the matches from the P15 and GS genomes. Only instances where the promoter and motif sequence elements were present in corresponding genomic regions in all *Giardia* isolates, and the matches also mapped to intergenic regions, were deemed probable ncRNA candidates and were further considered. The candidates were then inspected for other hallmark sequence elements (e.g. conserved box C/D and H/ACA sequences for snoRNAs; conserved snRNA elements, such as the 5′ splice site binding sequence for U1, ACAGAGA sequence for U6) and overall secondary structures (MFOLD and RNAalifold) to classify their function. This strategy identified the GlsR26 and GlsR27 box H/ACA snoRNAs, the GlsR28 ncRNA of unknown function and the new U1 and U6 snRNA candidates.

The novel *G. lamblia* U6 snRNA candidate then permitted prediction of new U2 and U4 snRNA candidates through its evolutionarily conserved ability to form extensive

intermolecular base pairs with these other snRNAs. Many previously identified *Giardia* ncRNAs have no assigned function (Chen *et al.* 2007); therefore, we reasoned that some of these may be snRNA homologues. From these, we generated a concatenated sequence file appropriate to serve as a library for our searches.

To identify the U4 snRNA, *Giardia* U6 nucleotides C21 to U60 are those predicted to be involved in U6/U4 snRNA base pairing and were, therefore, used as query in BLASTN searches of the concatenated ncRNA file, increasing expect thresholds to 10^4 to optimize search sensitivity for short sequences. This revealed extensive complementarity between the *G. lamblia* U6 snRNA and ncRNA 'Candidate'-11 (Chen *et al.* 2007), implicating it as a potential U4 snRNA candidate. Further MFOLD analysis and manual sequence inspection showed that Candidate-11 could form conserved intermolecular helices I and II with the U6 snRNA candidate and also a canonical U4 snRNA 5′ stem-loop (SL). These findings, in combination with compensatory mutations in *Giardia* GS isolate intermolecular helix II (Figure 3.10B) that maintain U6/U4 snRNA base pairing provided the evidence that Candidate-11 is the *G. lamblia* U4 snRNA.

Identification of a U2 snRNA candidate is more challenging because of the short and discontinuous base pairing between U2 and U6 snRNAs. Here, we used the Spin program (Staden freeware package, 1996) to search the concatenated library for any ncRNA that had the ability to form base pairs with the extended branch-point sequence found in *G. lamblia* introns (AACTAACAC, branch point 'A' underlined). This search revealed that uncharacterized ncRNA Candidate-14 (Chen *et al.* 2007) contains nucleotides '$_{26}$GUGUAGUU$_{33'}$ that are able to form extensive base pairs with the intron branch point with 'bulged' adenosine nucleotide configuration. Further analysis revealed that Candidate-14 could form canonical intermolecular helices I through III with the U6 snRNA, with

regions of pairing occurring at the same relative positions as other representative eukaryotic U2/U6 snRNA complexes. MFOLD analysis predicted U2-like helices IIa, III and IV in the 3′ half of Candidate-14, further implicating it as the *Giardia* U2 snRNA.

3.2.3 RT-PCR experiments

G. lamblia WB strain (clone 6, ATCC 30957) axenic trophozoites were cultured in modified TYI-S-33 medium, and *Giardia* genomic DNA and total RNA were extracted using DNeasy Kits (Qiagen) and TRIZOL Reagent (Invitrogen), respectively, according to the manufacturer's instructions. Reaction conditions used for reverse transcriptase (RT) and polymerase chain reaction (PCR) experiments on *Giardia* nucleic acid samples were performed as previously described (see section 2.2.2) unless otherwise specified, using oligonucleotide primers in Appendix IV.

RT-PCR amplification of branched *G. lamblia trans*-spliced introns was performed using methods described earlier (see section 2.2.2), however with some modifications. For the RT step, reverse primers were designed to anneal either downstream (primer R1) or upstream (R2) of motif sequences present in *trans*-spliced intron 5′ halves (see Figures 3.4A and A.2.5 and Appendix IV for primer sequences). Resulting cDNAs were then amplified in PCR reactions containing forward primers annealing upstream of BP sequences in corresponding *trans*-intron 3′ halves (primer F1) and reverse primers used for cDNA synthesis. Because this initial round of PCR generated very weak products, a second nested PCR reaction was performed using 5 μL (out of 50 μL total) of product from the first PCR as template, a nested forward primer (F2) binding downstream of the F1 primer on *trans*-intron 3′ halves and the same reverse primer used during RT and first PCR steps. Products from the second nested PCR were then cloned into vectors and subjected to

automated sequencing as described earlier (section 2.2.2). Branch point nucleotides were then determined by examining the first nucleotide upstream of 5′ SS sequences in clones.

3.2.4 Primer extension and northern blotting

Cellular expression of candidate ncRNAs was verified by northern blot (snRNAs only), and primer extension analysis (all ncRNAs) using ncRNA-specific reverse primers that anneal ~10 nt upstream of the conserved ncRNA motif sequences (see Appendix IV). Primers were 5′ end radiolabeled in 30 μL reactions containing 10 U T4 polynucleotide kinase (PNK, New England Biolabs), 1 X T4 PNK Buffer (70 mM Tris-HCl, 10 mM $MgCl_2$, 5 mM dithiothreitol (DTT), pH 7.6), 10 μCi [γ-^{32}P] ATP (3000 Ci/mmol, Perkin-Elmer) and 30 pmol gel-purified deoxyoligonucleotide primer. Labeling reactions were incubated at 37°C for 1 hour followed by phenol-chloroform extraction and ethanol precipitation of DNAs in the presence of 5 μg of linear polyacrylamide carrier. Precipitated DNAs were then resuspended in ddH_2O. Primer extensions were perfomed by incubating 1 pmol [^{32}P] 5′ end-labelled oligonucleotide primer (~10,000 cpm) with 10 μg *Giardia* total RNA in a 10 μL volume at 65°C for 10 minutes and then 47°C for 5 minutes. The remaining components were then added to constitute 20 μL reaction volumes containing: 1 X RT reaction buffer (50 mM Tris-HCl, 75 mM KCl, 3 mM $MgCl_2$, pH 8.3), 10 mM DTT and 200 U SuperScript II™ reverse transcriptase (Life Technologies) and reactions were incubated at 47°C for 1 hour. Msp I restriction enzyme-digested pBR322 plasmid fragments were ^{32}P end-labelled as described above and served as size standards for primer extension and northern blot experiments. Primer extension products, pBR322 Msp I digests and total *Giardia* RNAs (10 μg) for northern blotting were resolved by 8% urea polyacrylamide gel electrophoresis (denaturing PAGE). RNAs were then transferred to Amersham Hybond™-XL membranes using a Bio-Rad Trans Blot Cell apparatus,

according to the manufacturer's instructions. Radioactive gels and membranes were visualized using a GE Healthcare Typhoon phosphorimager.

DNA probes used for northern blots were created by PCR amplification of snRNA coding regions (as described in section 2.2.2) and radioactively [^{32}P] 5′ end-labelled as described above, using ~200 ng PCR product during labeling reactions. DNA probes (~100,000 cpm) were then heat denatured at 100°C for 5 min, placed on ice for 2 min and then added to RNA-cross-linked (northern blot) membranes that had been pre-equilibrated with hybridization buffer (500mM $NaPO_4$, pH 7.2, 7% SDS, 1 mM EDTA) and incubated at 68°C for 16 hours. Northern blot membranes were then washed once with 1 x SSC (150 mM NaCl, 15 mM sodium citrate), 0.1% SDS at 21°C for 10 min and then twice with 0.5 X SSC, 0.1% SDS at 68°C for 10 min. Membranes were then dried and visualized by phosphorimaging.

3.2.5 RNA end-mapping by random amplification of cDNA ends (RACE)

Mapping of mature ncRNA ends was performed using random amplification of cDNA ends (RACE) techniques. For ncRNA 3′ RACE, *Giardia* total RNA (2 μg) was polyadenylated with 5 U poly A polymerase (NEB) in 1 X PAP buffer (50 mM Tris-HCl, 250 mM NaCl, 10 mM $MgCl_2$) and 0.5 mM rATP at 37°C for 1 hour. First strand cDNAs were then generated using *Giardia* poly A RNA (2 μg) and oligonucleotide P-94 (Appendix IV) as reverse primer during reverse transcription using methods described earlier. Gene-specific forward primers and P-94 reverse primers were then used during PCR to amplify mature RNA 3′ ends.

To map 5′ ends and for detection of 5′ cap structures, we also performed RNA linker-mediated (RLM) 5′ RACE. Thirty micrograms of DNase I-treated *G. lamblia* total RNA was divided equally into three different samples as follows: (U) untreated, (C) treated

with 20 U calf intestine alkaline phosphatase (CIP, New England Biolabs (NEB)) or (CT) treated with 20 U CIP and subsequently treated with 10 U tobacco acid pyrophosphatase (Interscience). Following this, for each sample, RLM-5′ RACE oligomers were ligated onto available RNA 5′ ends in 100 µl reactions containing 10 µg *Giardia* treated RNA sample (aforementioned), 3 µg RLM-5′ RACE linker oligo, 1 mM adenosine triphosphate, 50 U T4 RNA ligase I (NEB), 1× RNA ligase buffer (NEB) and 20% wt/vol polyethylene glycol 8000. Ligation reactions were incubated for 1 h at 37°C and then used directly as the template for reverse transcriptase-polymerase chain reaction (RT-PCR) using adaptor-specific forward primers (PCR step) and ncRNA-specific reverse primers (RT and PCR step). Products generated during either RT-PCR or RACE experiments were agarose gel purified using eZNA Gel Extraction Kits (Omega Biotech) (when multiple amplicons were present) or directly cloned into the CloneJet vector (Fermentas) according to the manufacturer's protocol and subject to automated DNA sequencing (Macrogen).

3.2.6 *In vitro* U4/U6 snRNA complex formation

Regions encoding the mature *G. lamblia* WB isolate U4 and U6 snRNA candidate sequences, as determined by the end mapping experiments, were PCR-amplified from *Giardia* WB genomic DNA, using appropriate reverse primers and forward primers additionally containing the T7 viral promoter sequence. All PCR products were cloned and sequenced to verify their identities. Gel-purified PCR products then served as templates for *in vitro* transcription to generate unlabelled or [^{32}P] 5′ end-labelled transcripts using methods described above. *Giardia* U4 and U6 *in vitro* complexes were formed by assembling 20 nM radioactively labelled RNA transcript with 200 nM unlabelled RNA in the presence of 20 µM of each oligo oAH136 and oAH137, to optimize U4/U6 intermolecular base pairing (Brow and Vidaver 1995) in assembly buffer (50 mM NaCl, 20

mM HEPES, pH 7.0, 1.5 mM $MgCl_2$, 0.1 mM EDTA). Reactions (15 µl) were heated to 80°C for 2 min and were then allowed to slowly cool to room temperature and were placed on ice. Complexes were then resolved on 6% TBE native PAGE gels and visualized by phosphorimaging.

3.3 Results

3.3.1 Identification of a conserved sequence motif in *Giardia* ncRNA genes

ncRNAs display varying modes of genomic organization, expression and maturation in different eukaryotes. In *G. lamblia*, those ncRNAs identified to date are encoded as predicted single gene transcriptional units or as dicistronic gene clusters (Yang *et al.* 2005, Chen *et al.* 2007). However, the mechanisms underlying their expression and subsequent precursor transcript processing have yet to be examined. Consequently, we examined the genomic context of previously biochemically identified *Giardia* ncRNAs searching for conserved sequence elements that may be involved in their expression and/or processing. Genomic regions encoding previously annotated *G. lamblia* WB isolate box C/D and H/ACA snoRNAs, RNase MRP RNA, and other uncharacterized ncRNAs (Yang *et al.* 2005, Luo *et al.* 2006, Chen *et al.* 2007, Chen *et al.* 2011) were aligned and inspected for recurring sequence motifs (Table 3.1). Strikingly, this analysis uncovered a highly conserved 12 nt sequence motif overlapping or residing a few nucleotides downstream of the predicted 3′ ends of the mature RNAs. We also exploited the current availability of near-complete genome sequences of three *G. lamblia* isolates, WB, GS and P15, that display substantial sequence divergence (~77% nt identity between WB and GS in protein-coding regions) (Morrison *et al.* 2007, Franzen *et al.* 2009, Jerlstrom-Hultqvist *et al.* 2010). BLASTN searches using *G. lamblia* WB ncRNA genomic regions as queries readily

identifies orthologous P15 and GS genomic regions showing conservation of ncRNA sequences and in many cases even higher conservation of the 12 nt 3′ end sequence motif (Figure A.2.1).

The collection of 3′ end motif sequences from *Giardia* WB, GS and P15 genomes (n = 132 sequences) revealed the consensus: 5′-[T/A/C]C[C/A]TT[T/C][A/T/C][C/T/A]T[C/T/A]AA-3′ (Figure 3.1A). Thirty-nine unique variations of the sequence motif were identified, with 'TCCTTTACTCAA' being observed in 34/132 instances (Figure 3.1A, Table A.2.1). Motif variant prevalence is similar between *Giardia* isolates, and the motif displays strong sequence conservation with 6 of 12 positions being invariant. Because snoRNA 3′ end processing in some eukaryotes requires the formation of a SL structure (Chanfreau *et al.* 1998), we examined whether the identified *Giardia* motif may participate in the formation of such structures. MFOLD RNA secondary structure predictions (Zuker 2003) of motif variants either alone or in the context of adjacent flanking upstream and downstream sequences does not indicate significant secondary structural potential, and instead suggests the sequence motif may be exposed within single-stranded regions of an RNA primary transcript.

Table 3.1. A conserved 12 nucleotide sequence motif is located downstream of many ncRNAs and 5′ *trans*-spliced intron halves in *Giardia*

Comparison of DNA sequences surrounding ncRNA and *trans*-intron expressed sequences reveals a conserved motif located immediately downstream of mature RNA 3′ ends. Coding sequences (uppercase and bold) and flanking genomic sequences or intronic sequences (lowercase) for *Giardia* WB isolate ncRNAs and *trans*-spliced introns are shown with the conserved downstream sequence motif highlighted in green. Specific genomic locations for the displayed sequences are indicated as annotated for the *G. lamblia* WB isolate genome database. Predicted conserved snoRNA box elements for box C/D and H/ACA RNAs are highlighted in grey. *Trans*- spliced intron 5′ splice sites are underlined and intronic regions predicted to form base-pairing interactions with 3′ intron halves are those indicated in red text. NcRNA sequences identified by (Yang *et al.* 2005) [†], (Luo *et al.* 2006) [‡], (Chen *et al.* 2007) [ϕ] and (Chen *et al.* 2011) [Ω] are denoted. An asterisk [*] indicates ncRNAs identified in this study.

	Genomic Location (WB)	Genomic DNA Sequence Flanking Mature RNA 3′ Ends (WB isolate)
Box C/D RNAs		
GlsR1†	GLCHR03:412149-412256(+)	**TCTCCTGAGGCAGATGATGACTTTGCGACGGGCGGACGGAGGGACGCGTGACGAAGTTTGTCGTATTCTGAATTccttcatttaa**aattggg
GlsR2†	GLCHR05:2883797-2883904(+)	**GGCGATGGAGACAAAAGCAGTTACGTTCGCAACTCTCTGAGGGTTCCTGATGCTTCCTTGGATGTCCGAGCCT**tcctttacttaatcgaccg
GlsR4†	GLCHR05:2609094-2609201(-)	aaagtgaggcaggc**AGTCTCCATGACGAGAATTACGCCGCCCCAGTCTGACCCCTGACGAACGGCTTCTCTGATC**attcactcaacccgccg
GlsR5†	GLCHR03:1407731-1407838(-)	**ATGACAGGTTCTTGCCCCGTATGACCCTGCGATGAGTTATACAAAAGAACGCATCCAAGCCAACCGGCTGAGCTC**cttcactcaaatcctgc
GlsR6†	GLCHR03:957087-957194(-)	acgtaaaaaatgc**AATGATGGCTTGTTATCCCTGTCTGAGGTCAATACCTTGATTAGACGATTTGACAGAGC**atccttcactcaacaccctt
GlsR7†	GLCHR01:808626-808733(+)	tcttgccttcctc**CCGCGATGATTACCGAATCACAGCGATACACGATGAAGCACTCATAGTTACTCTGAGCGG**tcctttactcaacaagcaa
GlsR8†	GLCHR03:136649-136756(-)	gggtt**CGTAGATGAAGAGAGATAAATCAGCTACCGCTGAGCCCAACGTGAGGAAGAAACCGCCTTTCGTCTGACC**cttcactcaacagcccc
GlsR9†	GLCHR01:636600-636707(-)	**ACCCGTGATTTGCAACGCTTAGTCCGTGTTTCGGAGTGTCTTGCACGCTGATGAGTGAAAGCACACATGAGGTT**cctttaataaaatgcaga
GlsR10†	GLCHR05:3181795-3181854(+)	cg**AGAATGATGAGACGTGTTCCTCTCTCCTACAGACTCCCTGGGGA**TGCTATGTACACCTTA**CTGATTTA**cttcctttttctcaagggccat
GlsR13†	GLCHR05:4314721-4314828(-)	**TGGGAGCGACCTATCTTGAGGACGACGGCCGCCCGTCTTACCTTGTGACGTTTGCCGTCTTACAATGCTCTGACC**ctttactaagctgccg
GlsR14†	GLCHR01:1279750-1279857(-)	aataaata**AAATGATGACAATGCGCATTTGTCAGAAGGCTCACTTCTGATGATTCCTCTGTCCATTCCCCTGATCCT**ttcctcaacaggtat
GlsR15†	GLCHR01:1329992-1330099(-)	**CTCCTTGGTTCCTCGCAGAATGATTATCTGTCTCCGAGCAAGCACGACTATGAGCTTACTTATGAGATCTGACTCC**tttactcaatgttagt
Candidate-1ϕ	GLCHR01:1215994-1216101(-)	agcagacg**AAAAAATAAATGAAGACAGAACCACAGACCTGTACTGACCCTTGATGTTAGTTGTCGCTCTGATA**tcctttactcaatcgtgtc
Candidate-2ϕ	GLCHR01:33293-33400(+)	gtcgaaaataaag**TGATGATTC**GAATTACCGCCCGAGGGCCCTCGGGCTCCGCTGAGGACATGCTGGTCTGACTcctttgctcaacctttcc
Candidate-13ϕ	GLCHR01:1010277-1010384(-)	aaatgattactccaa**CACGACGGTCTACTGAGAACCCAGTATCTTTAGACTGCTGAGACAGTGTTATATGATT**tcctttacttaaggctctc
Candidate-23ϕ	GLCHR01:449998-450105(-)	atgagctattttg**TACCACTCTGACCGTGAGGCGTATGCCTAGGGCATGGAGAAGAGCAGACTTGAG**gcctgttcctttactcaattttgtg
Box H/ACA RNAs		
GlsR17†	GLCHR01:149373-149480(+)	**CTGAGCAATCCCCAGGACACAGGCGGAGCGGAAGGCACGGCTGCGCCACGCAGCCTAATCACCGCCCCTATAGTC**cttttctaaacgcgtgg
GlsR18†	GLCHR01:149494-149601(+)	**GCGTGCACAGGCCTACATCCAGGGTCATAGGTGGGGAGCGGATCCCGTCCATCCTCAATCCGGGCCCGCACAGTC**ctttactcaagcttact
GlsR19†	GLCHR05:2415660-2415767(-)	**TAGGCTGGCCAAGCATCGTTGATAGAAGCTGCTCTTGGTCACCGGAGGGTCTCCGGTTTCATACGCAGAGACATCC**ttcaattaaaaacttt
GlsR20†	GLCHR04:893921-894028(-)	**GGCGCGTCAGAGGCCGGCTAGAGCGCGACTGGTTGAGTTCCCAGAGCGATCTGGGTGATTAGCAGTCATACAGTC**ctttacttaagcctact
GlsR21†	GLCHR04:1151475-1151582(-)	**GATCGGTGTTATGCTTTGTTGGGATAGCAGGCCGTGCCAGTTGGACAGCCAAGGTCCACCTCTGGTTCGGCACACAT**ttattcaagacatct
GlsR22‡	GLCHR03:1225005-1225112(-)	**CGTTCCCTGGGGCGATAGCTCTTGCTGGCAGGTCTTGCAGTGTCCATACCCGGGCAACACGTTTTCCAGCTACACCT**ttactcaacgtgcac
GlsR23‡	GLCHR04:1890711-1890818(-)	**ATGCCGTGACGAGACACGCACTGGGTGGCCATTGCGTCTGCGGTAGATCCGCCGATTCCACAGCCCAGAAACACCC**tttactcaagctggct
GlsR24‡	GLCHR05:1301296-1301403(+)	**CGGGGCAGAGTCGGCCCTCCCAGAGCCCGCCGACGCCCCCGAGCGCCAGCCCGGCGCAGGGGCCCGGCCACACTC**atttattaaacagcgat
GlsR25‡	GLCHR05:1459391-1459498(+)	**TGCGTAGCCGATAGGTACGGGTGACCGTTTATCCCGGGCTCGTGTGGGCCCGGGTAGGCACGGTCAAAGAGTT**tccttcattcaatttttag
Candidate-16ϕ	GLCHR05:1006432-1006539(-)	**CGCTCTGCCAGATACGCCGACAGAAAGCACCAAGGAAGGATGTGGATCTCCATGTCTGCCGTGTGCGCGCATATC**ctttactcaatctgtgt
GlsR26*	GLCHR05:1459529-1459636(+)	**CCTTGCGTGCGCATATCTCCGGGATCTGCGCCGCGTCTGCTCGCGGCGATTTCGGTTATGCCGGCCCGAACACTCC**ttcattcaacaggccc
GlsR27*	GLCHR01:1371298-1371405(+)	**TATGGGCAGCGAAAGTACCAGAGCCAAAGAGTTCCTCTGATCGCCTGGCCGGAGCACATTTGTGATCTCCTATACCT**tcatttaattagcgt

Table 3.1. (Continued)

Spliceosomal snRNA Candidates and other ncRNAs

Gl U1 Cand*	GLCHR03:661194-661301(-)	**GTGCTGCGCATACCGCGCTGGCACTGGTCACGGGGCAGTGCTCTCAGACCTGCTACCGTACCCTTTTAATTTTTCCTTcacttaa**ggcccat
Gl U2 Cand	GLCHR04:581396-581503(-)	**TACATGCAAGGGGCAGCCGGGCTGTGAGGCAGCTGCCAGGATGGTCCTGCCCTTGTCCCGGCTGGCGCCGTCCACCTTtattcaa**gttttct
Gl U4 Cand	GLCHR02:1195038-1195145(+)	**AGGTGCGTGATCCCTCGGTGATGCCTTGAGTGTTGCTTCACCAAAGAACAACCACACGGCACAGCCGAATCTCTCATTttttaaa**ctttctc
Gl U6 Cand*	GLCHR04:1813414-813521(-)	**ACCAGCTTCAGTCTAGAGTCGCTGGGGGACCTCTGGTTTCGCGGGAGCCCGTTGGCGCGTGCTTGCACCCCGCTCCTttctcaa**tcttcgc
RNase MRPΩ	GLCHR01:479893-480000(+)	**GCCGCCACACTGACAGTTATGGTTGCAGGACAAGCTTAGCGAGTCCGAACTCGACAGGGATACTCTACAGCGTTCCTttattcaa**tcattga
TERC Cand*	GLCHR01:978028-978135(-)	**TTCTCGCAGTGGTCTCTGTCAAAGACTGCCGCACGGTACACCAGAAGCAAGGGGAAGGATCCCATCCACGCAGTCCTttacttaa**acatggg
Candidate-3ϕ	GLCHR01:702197-702304(-)	**CCAGGTCCAAGACCCGGGCAGTCTGTGCTGTGGGGCGCCGCTGTAGACGTCTTCCGAACACACCTGCGATAAAC**cctttatttaa**aagatta**
Candidate-5ϕ	GLCHR03:299427-299534(+)	gaggaaacgagtgtttcgccgggcataactggg**CATGCATTTTCCTTGCCCAGTCTGCCTCCATACTAATTTCTCCTTtactcaa**tcaggat
Candidate-12ϕ	GLCHR03:474659-474766(-)	aaagaaccccaa**CCCGATGACGAATAGCTGTCCTGGCGGAGGCGGTCATGACGACGAAGCCATCACGTAGGATC**ccttcactcaa**cctctgc**
Candidate-15ϕ	GLCHR02:350570-350677(+)	**GACAGCCGGAGGCCGGAGACGGAGCACGGTCAGGCGGGCGGGGTGCAGTGCCAGCCCCAGCCGCAGAGCGGCTTCCTTtactcaa**gatcggg
Candidate-17ϕ	GLCHR03:1601283-1601390(-)	**CGTTAAGCGAGGCTTGGCCCGTGCGACGATGAGGCTCCCTGCGGGGA**agccctgcggcgcgtcttaaggaggc**tccttcactcaa**cggcgtc
Candidate-21ϕ	GLCHR05:1896857-1896964(-)	**TCTCCCACCGGAGCACATATGCTGCAGGATGACCGGCGCCTGTCTCCCACCACGTGCCAGCTAAACTGCAGCCACATTtattcaa**cctttc
***Trans*-spliced Intron 5′ Halves**		
Hsp90 Intron	GLCHR05:2515303-2515410(+)	**AAGAAGAAGACGGGCATCAAGCTCATGGTCAAGAAAGT**gtatgttatgtt**tgtatgctgtatgtgtgcgagac**t**cc**tttactcaa**a**tttgcg**
DHC β Intron 1	GLCHR03:577636-577743(+)	**ATTTGACAAGG**gtatgttactg**ggtgaaacgctacttat**gtatgtatgcttatatgtcttcgcgctcaggcgc**tcctttactcaa**ttatcag
DHC β Intron 2	GLCHR05:4266366-4266473(+)	**tgtgtagtcgcagtatgccatt**attttataacgtgtatgtcattatgtcagtatgccagtgcgctggtgagtt**tcctttattcaa**atgttgt
DHC γ Intron	GLCHR03:967523-967630(+)	**ATTCCGTAGCGTTGCAATG**atatgttcac**aggtggtttggtgtgtatgcttggcgt**gtatgtgtgtatgttcc**tcctttactcaa**tacttgg

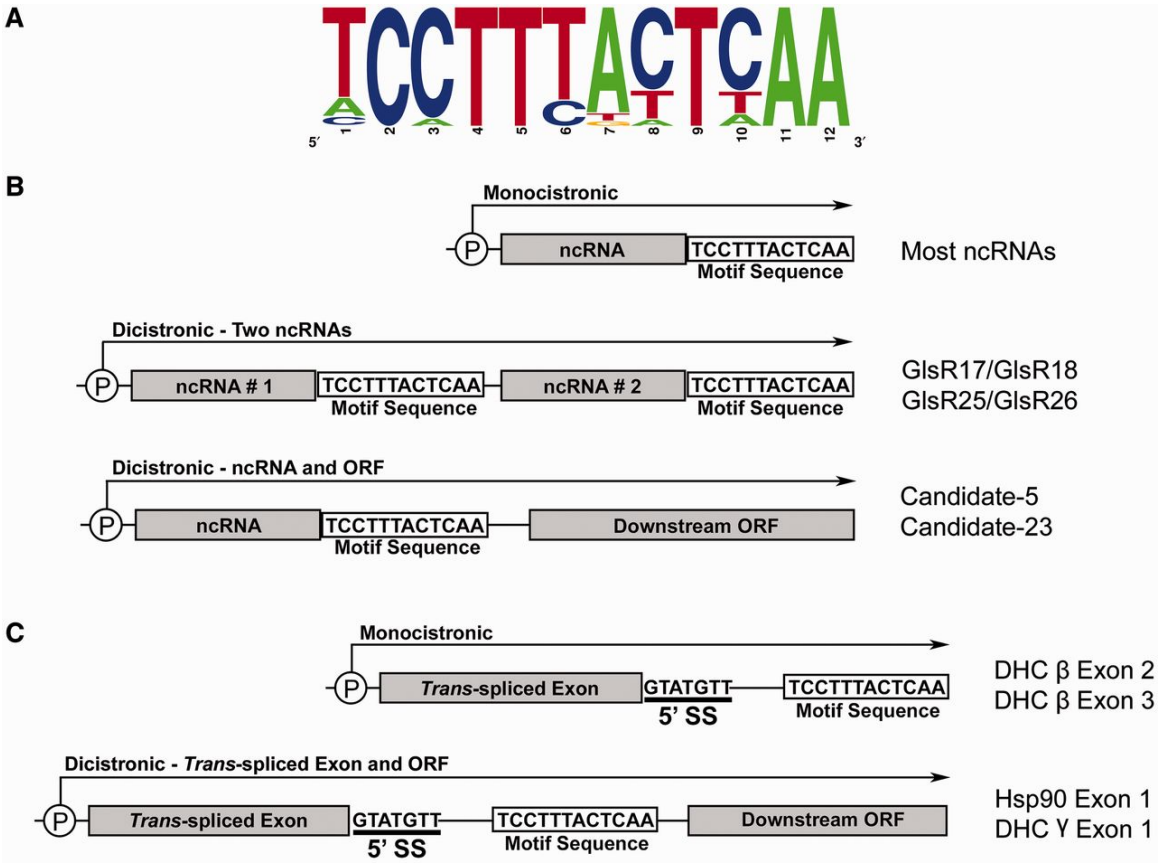


Figure 3.1. Identification of a 12 nt sequence motif within *G. lamblia* ncRNA and *trans*-intron containing genes.
(**A**) Motif sequences from *G. lamblia* WB, P15 and GS isolates (n = 132 sequences) were used to construct a WebLogo sequence logo frequency plot (Crooks *et al.* 2004). Nucleotide frequency at each motif position is denoted by the relative height of the letter. Genomic organization and mode of expression of *G. lamblia* ncRNAs (**B**) and 5′ *trans*-spliced intron halves (**C**) containing the 5′ splice site and showing the relative location of the processing motif. Promoter sites are indicated by a circled 'P' with extended arrows indicating predicted initiation sites, lengths and directionality of precursor transcripts. Representative examples of each mode of gene organization are indicated.

3.3.2 The conserved motif mediates 3′ end formation of *G. lamblia* ncRNAs

The conservation of the motif and its consistent position relative to predicted mature RNA 3′ ends suggests it may play an important role in *Giardia* ncRNA 3′ end formation. Additionally, several *Giardia* snoRNAs are encoded immediately upstream of annotated ORFs or other ncRNAs, with short intervening spacer sequences that contain the motif (Figure 3.1B). Thus, we hypothesized the motif may serve to either mediate transcription

termination or post-transcriptional cleavage of precursor ncRNAs. To examine this, we used RT-PCR and RACE techniques, in conjunction with DNA sequencing of amplified products, to detect precursor transcript species and to map mature RNA ends. RT-PCR experiments confirmed the presence of dicistronic precursor transcripts consisting of two different ncRNA species (Figure 3.2, lanes 1 and 3) or an ncRNA with a downstream ORF (Figure 3.2, lanes 5 and 7). Next, 3′ RACE experiments determined that the majority of mature ncRNA 3′ ends are consistently located at the third to the fifth position of the 12 nt motif sequence (Figures 3.3 and A.2.3). One explanation for the observed mature ncRNA 3′ end pattern is that endonucleolytic cleavage of mono- or distronic precursor transcripts occurs at a specific position within the motif sequence. Alternatively, mature ncRNA 3′ ends may be the result of cleavage (or termination) further downstream followed by 3′ to 5′ exonucleolytic trimming to positions 3 to 5 of the motif sequence.

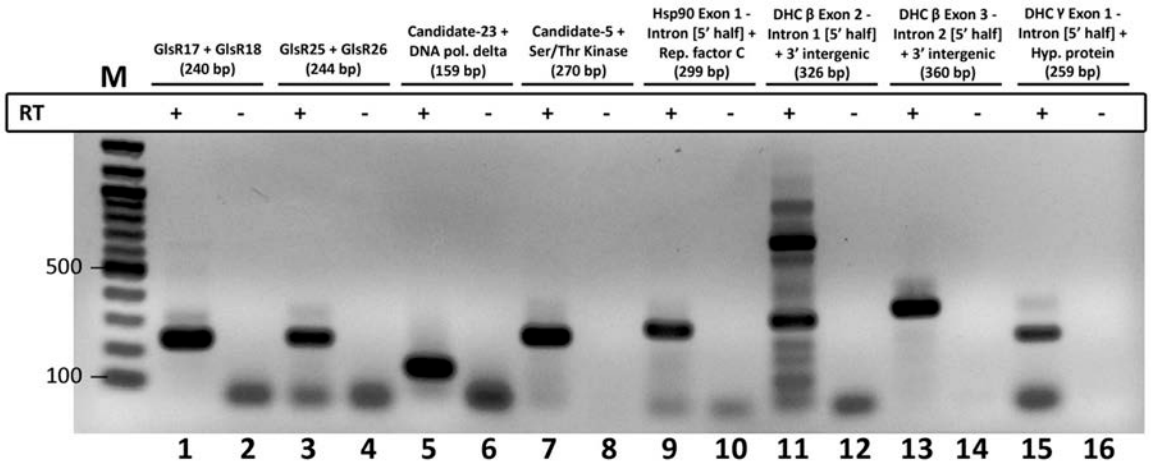


Figure 3.2. Dicistronic transcription of *Giardia* ncRNA and *trans*-spliced intron precursors.
RT-PCR detection of precursor transcripts containing two different ncRNAs (designated GlsR#) (lanes 1 and 3), ncRNA (designated Candidate-#) with downstream ORF (lanes 5 and 7) or *trans*-spliced intron 5′ half with unrelated downstream ORF (lanes 9 and 15). Products of expected size (sizes indicated in parentheses; also refer to Figure A.2.2) were sequenced to confirm their identity. Experiments were either performed with the addition (+) or omission (−) of RT enzyme during the cDNA synthesis reaction. M = molecular weight marker, bp = base pairs.

To distinguish between potential RNA motif processing pathways, we performed RLM 5′ RACE to map predicted pre-RNA 5′ ends for 2 different ncRNAs (Candidate 5 and Candidate 23) and the *trans*-spliced *Hsp90* 5′ half, using reverse primers annealing immediately 3′ of motif sequences (Figure 3.3A and Figure A.2.4). If motif processing occurs via downstream cleavage/termination followed by exonucleolytic trimming, we should only detect cDNA products with 5′ ends originating from the upstream ncRNA promoter (i.e. prior to any RNA cleavage). On the other hand, if endonucleolytic cleavage occurs within the motif sequence, we expected to detect some 5′ RACE products with 5′ ends located at some position(s) within the motif sequence. For the two ncRNAs examined, we found that a large proportion of the cDNA clones had 5′ termini ending abruptly at the sixth position of the 12 nt RNA sequence motif (Figure 3.3B and Figure A.2.4). The *Hsp90* 5′ half cDNA clones showed more heterogeneous 5′ ends terminating downstream of the motif (Figure A.2.4). Most importantly, no single 5′ RACE clone contained a 5′ end terminating at any position upstream of the sixth position of the motif, and correspondingly, no ncRNA 3′ RACE clone contains a mature 3′ end downstream of motif position five (Figure 3.3B). Together, these data suggest endonucleolytic cleavage occurs between positions five and six of the motif sequence to generate ncRNA 3′ ends.

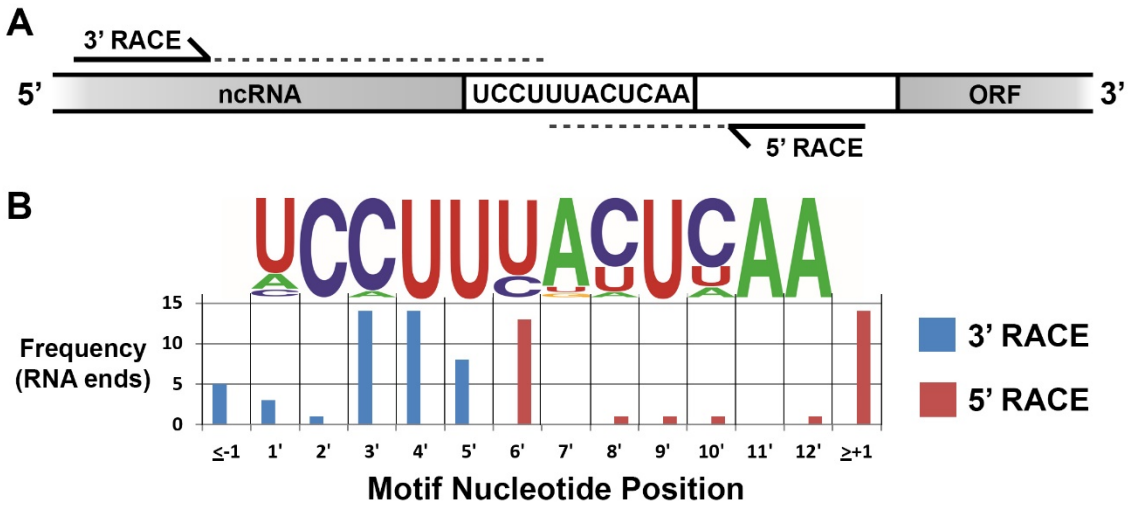


Figure 3.3. RNA end mapping suggests endonucleolytic cleavage within the motif. (**A**) Experimental strategy for 3′ and 5′ RACE RNA end mapping, showing the relative locations for primer binding (and extension products) relative to RNA motif sequences. (**B**) Frequency plot showing terminal nucleotide positions for individual sequenced clones from 3′ RACE of ncRNA or *trans*-intron 5′ fragments (blue bars) and 5′ RACE of RNA regions downstream of the motif (red bars) (see Figures A.2.3 and A.2.4 for clone sequences).

The two ncRNAs and *Hsp90* 5′ half precursor RNAs examined for motif cleavage are encoded upstream of unrelated ORFs (Figure 3.3A and Figure A.2.4). We next asked whether downstream ORF mRNAs were generated via polycistronic transcription with their upstream ncRNA followed by precursor RNA motif cleavage to produce the mature 5′ end of the ORF containing product or alternatively, by expression from an autonomous internal promoter(s) downstream of the motif. To test this, we performed RLM 5′ RACE experiments, a technique that assesses 5′ cap structure status of RNAs, to map mature 5′ ends of mRNAs encoded by downstream ORFs. Since canonical eukaryotic translation initiation requires a 5′ nucleotide cap structure, we examined only 5′ RACE products originating from capped mRNAs by using *G. lamblia* total RNA samples treated with DNase I, CIP and then TAP. Only capped RNAs will be RT-PCR amplified following the RNA linker ligation step (see Materials and Methods, section 3.2.3). In all three examined

cases, downstream ORF mRNA capped 5′ ends mapped to A-T rich sequence elements downstream of the motif (Figure A.2.4) which are similar to transcription initiation sites from other characterized *G. lamblia* protein-coding genes (Yee *et al.* 2000, Elmendorf *et al.* 2001). It seems likely that the majority of capped ORF mRNAs originate from independent promoters located downstream of (or overlapping with) motif sequences.

3.3.3 The conserved sequence motif has a role in the novel *Giardia* mRNA *trans*-splicing pathway

Given the common association of the sequence motif with previously annotated ncRNAs, we predicted that we should be able to use this motif as a tool to identify other genomic regions specifying novel ncRNAs. Initially, we performed simple BLASTN searches using individual motif sequence variants against the *G. lamblia* WB genome sequence database. Surprisingly, these searches revealed canonical motif sequences residing in the 5′ halves of all four known *Giardia trans*-spliced introns previously identified by our group and others (Kamikawa *et al.* 2011, Nageshan *et al.* 2011, Roy *et al.* 2012) (Figure 3.1C and Table 3.1). Motif sequences locate immediately downstream or overlap with those nucleotides predicted to be involved in intermolecular base-pairing interactions that mediate the *in vivo* association of the intron 5′ and 3′ halves.

We next performed RT-PCR experiments to characterize *trans*-spliced intron containing precursor transcripts. These experiments detected extended precursor mRNA species with 3′ ends extending beyond the conserved sequence motif, and in some cases, extending into downstream unrelated ORFs (Figures 3.1C and 3.2). Mapping of 3′ ends by 3′ RACE shows that precursor RNAs containing *trans*-spliced intron 5′ halves are also cleaved at their motif sequences (Figure A.2.3).

We next asked whether motif cleavage was an essential precursor step for *trans*-splicing to occur. To assess this, we employed an RT-PCR strategy capable of amplifying branched introns representing (at minimum) products of the first step of *trans*-splicing (Figures 3.4A and A.2.5A). Reverse primers annealing upstream or downstream of motif sequences of *trans* intron 5′ halves were used to generate cDNA extension products which bridged 5′ splice site-branch point junctions. RT products were then PCR amplified using forward primers annealing upstream of 3′ intron branch point (BP) sequences and reverse primers employed for cDNA synthesis followed by cloning and sequencing of amplified products. Conveniently, this strategy also detects the identity of the intron BP nucleotide and has been used successfully to map BP nucleotides in other eukaryotes (Gao *et al.* 2008, Zhang *et al.* 2011). RT-PCR detected products with sizes corresponding to both R1 and R2 primer annealing positions (Figures 3.4B) and sequencing of numerous clones verified that indeed the first step of *trans* splicing can occur without motif cleavage (Figure A.2.5B). Intron BP nucleotides have been mapped in very few eukaryotes. Thus, we assessed positions of intron branching by examination of the first nucleotide upstream of the 5′ most nucleotide of the 5′ intron half in RT-PCR sequence reads of branched RNA species. Interestingly, we found that branching may occur at two adjacent adenosine nucleotides within predicted intron BP sequences (Figures 3.4C and A.2.5B). Frequent misincorporation of an ‘A’ nucleotide (instead of ‘T’) at the BP nucleotide during reverse transcription has also been observed in other studies (Gao *et al.* 2008). Given the high conservation of *G. lamblia* intron BP sequences, we predict that other introns from this organism will also utilize these nucleotides for intron branching. However, we acknowledge that our positioning of forward primers (F2) during RT-PCR would not allow

for the detection of other upstream branching positions, and thus we do not rule out the possibility of other alternative BP nucleotides.

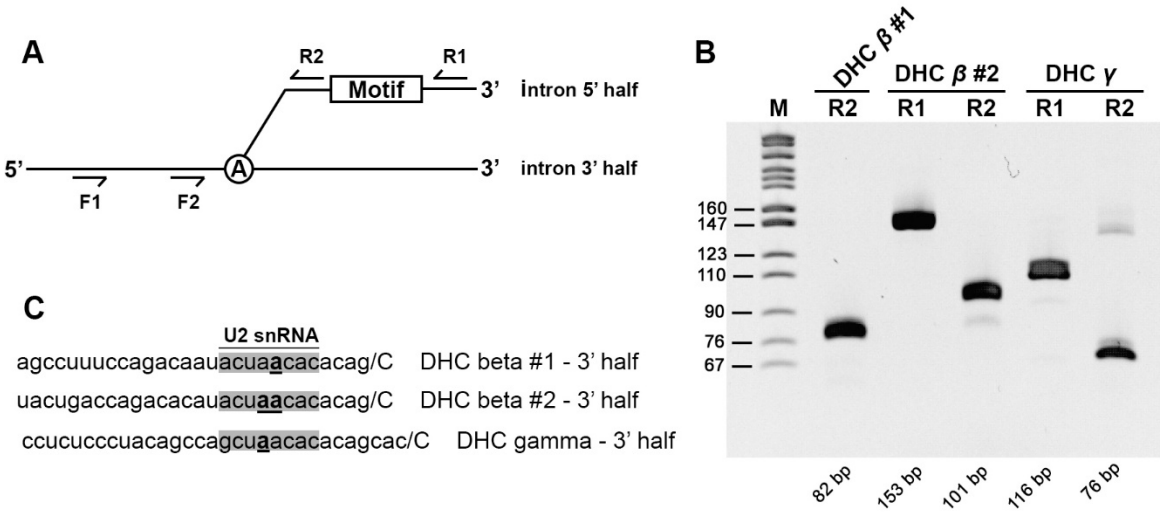


Figure 3.4. Detection of motif cleavage and BP nucleotides of *G. lamblia trans*-introns. (**A**) Experimental approach showing binding locations for forward (F1 and F2) and reverse (R1 and R2) primers used during RT-PCR. See Figure A.2.5A for more details. (**B**) RT-PCR products from nested PCR using F2 primers and reverse primers annealing downstream (R1) or upstream (R2) of the motif sequence. Names for *trans* introns and expected RT-PCR product sizes are shown above and below lanes, respectively. M = pBR322 Msp I size marker. Band sizes are indicated in base pairs (bp). (**C**) Branch point nucleotides determined by sequencing are indicated in bold text and underlined. Intron regions predicted to interact with U2 snRNA (Figure 3.10C) are highlighted in grey. See Figure A.2.5B for sequence traces confirming BP nucleotide positions.

In summary, these results indicate that some of the steps in the unusual *Giardia* mRNA *trans*-splicing pathway involve the generation of extended transcripts containing the conserved processing motif which may then be cleaved to generate intron 5′ halves whose ends reside directly adjacent to nucleotides predicted to mediate association to intron 3′ halves. However, while the first step of splicing may precede motif cleavage, it is unclear whether motif cleavage occurs at some point during or following the second step of splicing. Alternatively, due to the ability of RT-PCR to amplify even the most non-abundant RNA species, it is conceivable that motif cleavage occurs primarily before either step of *trans*-splicing and only rarely does either step of splicing precede motif cleavage.

Because all identified *Giardia trans*-spliced introns contain the conserved motif sequence downstream of the 5′ splice site, we also searched for any instances of motif sequences within annotated protein-coding genes that may indicate the presence of additional uncharacterized *trans*-spliced introns. Searches of the *G. lamblia* WB genome identified four cases in which a motif sequence could be found within a conserved protein coding gene (Figure A.2.6). RT-PCR experiments confirmed expression of each of the four motif-containing genomic regions; however, 3′ RACE experiments failed to detect products corresponding to transcripts terminating near motif sequences (Figure A.2.6). Unlike the previously characterized *trans*-spliced introns, these regions do not interrupt protein-coding continuity in these genes (consistent with these regions being exons), and it is interesting to note that the presence of the sequence motif alone does not always result in RNA cleavage. This may indicate a requirement for the association of motif recognition/cleavage factors with other ncRNA assembly or processing factors (or even spliceosomal components in the case of *trans*-spliced introns) that only occurs when the motif is located in the correct structural or spatial context.

3.3.4 A *G. lamblia* telomerase RNA candidate

We next used the high sequence conservation of the newly discovered motif to identify novel candidate *Giardia* ncRNAs, using the sequence pattern matching program 'Scan for Matches' (Dsouza *et al.* 1997). Our searches identified five novel ncRNA candidates that maintained promoter and motif sequences in all three isolates, mapped to intergenic regions and showed significant sequence conservation in all three isolates (Figure A.2.1-27, 28, 29, 32 and 34). To verify *in vivo* expression of these ncRNA candidates, primer extension and northern blot analysis of total *Giardia* RNA was

performed (Figure 3.5). Products of expected size were detected for all five ncRNA species (detected in isolate WB). Next, 5′ and 3′ RACE experiments were used to more accurately determine the sizes and map the mature 5′ and 3′ ends of each species (Figure A.2.3). Again, these experiments show that mature 3′ ends coincide with the conserved 3′ end sequence motif, and suggest that these RNAs may be processed by a similar mechanism to the *trans*-spliced intron containing transcripts.

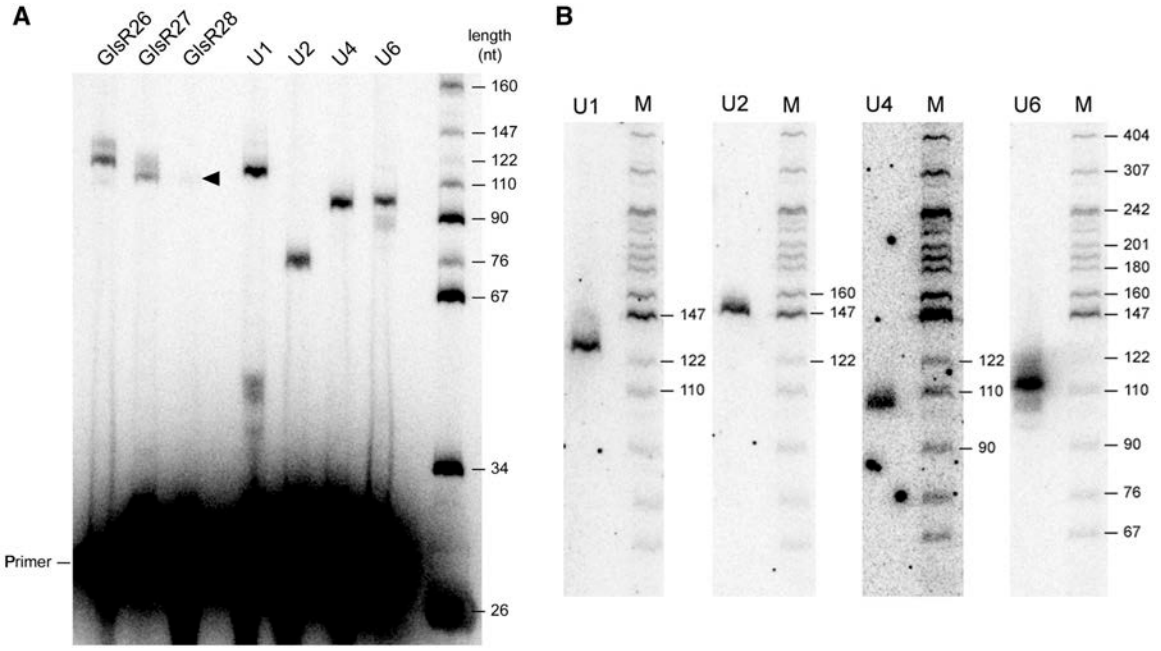


Figure 3.5. Detection of *G. lamblia* ncRNA expression.
(**A**) RT-primer extension experiments were performed using gene-specific ^{32}P-labelled oligonucleotides annealing directly adjacent to the RNA processing motif of each *Giardia* ncRNA (GlsR) or spliceosomal snRNA candidate. For each RT reaction, 10 μg of *G. lamblia* WB isolate total RNA was used as template with SuperScript II™ RT, and products were resolved by 8% denaturing PAGE and visualized by phosphorimaging. Rightmost lane is ^{32}P- labelled Msp I digested plasmid pBR322 size marker, with fragment sizes indicated in nucleotides (nt). An arrowhead indicates a faint primer extension product of expected size for the Telomerase RNA candidate (GlsR28) (**B**) Northern blot analysis of *Giardia* snRNAs. DNA probes specific for U1, U2, U4 or U6 snRNA candidate sequences were hybridized to *Giardia* total RNA that was fractionated by 8% denaturing PAGE. The DNA size ladder (M) is the same as in part (a).

Two of the novel ncRNAs, designated GlsR26 and GlsR27, show the conserved structural features of box H/ACA snoRNAs (Figure 3.6). GlsR27 is predicted to guide the

formation of pseudouridine (Ψ) modification at U1745 of the *Giardia* large subunit rRNA (Figure 3.6B). The GlsR26 candidate is encoded immediately downstream of the GlsR25 box H/ACA snoRNA genomic sequence (Luo *et al.* 2006). This arrangement is similar to the previously reported organization of the GlsR17/GlsR18 snoRNA gene cluster (Yang *et al.* 2005) (Figures 3.1B and 3.6C).

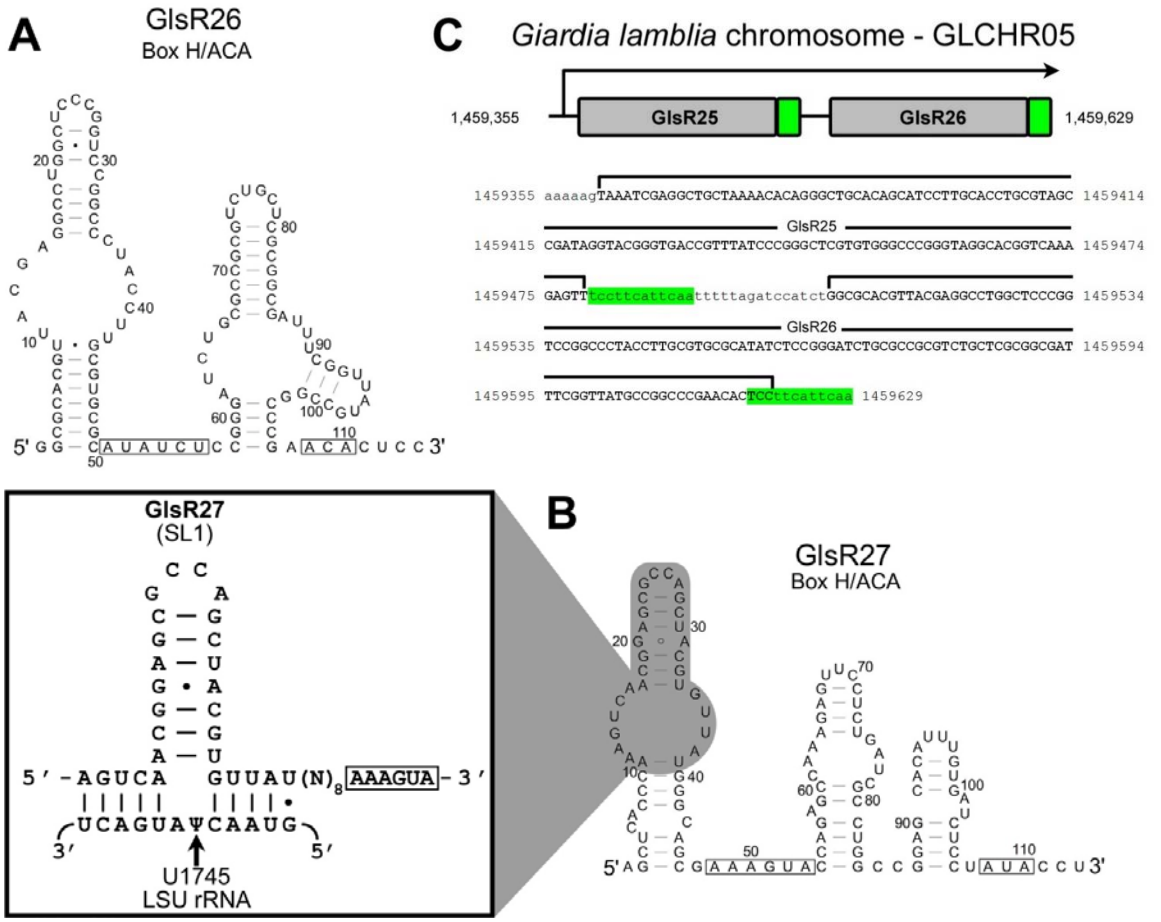


Figure 3.6. Two novel *G. lamblia* box H/ACA snoRNAs.
(**A** and **B**) Secondary structures based on MFOLD (Zuker 2003) structural predictions for GlsR26 and GlsR27 snoRNAs. Mature ends were determined by RACE experiments and the predicted "H" and "ACA" box sequence elements are in open boxes. (**B**, inset) The "pseudouridine pocket" formed by the predicted pairing of the guide region of GlsR27 stem-loop 1 (SL1) to the *G. lamblia* 28S large ribosomal subunit (LSU) rRNA. This interaction is predicted to target U1745 for pseudouridine formation. (**C**) Organization of the GlsR25-GlsR26 snoRNA gene cluster is schematically represented with genomic DNA sequence from *G. lamblia* WB isolate displayed below. A line with arrowhead indicates a predicted single transcription initiation site to produce a precursor polycistronic transcript containing both GlsR25 and GlsR26 snoRNAs. Nucleotide sequences found in the mature snoRNAs are in uppercase bold and intervening sequences in lowercase. The locations of processing motif sequences are highlighted in green.

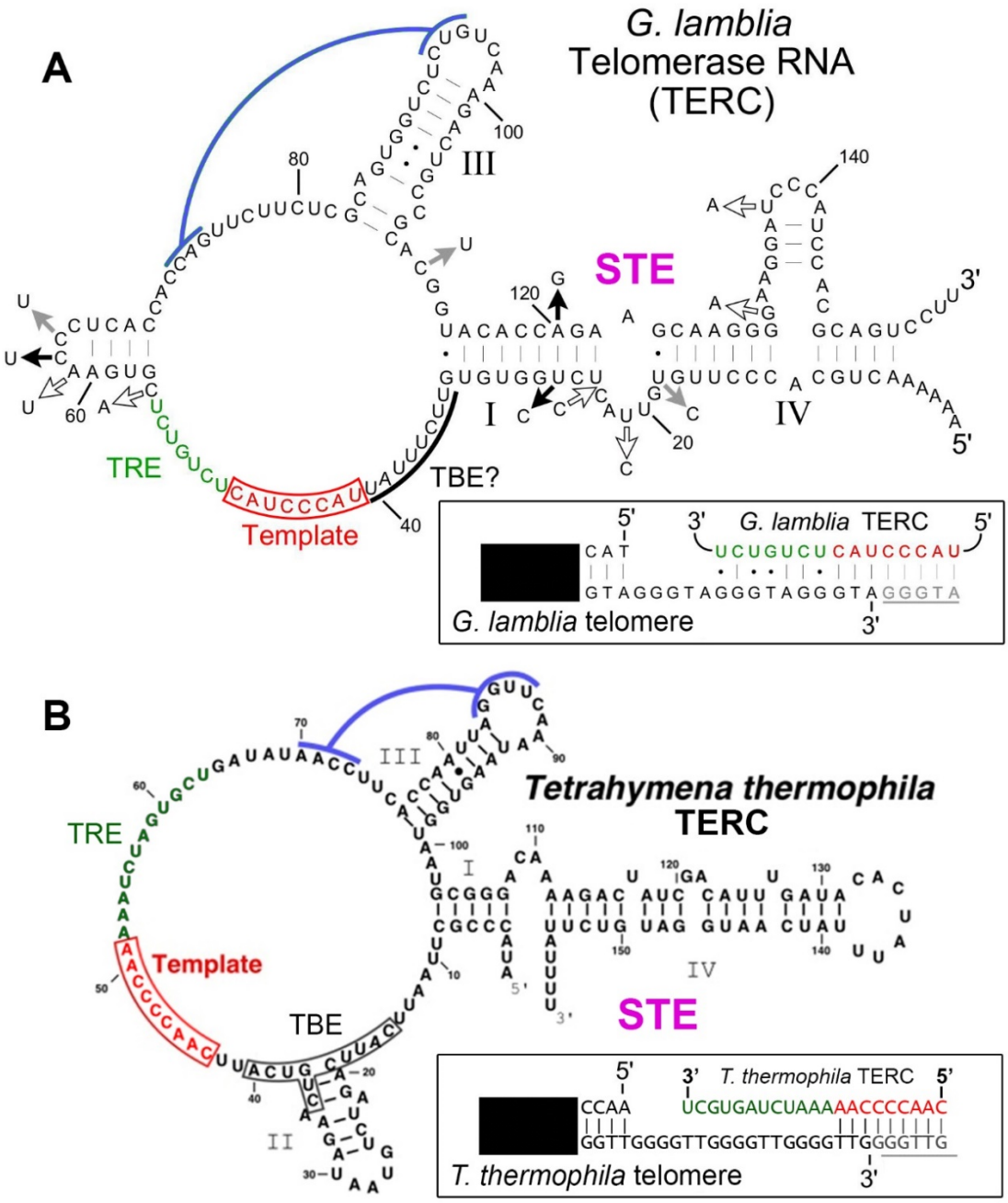


Figure 3.7. A putative *G. lamblia* telomerase RNA component (TERC).
(**A** and **B**) Secondary structure prediction (MFOLD) of the *G. lamblia* telomerase RNA candidate (TERC) (GlsR28, WB isolate sequence is shown) is compared to TERC from *Tetrahymena thermophila* (Podlevsky et al. 2008). TERC template regions are boxed and in red text and template boundary elements (TBE) are indicated. TRE = template recognition element, STE = Stem terminus element. RNA nucleotides involved in forming pseudoknots are over-lined in blue. Nucleotide substitutions occurring in *G. lamblia* P15 and GS isolates are denoted by grey and black arrows, respectively. (**A** and **B**, inset) The hypothetical base-pairing interaction between *G. lamblia* TERC template/TRE and telomere repeat sequence and the equivalent interactions in *T. thermophila* is shown with the next telomere repeat to be added in grey text and underlined.

Initially, the novel GlsR28 ncRNA was not classified into any RNA functional group (Hudson *et al.* 2012). However, subsequent 5′ and 3′ RACE end mapping experiments and further inspection revealed that a portion of GlsR28 is complementary to the *G. lamblia* telomeric repeat sequence 'TAGGG' (Le Blancq *et al.* 1991), implicating this ncRNA as the missing *G. lamblia* telomerase RNA component (TERC) (Figures 3.7 and A.2.3). TERCs minimally contain three elements: i) a template region capable of specifying the sequence of telomere repeat addition, ii) a pseudoknot and iii) a stem terminus element (STE) (Blackburn and Collins 2011). Beyond these features, TERCs may vary drastically in their lengths (140-1500 nt), primary sequences and accessory domains (Blackburn and Collins 2011). Interestingly, MFOLD predictions suggest GlsR28 may fold into a secondary structure resembling the TERC from *Tetrahymena thermophila*, with the putative GlsR28 template and pseudoknot elements in the same relative locations as in *T. thermophila* (Figure 3.7). Additionally, the 5′ and 3′ ends of GlsR28 are predicted to form a long, discontinuous helix which resembles the terminal stem-loop (STE) from *T. thermophila* TERC. Supporting our inferred GlsR28 secondary structure, nucleotide changes in the various *G. lamblia* isolates occur in RNA single-stranded regions or result in compensatory changes which maintain RNA secondary structures. Taken together, these characteristics strongly suggest GlsR28 is the authentic *G. lamblia* TERC; however, additional experiments will be required to verify the functionality of this ncRNA.

3.3.5 Identification of novel *Giardia* U1 and U6 snRNA candidates

Most surprising was our finding that the remaining two ncRNAs possess conserved sequence motifs and secondary structures diagnostic of U1 and U6 spliceosomal snRNAs (Figure 3.10). The predicted *Giardia* U1 snRNA structure adopts the typical 'cloverleaf'

secondary structure and contains a predicted U1A binding site sequence that is a close match to the *Saccharomyces cerevisiae* 'CACAUAC' sequence (Tang and Rosbash 1996). This contrasts to the previously identified U1 candidate (Chen *et al.* 2008) that has an atypical predicted secondary structure lacking a recognisable U1A binding site (compare Figure 3.10A and Figure A.2.7A and A.2.7E). We mapped the 3′ end of the new U1 RNA candidate as directly downstream of the predicted Sm site, similar to the *Candida albicans* U1 snRNA, which also lacks the SL IV structure (Mitrovich and Guthrie 2007) that is commonly found in the U1 RNAs of other eukaryotes (Figures 3.10A and 3.11A). Another noteworthy feature of the *Giardia* U1 RNA is the lack of a conserved U1-70 kDa protein binding site sequence in SL I. A *bona fide* U1 snRNA is expected to base pair to the 5′ splice site of *Giardia* spliceosomal introns, and we observe extensive base pairing potential of the 5′ end of the U1 snRNA to the highly conserved *Giardia* 5′ splice site sequence (Figure 3.10A). The non-canonical U•U pairing is also observed in *S. cerevisiae* U1-5′ splice-site interactions at the same relative position (Siliciano and Guthrie 1988). In humans (Reddy *et al.* 1981) and *S. cerevisiae* (Massenet *et al.* 1999), the two adjacent 'U' residues in U1 snRNA are converted to Ψ, which may be important for the function of U1 snRNA in splicing (Karijolich and Yu 2010). As the identified *Trichomonas vaginalis* snRNAs are not 5′ capped (Simoes-Barbosa *et al.* 2008), we examined the capping status of the *Giardia* snRNAs using RLM 5′ RACE (Figure 3.8). Only CIP plus tobacco acid pyrophosphatase-treated *Giardia* RNA samples generated PCR amplicons containing snRNA 5′ end sequences after linker addition, RT-PCR and sequencing. These results suggest the majority of *Giardia* U1 and U2 snRNAs are 5′ capped; however, experimental results for the U4 and U6 snRNAs are ambiguous (see Figure 3.8).

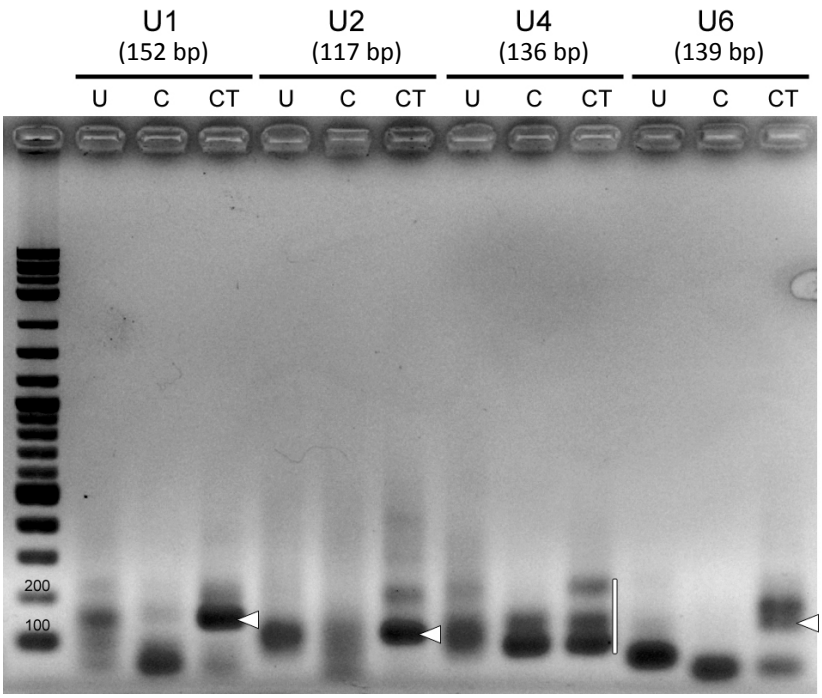


Figure 3.8. RNA linker mediated (RLM) 5′ RACE analysis of snRNA candidates.
Giardia WB total RNA was untreated (U), CIP-treated (C) or CIP plus TAP-treated (CT) before ligation of a common RNA linker to RNA 5′ ends followed by RT-PCR amplification (see methods for more details). Only products indicated with open arrowheads from "CT" treatments were determined to be snRNA products by DNA sequencing following gel extraction of bands, thus indicating the presence of 5′ nucleotide cap structures. Other sequences were determined to be non-specific (i.e. non-snRNA) amplification products. Sequences appearing in the "U" and "C" treatments determined they were also non-specific RT-PCR amplicons and did not correspond to uncapped versions of the snRNAs. U4 "CT" treatment products (white bar) were individually gel-extracted, cloned and sequenced which identified several non-specific amplification products and a single U4 sequence clone. Expected sizes for snRNA RT- PCR products and DNA marker bands are given in base pairs (bp).

The U6 snRNA candidate contains the conserved 'ACAGAGA' and invariant 'AGC' trinucleotide sequences found in all known U6 snRNAs (Guthrie and Patterson 1988) (Figure 3.10B and C). The predicted Mg^{2+} binding site in U6 intramolecular SL (ISL) (boxed region, Figure 3.10C) shows a sequence differing from the typical non-canonical C•A pair followed by bulged 'U' residue. Similar divergence within the U6 ISL is also observed in *C. albicans*, which contains a bulged 'C' instead of a 'U' (Mitrovich and Guthrie 2007), and in *Trypanosoma cruzi*, which has an unusual C•C pair instead of C•A

(Ambrosio *et al.* 2007) (Figure 3.9). Some key differences in the *Giardia* U6 snRNA candidate that we report here, as compared with the different U6 snRNA candidate previously reported (Chen *et al.* 2008), is the presence of complete 'ACAGAGA' and 'AGC' trinucleotide sequences that are strictly conserved between *Giardia* isolates (compare Figure 3.10C and Figure A.2.7D and F), unlike the previously predicted U6 snRNA that displays unexpected substitutions in the three isolates disrupting these functionally critical sequence elements. Our identified U6 snRNA also displays more robust base-pairing potential in the U6 ISL. We also note that sequence differences evident in the U1 and U6 snRNAs that we have identified in the three different *Giardia* isolates either occur in single-stranded regions or alternatively show compensatory mutations or formation of G•U wobble pairs that maintain the base-pairing interactions in the predicted secondary structures (Figure 3.10).

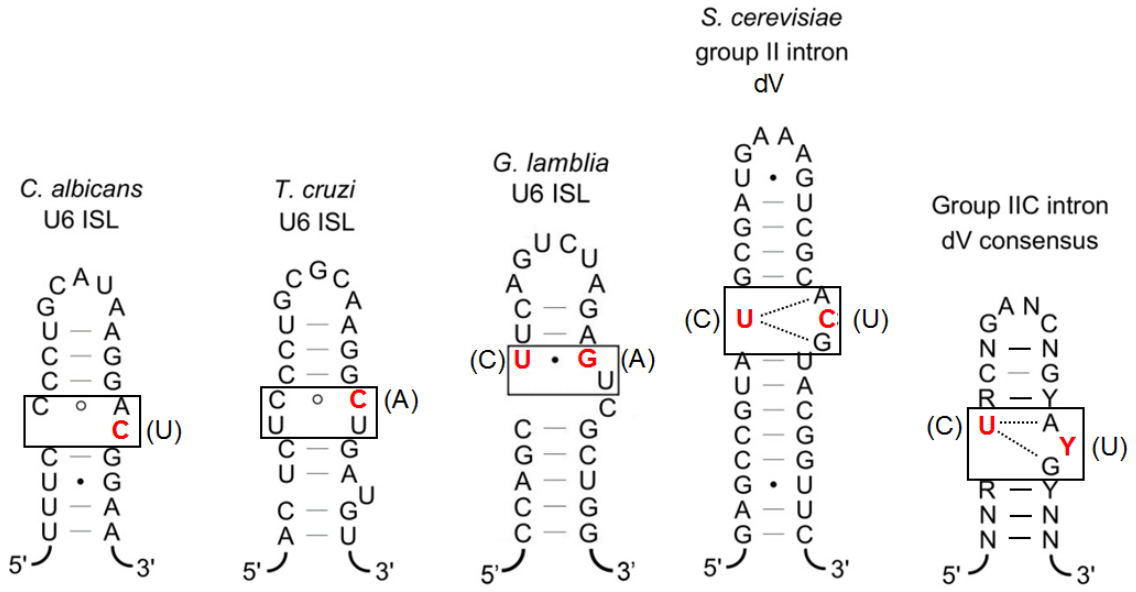


Figure 3.9. Sequence divergence within the intramolecular stem loops (ISL) in U6 snRNAs from diverse eukaryotes compared to domain V of group II introns.
Predicted secondary structures for the U6 ISL from the human pathogens *C. albicans* (Mitrovich and Guthrie 2007), *T. cruzi* (Ambrosio *et al.* 2007) and *G. lamblia* (this study). These are compared to domain V (dV) from yeast mitochondrial ai5γ group II intron (Zhang and Doudna 2002) and the group IIC intron consensus (Michel *et al.* 2009). Non-canonical interactions in the predicted Mg^{2+} binding sites are highlighted with boxes where those nucleotides differing from the eukaryotic consensus are indicated in red text, with eukaryotic consensus nucleotides at those positions indicated in brackets.

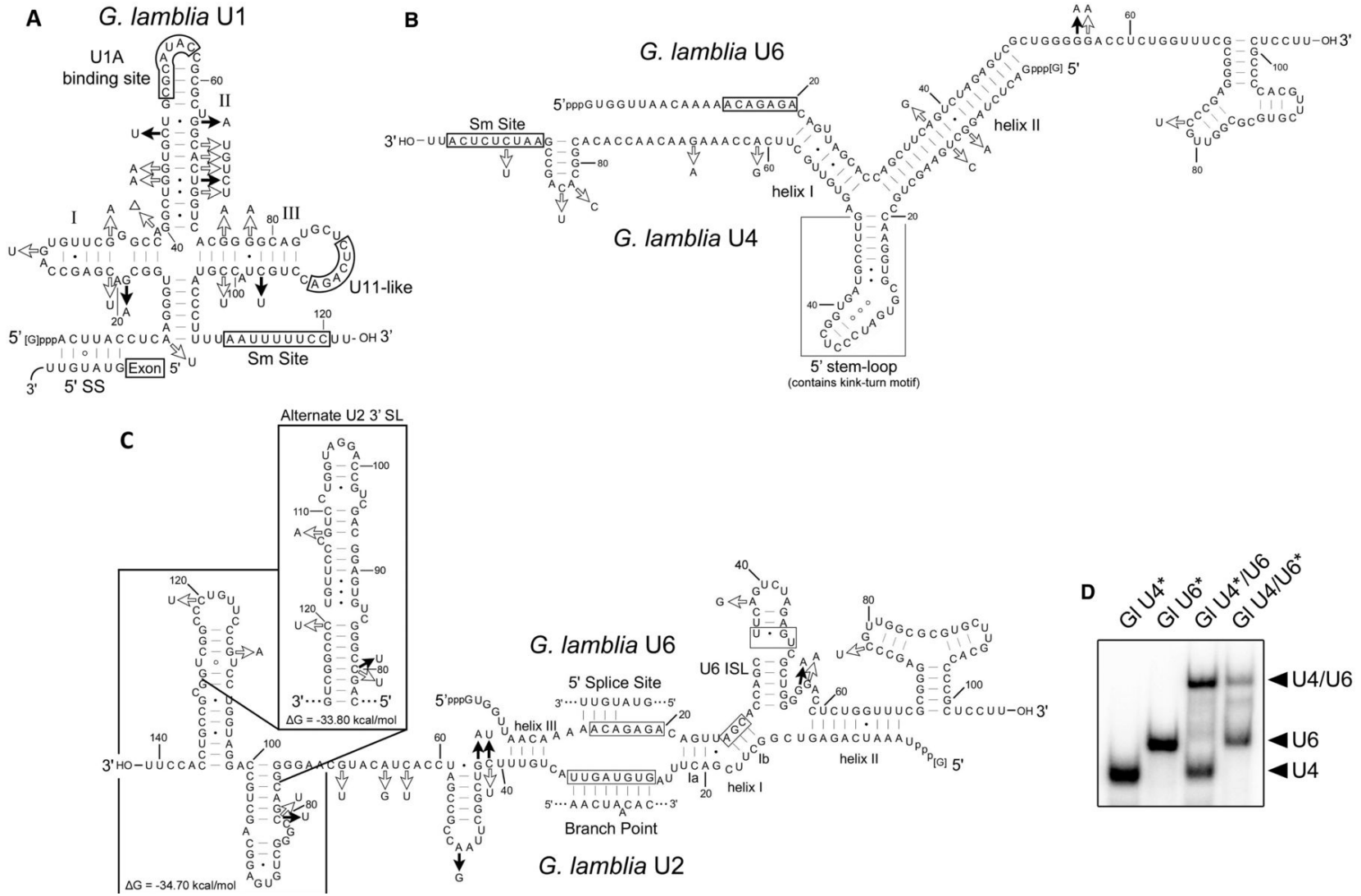


Figure 3.10. Evolutionarily divergent spliceosomal snRNAs in *G. lamblia*.

Figure 3.10. Evolutionarily divergent spliceosomal snRNAs in *G. lamblia*.
MFOLD secondary structural predictions for *G. lamblia* WB isolate snRNA candidate sequences are shown with arrows denoting nucleotide sequence differences observed in the P15 (filled arrows) and GS (open arrows) isolates. The snRNAs are shown base pairing to the 5′ splice site (**A**) and branch-point sequence (**C**) of the Hsp90 *trans*-spliced intron (Nageshan *et al.* 2011). The two alternative SL structures for the 3′ most terminal portion of the U2 snRNA candidate are shown (**C**, inset) with predicted free energies. In the *Giardia* U4/U6 interaction (**B**), the 5′ stem loop structure (boxed) contains a kink-turn motif. (**D**) *G. lamblia* U4/U6 snRNA transcripts form a complex *in vitro*. *Giardia* WB isolate *in vitro* synthesized U4 and U6 transcripts were incubated either individually or together and then fractionated by 6% native PAGE and visualized by autoradiography. The asterisk indicates which transcript is radioactively ^{32}P end-labelled.

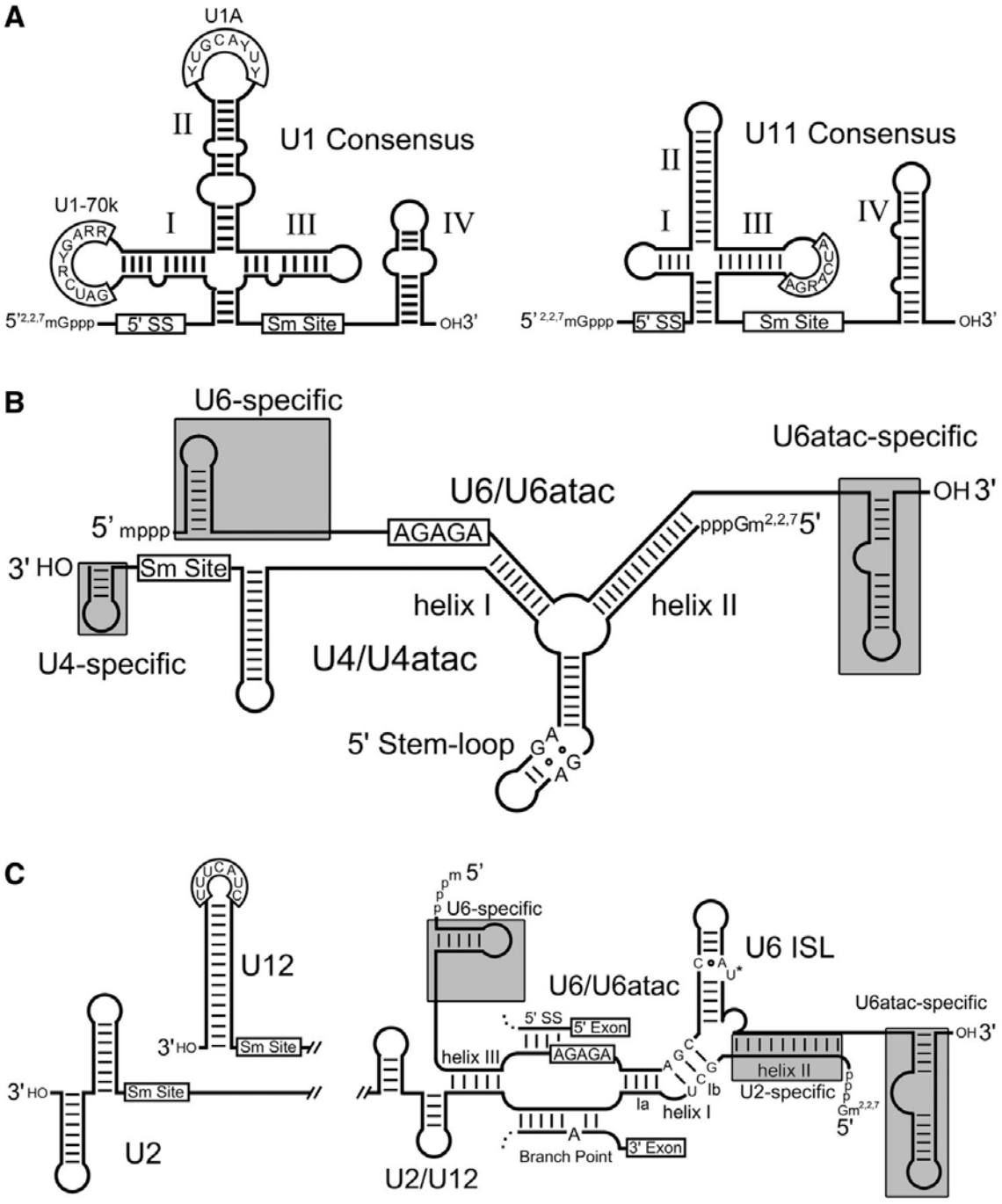


Figure 3.11. Comparison of major and minor spliceosomal snRNAs.
Consensus secondary structures and conserved sequence elements (open boxes) characteristic of major and/or minor spliceosomal snRNAs are shown. (**A**) Comparison of major spliceosomal U1 to minor spliceosomal U11 snRNA, where the 5′ splice site interacting sequence (5′ SS), Sm protein binding (Sm site) or U1 specific protein binding sites are boxed. A conserved sequence of unknown function in SL III of U11 is also boxed. (**B** and **C**) Structures that distinguish major (U2, U4 and U6) and minor (U12, U4atac and U6atac) spliceosomal snRNAs are highlighted in grey boxes. The

snRNAs (excluding U6) are shown containing 5′ trimethylguanosine caps as occurs in most characterized eukaryotes.

3.3.6 Identification of novel *G. lamblia* U2 and U4 snRNA candidates

The surprising finding of a novel U6 snRNA candidate stimulated us to search for the interacting *G. lamblia* U2 and U4 snRNA candidates. In the U6/U4•U5 tri-snRNP particle, U6 snRNA forms an evolutionarily conserved three-way helical junction with U4 snRNA involving nucleotides immediately downstream of the 'ACAGAGA' sequence in U6 snRNA. This region of U6 base pairs extensively to the 5′ half of U4 snRNA (Wahl *et al.* 2009) (Figures 3.10B and 3.11B). Formation of the intermolecular U4/U6 snRNA helices I and II stabilizes a 5′ SL in U4 snRNA that contains the RNA structural motif known as the kink-turn (Vidovic *et al.* 2000, Klein *et al.* 2001). Although the primary sequence of nucleotides involved in U4–U6 snRNA intermolecular pairing are conserved in U6 snRNAs, the interacting region of U4 is somewhat more divergent (Simoes-Barbosa *et al.* 2008), but maintains the ability to form intermolecular helices I and II that are critical to spliceosome function (Vankan *et al.* 1992, Wolff and Bindereif 1993). Using the sequence downstream of the '$_{14}$ACAGAGA$_{20}$' sequence in the *G. lamblia* U6 snRNA candidate as query, we searched the library of previously characterized *Giardia* ncRNAs of unknown function for a potential U4 snRNA candidate that could base pair with the U6 snRNA candidate. Our analyses revealed that the ncRNA Candidate-11 (Chen *et al.* 2007) is capable of forming the extended base-pairing interaction with the *Giardia* U6 snRNA candidate, generating a canonical helical junction containing helices I and II of similar length to those found in other eukaryotic U4/U6 snRNA complexes. The predicted interaction is further substantiated by compensatory mutations in the U6 and U4 snRNA sequences from the *G. lamblia* GS isolate that maintain contiguous base-pairing between

the two molecules (Figure 3.10B). Furthermore, the intermolecular pairing between the *Giardia* U6 and U4 snRNA candidates would allow for the formation of a 5′ SL in the U4 snRNA candidate containing a typical kink-turn motif with the sheared G-A base pairs (Klein *et al.* 2001). Inspection of the potential interaction of the previous U4 and U6 candidates (Chen *et al.* 2008) reveals significantly weaker base-pairing potential between the two molecules, particularly intermolecular helix II, and no conserved kink-turn motif in the U4 snRNA 5′ SL region (compare Figure 3.10B and Figure A.2.7F). In eukaryotes, the protein Snu13p is an important U4 and box C/D snoRNP assembly factor that binds to kink-turns found in box C/D RNAs and U4 snRNA. The previous identification of a well-conserved *G. lamblia* Snu13p homologue (Russell *et al.* 2005) that has been experimentally demonstrated to interact with kink-turn motifs (Biswas *et al.* 2011) predicts its role in *G. lamblia* U4 snRNP assembly and is consistent with finding a conserved kink-turn in our new U4 snRNA candidate.

Next, we assessed the ability of the U6 and U4 snRNA candidates to form a complex *in vitro*. Using full-length in vitro synthesized transcripts of complete U4 and U6 sequences (as determined by the 5′ and 3′ RACE end mapping experiments), we performed gel mobility shift assays using either radioactively end-labelled U4 incubated with unlabelled U6 or labelled U6 incubated with unlabelled U4. In both cases, U4/U6 complexes were readily observed (Figure 3.10D).

After the recruitment of the U6/U4•U5 tri-snRNP to the intron, U6/U4 intermolecular base pairs are unwound, allowing for the formation of U2/U6 snRNA intermolecular helices I through III and the U6 snRNA ISL (Will and Luhrmann 2011). During this remodelling, the ‘ACAGAGA’ sequence in U6 snRNA forms base pairs with

the intron 5′ splice site, and a sequence in U2 snRNA pairs with the branch point sequence, juxtaposing intron elements for the first transesterification reaction. Using the same strategy as described for identification of the U4 snRNA candidate, we determined that the previously identified ncRNA Candidate-14 of unknown function (Chen *et al.* 2007) is capable of forming all of the conserved interactions with the U6 snRNA candidate, displaying close adherence to those interactions characterized in other eukaryotes. This includes a discontinuous helix I containing the U6 'AGC' nucleotides implicated in magnesium ion binding (Yuan *et al.* 2007) (Figure 3.10C). The U2 snRNA candidate exhibits an extended and canonical interaction with the highly conserved *Giardia* intron branch-point sequence (AACUAACAC, branch point 'A' underlined) found within the ten currently identified *G. lamblia* introns. This interaction is significantly different from the interaction between the previously proposed *G. lamblia* U2 snRNA candidate (Chen *et al.* 2008), and the conserved branch point sequence in which only a limited base-pairing interaction was possible that unexpectedly only includes intron nucleotides 5′ upstream of the branch point 'A' residue involved in the first transesterification reaction (Figure A.2.7B). Nucleotide changes observable in the novel U6 and U2 snRNA candidates from the *G. lamblia* GS and P15 isolates occur in predicted single-stranded regions; therefore, the intermolecular U2/U6 snRNA interactions are strictly maintained in the other *Giardia* isolates. Curiously, the predicted interaction of our U2 snRNA candidate with *G. lamblia* intron BP sequences may bulge either two intron adenosines. This is also consistent with our earlier BP nucleotide mapping experiments in which either adenosine may participate in intron branching (Figure 3.4C).

Next, we assessed the expression of the U2 and U4 snRNA candidates by primer extension and northern blot analysis (Figure 3.5A and B) and mapped their mature ends using 5′ and 3′ RACE. The experiments generated products of expected size for the *G. lamblia* U2 and U4 snRNA candidates. When using excess oligonucleotide primer, similar signal intensities for extended cDNA products from U2, U4 and U6 snRNAs are observed (Figure 3.5A), indicating all three snRNA candidates are likely present at similar levels *in vivo*. These experiments also seem to indicate size homogeneity (discrete ends) for each mature snRNA species. Finally, BLASTN searches of individual snRNA sequences against *G. lamblia* WB, P15 and GS isolate genomes identified only one match per genome, indicating these snRNAs are encoded by single copy genes.

Collectively, these data strongly suggest a functional role for the previously identified but uncharacterized ncRNA candidates as authentic *G. lamblia* U2 and U4 spliceosomal snRNAs. Identification of U2 and U4 snRNA candidates capable of forming evolutionarily conserved base-pairing interactions with the U6 snRNA candidate and conserved intron sequence elements also further validates our correct identification of a bona fide *Giardia* U6 snRNA.

3.4 Discussion

3.4.1 *Giardia* snRNA candidates are evolutionarily divergent with properties of U2-type major and U12-type minor spliceosomal snRNAs

The identified snRNA candidates display the core structural features and nucleotide motifs conserved in spliceosomal snRNAs; however, they have noteworthy structural

simplification lacking some of the evolutionarily conserved domains. Curiously, all show features also resembling U12-type (minor) spliceosomal snRNAs.

Nucleotide co-variation in the U1 snRNA sequences from *Giardia* WB, P15 and GS isolates strongly support the proposed cloverleaf secondary structure with SLs I to III having lengths similar to U1 snRNAs found in other eukaryotes (Figure 3.10A). The 5′ terminal sequence '$_{1}$ACUUAC$_{6}$' is predicted to form base pairs with the conserved 5′ splice site found in *G. lamblia* introns ('[G/A/C]UAUGUU') similar to the interactions that occur in *S. cerevisiae*, and a conventional Sm protein binding site is located at the 3′ end of the RNA. Beyond these features, the U1 snRNA candidate is divergent, lacking SL IV and a recognizable U1-70 kDa protein binding sequence (AUCACGAA) (Surowy *et al.* 1989). In fact, the shortened *Giardia* SL I loop sequence is more similar in size to the loops observed in the corresponding regions of U11 snRNAs. Even more intriguing is that the *Giardia* U1 snRNA candidate contains a SL III loop sequence '$_{87}$CUCAGA$_{92}$', which is similar to the conserved 'AUCARGA' sequence of unknown function which we note in the equivalent region of U11 snRNAs from diverse eukaryotes (Figure 3.11 and Figure A.2.8). The *Giardia* U1 snRNA candidate SL II sequence '$_{51}$CGCAUAC$_{57}$' (boxed, Figure 3.10A) is conserved between *Giardia* isolates and divergent relative to the eukaryotic consensus (UGCACUC, identical positions in bold) (Guthrie and Patterson 1988). Interestingly, it most closely resembles the U1A binding site sequence present in *S. cerevisiae* U1 snRNA (CACAUAC) (Mitrovich and Guthrie 2007); however, it is also akin to the sequence present in *T. vaginalis* U1 snRNA (UGCAUAU) (Simoes-Barbosa *et al.* 2008), the most closely related eukaryote to *Giardia* in which snRNAs have been characterized. The apparent lack of a U1-70 kDa binding site and a divergent U1A protein binding site

sequence prompted us to search for homologues of these proteins in *G. lamblia*. Consistent with previous reports (Collins and Penny 2005), we could not identify clear homologues for either U1-70 kDa or U1A, or U11 snRNP-specific minor spliceosomal proteins in *G. lamblia* (Russell *et al.* 2006), suggesting these proteins are either highly divergent or absent.

Analysis of the *G. lamblia* U6 candidate in complex with U4 also reveals some intriguing similarities to U12-dependent spliceosomal snRNAs (Figures 3.10B and 3.11B). The U6 candidate lacks the upstream U6 snRNA-specific SL I (boxed, Figure 3.11B), a structure which is not present in minor spliceosomal U6atac snRNAs (Patel and Steitz 2003), and instead has a 5′ end position identical to U6atac RNAs (Figure A.2.9C). Likewise, we note that the extreme 5′ terminal sequence of the *Giardia* RNA most closely resembles U6atac RNAs. Additionally, the *Giardia* U6 snRNA candidate has an extended 3′ terminal region containing a terminal complex SL structure, more characteristic of U6atac snRNAs (Padgett and Shukla 2002) and a structure usually not present in U6 snRNAs. The *Giardia* U4 snRNA candidate lacks a 3′ terminal SL downstream of its predicted Sm protein binding site, which is also absent in U4 snRNA from *S. cerevisiae* (Brow and Guthrie 1988) and *C. albicans* (Mitrovich and Guthrie 2007). Interestingly, in metazoans, this terminal SL is present in U4 snRNA (Vankan *et al.* 1992) but not in minor spliceosomal U4atac snRNA (Padgett and Shukla 2002). At the primary sequence level, the *Giardia* U4 is divergent but displays some similarity to U4 and U4atac RNAs of other species (Figure A.2.9B).

Inspection of the *Giardia* U2 snRNA candidate and the U2/U6 interaction also shows major/minor spliceosomal characteristics. For example, the 5′ terminal nucleotides

of the U2 snRNA candidate are predicted to form the extended nine base pair helix II (Figure 3.10C) with the U6 snRNA candidate, which is typically observed in major spliceosomal U2/U6 snRNA complexes (Madhani and Guthrie 1992) but not in the minor spliceosomal U6atac/U12 snRNA counterpart (Shukla and Padgett 1999). The U2 candidate nucleotides in the region forming U2/U6 helices I and III and comprising the intron branch-point interacting sequence (Figure A.2.9A-1 and A-2) also show significantly higher sequence identity to other U2 snRNAs than to U12 RNAs (e.g. 26/39 identical nucleotides when comparing *Giardia* U2 positions 13–51 to the human U2 sequence). The downstream 3′ half of the *G. lamblia* U2 snRNA candidate is somewhat unusual, as it seems to lack a canonical Sm protein binding site before the 3′ terminal SL element(s). Sm protein homologues have been identified in *G. lamblia*, and phylogenetic analysis indicates they are divergent (Nixon *et al.* 2002). It is, therefore, plausible that lineage-specific non-canonical Sm sites may exist. Secondary structural predictions indicate that the 3′ terminal ~70 nt of the U2 snRNA candidate may fold into two distinct structural conformations, with nearly identical predicted thermodynamic stabilities (Figure 3.10C). In one conformation, the 3′ terminus folds into a dual SL structure resembling SLs III and IV of U2 snRNAs (Figure 3.11C, U2) (Patel and Steitz 2003). In the other conformation, the *Giardia* U2 snRNA candidate forms a single extended SL element reminiscent of SL III of U12 snRNAs but lacking a recognizable U12 65 kDa protein binding site sequence (CUACUUU) in the loop region (Benecke *et al.* 2005) (Figure 3.11C, U12), consistent with the lack of a recognizable coding region for this protein in the *Giardia* genome. The intriguing possibility exists that both conformations may be functionally relevant in

Giardia, and this observation further emphasizes the 'major/minor' hybrid nature of the *Giardia* snRNAs.

Examination of features of the predicted *Giardia* U6 snRNA candidate ISL region shows noteworthy differences from typical U6 ISL structure (Figures 3.10C and 3.11C). The Mg^{2+} binding site in U6 ISL usually contains a non-canonical C•A wobble pair and bulged uridine residue involved in metal-ion coordination (Huppler *et al.* 2002), a feature present in U6 and U6atac snRNAs (Patel and Steitz 2003). The *G. lamblia* U6 snRNA candidate is instead predicted to contain a U•G wobble pair followed by bulged uridine and cytidine (we note that alternative pairing interactions are also possible). In trypanosome species, sequence variations are also observed in the U6 ISL Mg^{2+} binding site (Figure 3.9), and curiously these organisms, like *Giardia*, have relatively few introns and can *trans*-splice precursor mRNAs. The *C. albicans* U6 ISL also differs by having a bulged cytidine instead of uridine. It seems that organisms containing relatively few introns and possessing more evolutionarily divergent spliceosomes display sequence variation in this region of U6. It is also interesting that the equivalent structural region of group II introns, the Mg^{2+} binding site within domain V, also show alternative sequences and non-canonical interactions (Figure 3.9).

In summary, the *Giardia* spliceosomal snRNAs show some novel characteristics, in particular, a surprising number of structural similarities to both major and minor spliceosomal snRNAs. The observation of highly conserved splice site sequence motifs in the currently identified *Giardia* introns that most closely match the consensus sequences of major (U2-type) introns would initially lead one to predict the existence of a major rather than minor spliceosome in *Giardia*. However, other features of these introns make their

classification less than straightforward. Recently, we identified an 'AT-AC' intron (Roy *et al.* 2012), and the first intron identified in *Giardia* in a ferredoxin gene was a 'CT-AG' intron (Nixon *et al.* 2002). The collection of characterized *Giardia* introns also show an apparent fusion of branch point and 3′ splice site sequences that highly constrains the distance between the branch point adenosine and 3′ splice site. These are features commonly observed in minor (U12-type) introns; therefore, like the snRNA candidates, the introns are also showing hybrid features of major and minor spliceosomal introns.

3.5 Conclusions

Collectively, we have used bioinformatic and molecular techniques to identify novel *G. lamblia* ncRNAs and characterize their expression and processing strategies, which to date are largely unexplored in this organism. In addition to identifying novel *G. lamblia* snRNAs, we find that a large number of *G. lamblia* ncRNAs (including snRNAs) are initially transcribed as longer mono- or di-cistronic precursors that are subsequently processed at the conserved 12 nt RNA sequence motif present at the 3′ downstream regions of mature ncRNAs. Surprisingly, we also identify motif sequences residing in the 5′ halves of the four known *G. lamblia trans*-introns, indicating an unexpected common RNA processing pathway for *Giardia* ncRNAs and *trans*-spliced introns. While not essential for the first step of *trans*-splicing, we speculate that such positioning of motif cleavage sites to liberate intron 5′ halves from longer precursor transcripts may allow for more efficient association of *trans*-intron halves for splicing, particularly when initially transcribed with downstream ORFs (e.g. Hsp90 exon1-intron 5′ half + replication factor C subunit 5; see Figure 3.2, Figure A.2.4).

Chapter 4: Conservation of Spliceosomal Intron Structures and snRNA Divergence in Diplomonad and Parabasalid Lineages

4.1 Introduction

Eukaryotic nuclear genomes contain spliceosomal introns which divide protein coding sequences on to separate exons. Exons must then be ligated during precursor mRNA splicing, before mRNA transit to the cytoplasm for protein translation. Spliceosomal introns have been identified in virtually all eukaryotes, however, intron density is remarkably variable in different species. Some intron-poor species contain only a few introns per genome (Morrison *et al.* 2007, Lee *et al.* 2010) whereas some intron-rich species have on average several introns per kilobase of gene sequence (Lander *et al.* 2001). Intron length may also differ substantially, from introns as short as 18 nt in the nucleomorph genome of *Bigelowiella natans* (Gilson *et al.* 2006) to mammalian gene introns which can reach many tens of kilobases in size (Lander *et al.* 2001).

Thus far, two separate classes of spliceosomal introns have been identified in eukaryotes: the major/U2-type and minor/U12-type spliceosomal introns. U2-type introns have been identified in all fully-sequenced nuclear genomes, whereas only a small subset of eukaryotes have been found to contain U12-type introns (Russell *et al.* 2006, Bartschat and Samuelsson 2010). However, the distribution of U12-type introns in evolutionarily-diverse eukaryotes reveals an ancient origin for U12-type introns and suggests they were likely present in the last eukaryotic common ancestor (LECA) (Russell *et al.* 2006, Bartschat and Samuelsson 2010).

Removal of U2-types introns is catalyzed by the major/U2-dependent spliceosome consisting of five evolutionarily conserved small nuclear RNAs (snRNAs) U1, U2, U4, U5 and U6 and dozens to several hundred spliceosomal proteins (Will and Luhrmann 2011).

U12-type introns are excised by a distinct minor/U12-type spliceosome that contains both shared U2-dependent and unique U12-dependent spliceosomal proteins, the common U5 snRNA and uniquely the U11, U12, U4atac and U6atac snRNAs which are functionally analogous to the U1, U2, U4 and U6 snRNAs, respectively (Patel and Steitz 2003). U2- or U12-type spliceosomal introns are distinguished by distinctive 5′ and 3′ splice sites (SS) and internal branch point (BP) sequence motifs that are recognized via specific RNA-RNA base-pairing interactions with U2- or U12-dependent spliceosomal snRNAs (Wahl *et al.* 2009).

Most spliceosomal introns are only positionally-conserved in closely-related taxa. However, ~25% of introns in *Arabidopsis thaliana* occupy the same position in orthologous genes in humans (Rogozin *et al.* 2003) and some of the introns in intron-reduced protist species show conservation in orthologous genes from distantly-related, intron-rich species (Russell *et al.* 2005, Vanacova *et al.* 2005). For example, the *Rpl7a* gene intron in *Giardia lamblia* is also present in orthologous genes in animals and some Amoebozoans, suggesting that this intron was present in a common ancestor of the Excavata and Unikont eukaryotic 'supergroups' (Russell *et al.* 2005). Reconstruction of ancestral eukaryotic intron density based on patterns of intron gain and loss in 99 eukaryotes suggests that the last eukaryotic common ancestor (LECA) was probably intron-rich (Csuros *et al.* 2011) and already endowed with a complex spliceosomal apparatus (Collins and Penny 2005). The identification of ancient spliceosomal introns with functions conserved in diverse eukaryotes would indicate very early intron fixation and beneficial intron function.

Diplomonads are a group of early-branching protists with characterized species containing highly reduced nuclear genomes and apparently few spliceosomal introns

(Morrison *et al.* 2007, Xu *et al.* 2014). The first ~10 characterized spliceosomal introns in diplomonads were identified in *G. lamblia* and they contain extended highly-conserved 5′ splice site sequences, with fused BP and 3′ SS sequences (Nixon *et al.* 2002, Russell *et al.* 2005, Kamikawa *et al.* 2011, Roy *et al.* 2012). *Trichomonas vaginalis*, a parabasalid (diplomonad sister group), shares the same general spliceosomal intron structure and splice site sequence motifs as *G. lamblia* (Vanacova *et al.* 2005) and *a priori*, one would therefore predict that other diplomonads will share these conserved intron features.

In this study, we used bioinformatics to identify spliceosomal introns in another diplomonad species, *Spironucleus vortens*. We first identified a set of introns by specifically examining ribosomal protein (RP) genes which then allowed us to design search parameters to identify additional introns in this organism. Intron consensus sequences from *Spironucleus* introns then aided in bioinformatic prediction of U1, U2 and U5 spliceosomal snRNAs (snRNAs) from *S. vortens* and *Spironucleus salmonicida* genomic sequences. We find striking conservation of intron structural properties in both diplomonads and a parabasalid and observe that many of the remaining spliceosomal introns in diplomonads are ancient.

4.2 Materials and Methods

4.2.1 Identification of *S. vortens* spliceosomal introns

Intron-poor eukaryotic genomes may have their introns concentrated within ribosomal protein coding genes (Bon *et al.* 2003, Russell *et al.* 2005), thus we reasoned that spliceosomal introns may possibly interrupt RP genes in *S. vortens*. Consequently, the complement of 80 ribosomal protein sequences from *Saccharomyces cerevisiae* was

downloaded from the Ribosomal Protein Gene Database (http://ribosome.med.miyazaki-u.ac.jp/) (Nakao *et al.* 2004) and each RP sequence was used as query in TBLASTN searches against the *S. vortens* expressed sequence tag (EST) library on the NCBI website and matching ESTs encoding RP sequences were obtained. In most cases, these searches unambiguously identified a matching *S. vortens* RP ortholog, however, 11 of the 80 *S. cerevisiae* RP proteins sequences did not identify obvious RP gene orthologs (data not shown). Next, the *S. vortens* RP EST sequences were used as queries in BLASTN searches against the *S. vortens* genomic sequences from the NCBI trace archive and for positive hits, 500 nt of additional upstream and downstream sequence was also downloaded (when possible). Genomic trace sequences were then aligned with corresponding ESTs manually and inspected for introns disrupting coding sequences. These analyses identified the *Rpl7a*, *Rpl30*, *Rps4*, *Rps12* and *Rps24* RP gene introns.

To identify additional (and possible non-RP gene) introns, we utilized the pattern-matching software 'Scan for Matches' (Dsouza *et al.* 1997) in conjunction with the newly-identified *S. vortens* fused branch point and 3′ SS sequence consensus: 5′-RCTAACAARYTAG-3′ obtained from the identified RP gene introns. *S. vortens* raw genomic sequence reads were downloaded from the NCBI trace database (130 Mb genomic sequences) and made into a concatenated file which served as the local database for our searches. Next, we searched our local database using Scan for Matches and the pattern: 500…RCTAACAARYTAG…500 (where 'R' and 'Y' represent a purine and a pyrimidine, respectively). We examined the hits for the presence of a potential 5′ SS in the regions upstream of the BP/3′ SS sequence. Next, sequences from the region downstream of the BP/3′SS from each unique hit were translated in the three possible reading frames using the

ciliate genetic code (usual stop codons TAA and TAG codons are instead glutamine) (Keeling and Doolittle 1997) and used as queries in BLASTP searches against the non-redundant protein sequence database on the NCBI website to determine if they encoded conserved protein coding sequences. This strategy identified the *FolC-like* gene intron, the predicted hypothetical gene introns and the predicted *Rps15* gene 5′ UTR intron.

4.2.2 Bioinformatic prediction of *Spironucleus* snRNAs

Spliceosomal small nuclear RNAs (snRNAs) were predicted in *S. vortens* and *S. salmonicida* genomic sequences using a combination of sequence motif and covariation model (CM) search strategies. Initially, optimized alignments of U1, U2, U4, U5 and U6 snRNA sequences from phylogenetically-diverse eukaryotes were downloaded from the Rfam database (http://rfam.xfam.org/) and used to generate CMs using the cmbuild tool from the Infernal software package (Nawrocki and Eddy 2013). Next, individual U-snRNA CMs were employed in cmsearch (Infernal software package) queries to identify snRNA-like sequences in *S. vortens* and *S. salmonicida* local DNA databases, with cmsearch *E* value cut-offs set to 10. In anticipation that *Spironucleus* snRNAs may be highly divergent (as observed for the *G. lamblia* snRNAs, see Chapter 3), all resulting cmsearch hits were examined manually for evolutionarily-conserved secondary structures (e.g. a ‘cloverleaf’ structure for U1 snRNA) or expected sequence motifs (e.g. 5′ SS binding sequence for U1 or BP interacting sequence for U2 snRNA).

U5 snRNA candidates were identified using Scan for Matches queries specifying the conserved U5 snRNA loop I sequence ‘UGCCUUUUACY’ (allowing two mismatches) flanked by nucleotides capable of forming a 6 base pair helix (allowing G•U wobble pairs). For each hit, 100 nt of upstream and downstream sequence was then examined for the

ability to form a longer stem-loop I consisting of conserved 1a/1b/1c helices and IL1 and IL2 internal loops, and the presence of a canonical Sm binding site ($RAU_{4-6}GR$, where R is a purine).

4.2.3 Intron secondary structure and RP gene intron conservation in eukaryotes

To identify possible conserved intron secondary structures, the collection of *S. vortens cis*-introns identified here and annotated *T. vaginalis* introns (retrieved from TrichDB.org) were used as input for MFOLD (Zuker 2003) secondary structure predictions. MFOLD parameters were modified to force intron regions predicted to interact with spliceosomal machinery (i.e. 5′ SS, BP and 3′ SS) to be single stranded and RNA folding temperatures were set to either 21°C or 37°C for *S. vortens* and *T. vaginalis* introns, respectively, based on the optimal growth temperatures for each organism. For each intron, the top three MFOLD energy secondary structural predictions were then examined for secondary structural potential (i.e. extended helices) and total single-stranded distances (excluding loop nucleotides) were determined.

To determine the phylogenetic conservation of *S. vortens Rps4* and *Rps24* introns in eukaryotes, orthologous *Rps4* and *Rps24* genes from representative eukaryotes were examined for intron insertion at the same relative position as *S. vortens* using the gene browser tool on the NCBI website (http://www.ncbi.nlm.nih.gov/gene/). Only introns found in the same phase and relative position of the RP gene-coding sequences were considered to be homologous introns. RP gene intron distribution was then mapped using a recent proposed eukaryotic tree from Burki (2014).

4.3 Results

4.3.1 Spliceosomal introns in RP and non-RP genes in *S. vortens*

Only eleven spliceosomal introns have been identified in the diplomonad *Giardia lamblia* (Nixon *et al.* 2002, Russell *et al.* 2005, Morrison *et al.* 2007, Kamikawa *et al.* 2011, Nageshan *et al.* 2011, Roy *et al.* 2012, Kamikawa *et al.* 2014), revealing both a remarkable paucity of spliceosomal introns and also proportionally large number of *trans*-spliced introns in this organism. Despite this, very little is known about spliceosomal intron structure from any member of the diplomonad genus *Spironucleus*. However, very recently, sequencing the genome of *S. salmonicida* uncovered three experimentally-confirmed *cis*-spliceosomal introns in genes encoding ribosomal proteins L30 (*Rpl30*) and S24 (*Rps24*) and a hypothetical protein (Xu *et al.* 2014). To further expand our knowledge of spliceosomal intron structure in diplomonads, we searched the preliminary nuclear genomic DNA sequence data from *Spironucleus vortens* for spliceosomal introns. Initially our search strategy employed the conserved *G. lamblia* 5′ splice site (SS) sequence 'VTATGTT' and fused branch point (BP) and 3′ SS sequence 'VCTRACACRCAG' ('R' is a purine; 'V' is an A, C, or G nucleotide) (Roy *et al.* 2012), but these searches did not identify any introns in *S. vortens*. Thus, we reasoned that intron splice site sequences differ in *S. vortens* compared to *G. lamblia* and *T. vaginalis* introns.

Ribosomal protein (RP) genes are highly-conserved protein-coding sequences readily recognizable in eukaryotic genomes. Notably, some intron-poor eukaryotes (e.g. *S. cerevisiae*) contain a large proportion of their spliceosomal introns within RP genes (Spingola *et al.* 1999, Bon *et al.* 2003) and the few *cis*-spliced introns in *G. lamblia* and *S. salmonicida* interrupt RP genes (Russell *et al.* 2005, Xu *et al.* 2014). Therefore, we determined whether protein coding continuity in RP genes is interrupted by one or more spliceosomal introns in *S. vortens*. RP genes have not been previously annotated in *S. vortens*, so we initially performed homology searches using the 80 RP genes from *Saccharomyces cerevisiae* (Nakao *et al.* 2004) as queries for TBLASTN searches against the *S. vortens* raw genomic sequence data. These searches identified 69 predicted RP gene sequences in *S. vortens* (data not shown). Next, the *S. vortens* RP gene sequences were individually aligned with corresponding expressed sequence tag (EST) data to determine if they contained intervening sequences not present in mature mRNAs. This analysis identified single spliceosomal introns interrupting conserved regions of RP genes *Rpl7a*, *Rpl30*, *Rps4*, *Rps12* and *Rps24* (Figure 4.1A and A.3.1 and Table A.3.1). The *S. vortens Rpl30* and *Rps24* introns occur at the same positions as in the *S. salmonicida* orthologs and in all cases, intron sequences contain an in-frame stop codon and/or introduced a frame shift (in the downstream coding region) that would result in a truncated ribosomal protein (Figure A.3.1). We also observed *S. vortens* RP gene sequence variants during the analysis that appear to be allelic variants, based on the high-level of nucleotide sequence similarity (Figure A.3.2) and identical chromosomal context. Allelic variants of the *Rpl7a*, *Rps4* and *Rps12* genes contained intron sequence differences and thus were included in the subsequent intron analyses (Figure 4.1A and A.3.2).

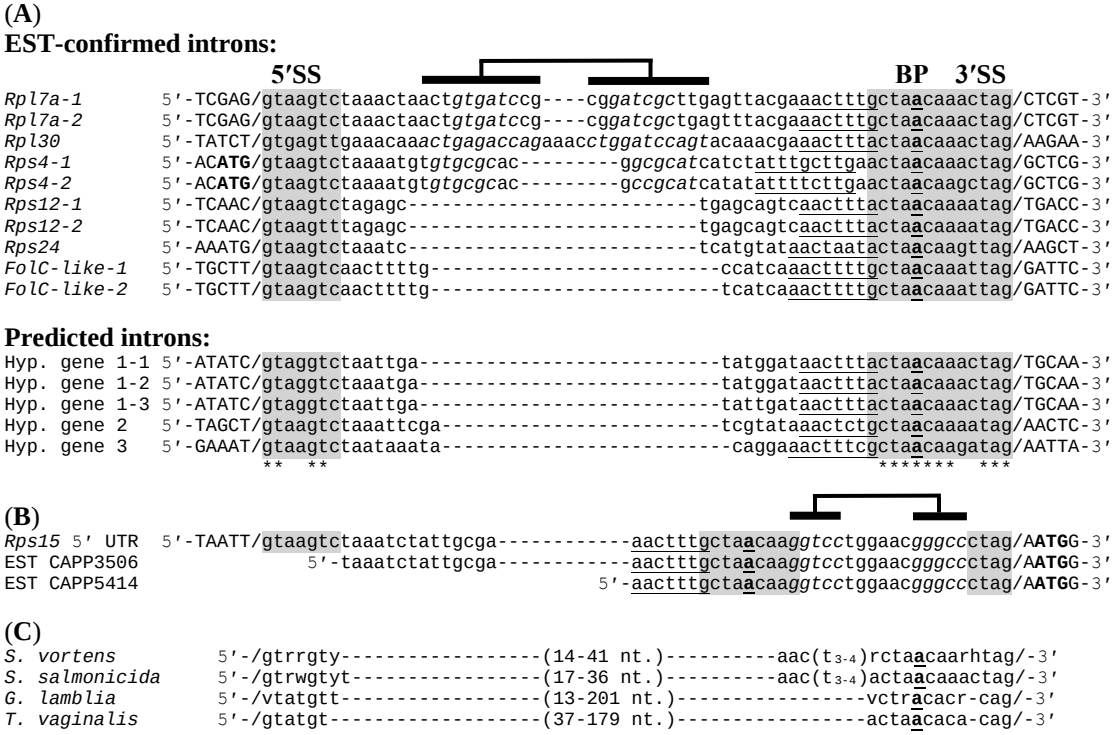


Figure 4.1. *Cis*-spliceosomal introns in *S. vortens*.
(**A**,**B**) EST-confirmed (RP and *FolC-like* genes) and predicted (hypothetical genes) spliceosomal intron sequences from *S. vortens* genomic sequences were aligned using ClustalW2 software (see Table A.3.1 for gene sequences and accession numbers). The 5 nt of exonic sequences flanking each intron are in upper case with a slash representing the exon-intron boundary and mRNA start 'ATG' codons in bold. Predicted intron 5′ and 3′ splice sites (SS) and branch point (BP) sequences are highlighted in grey with the putative reactive branch point adenosine in bold and underlined. A conserved pyrimidine-rich motif ('AAC[T/C]$_{3-4}$R') found upstream of the branch point sequence is underlined. Nucleotide identities shared between all aligned introns are indicated by asterisks under the alignment. Potential base pairing nucleotides within introns are in italics with black bars above alignments indicating interacting regions. (**B**) A predicted 5′ UTR intron in the *S. vortens Rps15* gene contains a 15 nt insertion interrupting the fused BP and 3′ SS sequence. Two ESTs confirm the accumulation of precursor mRNAs containing the unspliced intron. (**C**) Consensus sequences from the identified *S. vortens* spliceosomal introns are compared to those from the related diplomonad *Spironucleus salmonicida* (Xu *et al.* 2014), *Giardia lamblia* (Roy *et al.* 2012) and parabasalid *Trichomonas vaginalis* (Vanacova *et al.* 2005). An 'R' indicates a purine, 'Y' is a pyrimidine, 'W' is A/T, 'V' is A/C/G and 'H' is A/C/T.

Our non-sequence biased identification of introns in *S. vortens* RP genes revealed a consensus sequence 'GTAAGTY' at the 5′ SS, and a branch point fused to 3′ SS sequence, 'RCTAACAARHTAG' (predicted BP 'A' is underlined, 'R' is purine, 'Y' is pyrimidine, 'H' is A, C or T) (Figure 4.1). Using these identified conserved intron sequence features,

we next searched for additional (non-RP gene) introns in genomic sequences by employing the sequence pattern matching program 'Scan for Matches' (Dsouza *et al.* 1997). This search strategy uncovered an additional EST-verified intron interrupting a 'bifunctional folylpolyglutamate synthase-like' (*FolC-like*) gene and three additional putative introns individually interrupting three different predicted protein-coding genes of unknown function (Figure 4.1A and Table A.3.1). Although intron insertion sites may be located outside of protein coding sequences, it should be noted that *S. vortens* only utilizes one stop codon in its genetic code (Keeling and Doolittle 1997) increasing the likelihood of inaccurate protein-coding gene prediction. In the absence of EST data to confirm gene expression and due to the lack of ORF conservation in other characterized species, these three additional putative introns will require further experimental verification.

The collection of *S. vortens* spliceosomal introns are short and relatively uniform in size, ranging from 40 to 67 nucleotides, and are primarily located proximal to the 5′ ends of the ORFs. The *Rps4* intron is located immediately downstream of the 'ATG' start codon, (a so-called 'start codon intron') an intron location often observed in the RP gene sequences of other eukaryotes (Nielsen and Wernersson 2006). We also note that most *S. vortens* introns are phase '0' introns (5 of the 6 EST-confirmed introns).

The *S. vortens* intron sequences display extended sequence conservation of intron splice sites. In addition to standard 'GT-AG' boundaries, the introns display a 7 nt conserved 5′ SS and 13 nt fused BP + 3′ SS sequence, akin to the spliceosomal introns in *G. lamblia* and those recently identified in *S. salmonicida* (Figure 4.1C). Interestingly, we also note that some of the *S. vortens* introns display substantial sequence similarity to each other in the internal region between the 5′ SS and BP + 3′ SS. For example, in an optimized

alignment of the *Rpl7a* and *Rpl30* introns there is ~70% nucleotide identity over the entire intron length (44 out of 63 nt positions), and ~63% identity (27 of 43 positions) when excluding the 5′ SS and BP + 3′ SS sequence elements from the comparison (Figure 4.1A). Furthermore, 3 of 5 RP gene introns and 2 of the introns in the putative protein-coding genes contain the sequence 'TAAA' starting at intron position +8 which would extend the 5′ SS consensus to 'GTARGTYTAAA' for these introns. Also evident is a recurring pyrimidine tract-containing sequence motif, directly upstream adjacent to the intron branch point sequence, with consensus sequence 'AAC[T/C]$_{3\text{-}4}$R' (Figure 4.1A, underlined). The *FolC-like* gene intron contains an additional copy of this motif downstream adjacent to its 5′ SS sequence (Figure 4.1A). Notably, the three confirmed *S. salmonicida* introns (Xu *et al.* 2014) also display the 'AAC[T/C]$_{3\text{-}4}$R' motif sequence (Figure 4.1C) and a similar A-T extended 5′ splice site motif 'GTATGTTTAAC.'

4.3.2 A 5′ UTR intron remnant in the *S. vortens Rps15* gene?

Based on intron sequence conservation, we also identified an intron-like sequence in the 5′ UTR region of the *Rps15* gene through a BLASTN search using the newly-identified *Rpl7a* intron as query. The sequence was a plausible intron candidate due to the presence of a canonical and extended 5′ SS sequence 'GTAAGTCTAAA', BP and pyrimidine-tract (underlined) motif sequence 'AACTTTGCTAACAA' (Figure 4.1B), as found in the *Rpl7a* intron (Figure 4.1A and B). However, unlike the other *S. vortens* introns, the candidate *Rps15* intron's 3′ SS sequence motif 'TAG' is not fused to the BP sequence and instead is displaced downstream by 15 nt. The distance between the BP 'A' and 3′ SS is a highly conserved property of all identified *G. lamblia* introns (Russell *et al.* 2005, Roy *et al.* 2012). Experiments in *T. vaginalis* in which the BP motif ('ACTAAC') was moved

2 or 7 nt upstream of its conserved position abolished splicing in an *in vivo* assay (Vanacova *et al.* 2005) indicating a requirement for the precise spacing of these intron elements in the splicing reaction mechanism in these organisms. We therefore predict that the 'inserted' sequence in the *Rps15* intron-like element prevents splicing of this region.

Closer examination of this insertion sequence reveals an inverted repeat that could form an RNA stem-loop element containing 5 consecutive base-pairs (Figure 4.1B, italic sequence) in the mRNA transcript. This would bring the BP and 3′ SS-like sequence in closer spatial proximity and suggests the alternative possibility of a functional role of the stem-loop element in splicing of this *Rps15* intron-like element. We examined the ESTs generated from *Rps15* transcripts, most of which are 5′ end-truncated, but found two ESTs that have 5′ end terminal sequences still containing *Rps15* intron-like element sequence (Figure 4.1B). This result is consistent with either an inability to splice this region or inefficient intron removal resulting in significant precursor mRNA accumulation and detection.

4.3.3 Base pairing potential in *S. vortens* and *T. vaginalis* introns

The collection of 'long' *cis*- and *trans*-introns in *G. lamblia* display extensive secondary structural potential which appears to constrain the single stranded distance between splice donor and acceptor sites to 35-45 nt – a similar length to the characterized short *G. lamblia cis*-introns (Figures 4.2B and 2.3). Therefore, we examined *S. vortens* introns for similar internal base pairing potential. Intriguingly, while the length of the 'short' *S. vortens* introns cluster uniformly at 40-42 nt, MFOLD secondary structure predictions indicate that the longer EST-confirmed *Rpl7a*, *Rpl30* and *Rps4* introns may form stable stem loops, thus bringing the splice sites within a similar ~41 nt single-stranded

distance (Figures 4.1A and 4.2). We also found that the 'long' *S. salmonicida Rpl30* intron (Xu *et al.* 2014) is capable of forming a stem loop, making its single-stranded length of 41 nt similar to the other short *S. salmonicida* introns (43 nt) (Figure A.3.3).

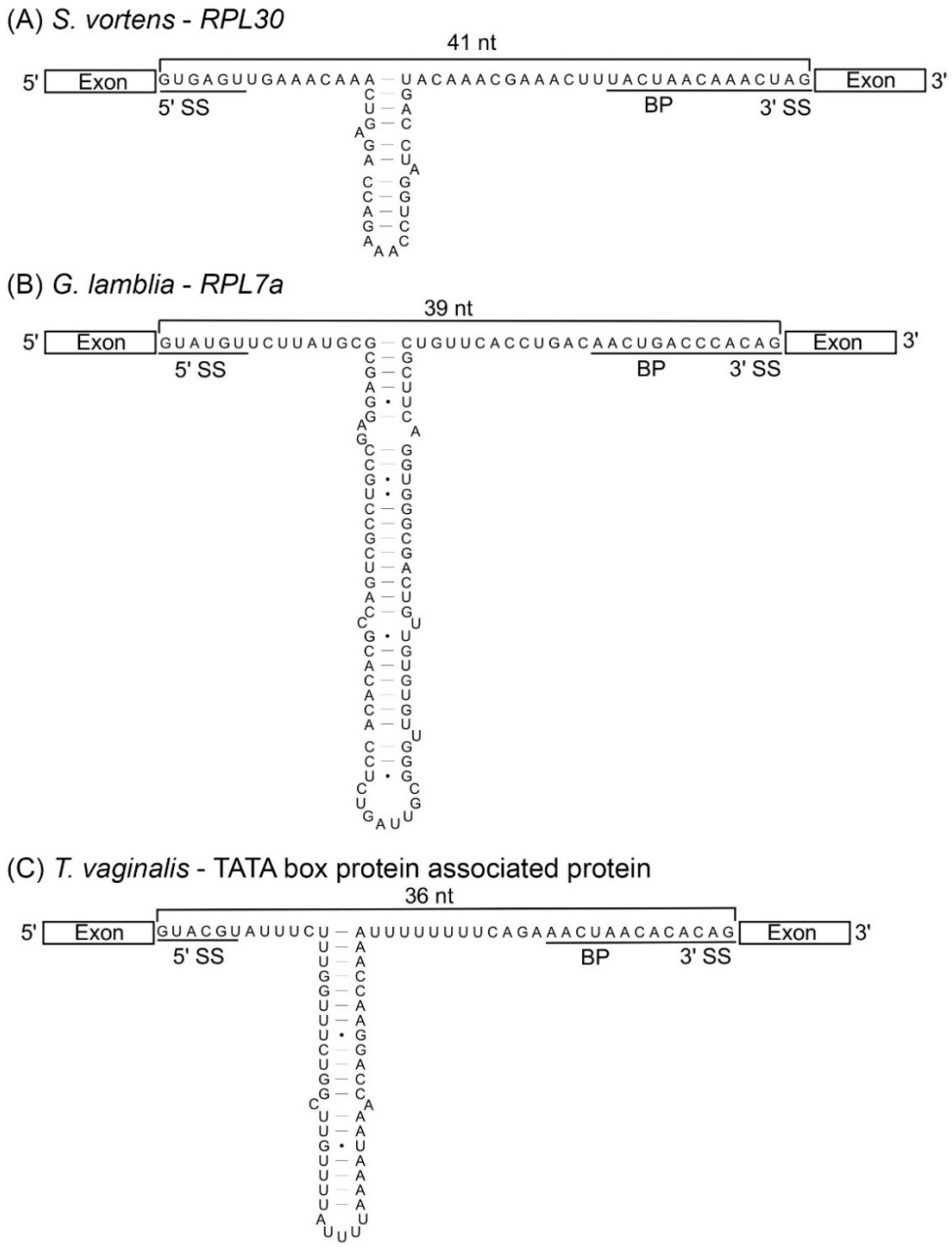


Figure 4.2. Base pairing of long *cis*-introns in diplomonads and a parabasalid.
Secondary structural predictions of representative *cis*-spliceosomal introns from *S. vortens*, *G. lamblia*, and *T. vaginalis* are shown with putative 5′/3′ splice site (SS) and branch point (BP) motifs underlined. Lengths of 'single-stranded' distances between splice donor and acceptor sites are indicated in nucleotides (nt) above intron sequences.

Stimulated by our discovery of structural potential in *Spironucleus* introns, we next examined secondary structural potential in *T. vaginalis* introns to assess whether similar intron base pairing is a more wide-spread phenomenon in other Metamonada. We found that most *T. vaginalis* introns were either uniformly short (25 nt) or else were longer (>50 nt) but possessed the ability to form extended stem loops, making intron single-stranded lengths between 25 and 44 nt (median length 37 nt) upon intron folding (Figures 4.2C and A.3.2).

4.3.4 Bioinformatic identification of *Spironucleus* spliceosomal snRNAs

During pre-mRNA splicing, intron substrates are recognized by the spliceosome in part by RNA-RNA intermolecular base pairing involving U1, U2, and U6 snRNAs with intron 5′ SS and BP sequences (Wahl *et al.* 2009). Using the newly identified *Spironucleus* intron 5′ SS and BP sequence motifs, we searched for snRNA-like sequences in *S. vortens* and *S. salmonicida*. Initially, we performed BLASTN searches using the *G. lamblia* U1, U2, U4 and U6 snRNAs (Hudson *et al.* 2012) (Figure 3.3) as queries against *S. vortens* and *S. salmonicida* genomic sequences; however, these searches did not yield any plausible snRNA candidates. Co-variation models (CMs) are probabilistic models that combine consensus RNA sequence and secondary structures from a known RNA family to identify similar sequences in a DNA/RNA database. Moreover, CMs have been successfully employed to predict snRNA-like sequences in genomic DNA sequences from diverse eukaryotes (Lopez *et al.* 2008). Thus, we performed CM searches for snRNA-like sequences in *Spironucleus* DNA sequences using the Infernal software package (Nawrocki and Eddy 2013) and CMs generated with U-snRNA sequences from the Rfam database

(Burge *et al.* 2013). These searches identified a potential U1 snRNA candidate for *S. vortens* and U2 snRNA candidates for *S. vortens* and *S. salmonicida* (Figure 4.3).

The *S. vortens* U1 snRNA candidate displays a characteristic 'clover leaf' secondary structure with conserved stem loops (SL) I to IV, typical of U1 snRNAs in most eukaryotes (Figure 4.3A). Canonical U1 snRNA binding sites for U1-70k (GAUCA) and U1A (AUUGCAC) (Pomeranz Krummel *et al.* 2009) proteins are evident in the loop regions of SL I and II, respectively as well as a predicted Sm protein binding site upstream of SL IV (Figure 4.3A). The 5′ end of U1 snRNA is expected to form base pairs with the intron 5′ SS and indeed we observe the 5′ end nucleotides of the *S. vortens* U1 snRNA 'ACUUAC' are complementary to the 'GURRGU' 5′ SS sequence consensus of *S. vortens* spliceosomal introns (Figures 4.1 and 4.3A). Interestingly, our searches also identified several other *S. vortens* U1 snRNA-like isoforms which contained noteworthy sequence changes (Figure A.3.4). We observe that several of the changes occur in functionally important regions of U1 (i.e. U1-70k and Sm protein binding sites) and/or destabilize secondary structures, suggesting that some of these isoforms may be non-functional pseudogenes.

Examination of the *S. vortens* and *S. salmonicida* U2 snRNA candidates reveals secondary structural features of U2 snRNAs from other representative eukaryotes, with identifiable SLs I, IIa/IIb and III and predicted Sm protein binding sites (Figure 4.3C and 4.3D). However, the SL IV found in most other U2 snRNAs (Patel and Steitz 2003) appears to be missing in both the *S. vortens* and *S. salmonicida* U2 candidates. The 5′ half of both *Spironucleus* U2 candidates contain branch point-interacting sequences that would generate the expected bulged intronic catalytic adenosine upon interaction (Figure 4.3C and

4.3D). We also note that the conserved U2 snRNA 'GCU' and 'GAUC' nucleotides involved in formation of U2-U6 intermolecular helix I (Burke *et al.* 2012) are conserved in both *Spironucleus* U2 candidates. Also, interestingly, while the first ~45 nt of the *S. vortens* U2 snRNA candidate displays high sequence conservation to the *S. salmonicida* U2 candidate (36/45 nucleotide identity), the remaining downstream sequences are divergent yet both maintain the ability to form structurally-conserved SL IIa/IIb and an extended SL III. CM searches did not identify any plausible U5 snRNA candidates. However, U5 snRNAs are typified by a long stem-loop containing the highly-conserved loop I sequence 'UGCCUUUUACY' involved in binding exons during the splicing reaction (Newman and Norman 1992). Thus, we reasoned that 'Scan for Matches' may be more successful in finding U5 snRNA-like sequences in *Spironucleus spp.* DNA sequences by searching for instances of the canonical loop I sequence motif or variants (allowing 2 substitutions), flanked by sequences capable of forming a 6 bp apical stem Ic expected in U5 snRNA structures. One pattern match in *S. salmonicida* displayed a perfect loop I sequence (UGCCUUUUACU) and upon closer examination, was capable of not only forming stem Ic, but also a canonical extended SL I containing internal loops (IL) 1 and 2 followed by a predicted Sm protein binding site (Figure 4.3B). So far, this strategy has not identified obvious U5 snRNA-like sequences in *S. vortens*.

Based on the two *Spironucleus* U2 candidates, we anticipated that U1 and U5 snRNAs may also be similar in either species. Consequently, we performed reciprocal BLASTN searches using *S. vortens* U1 and *S. salmonicida* U5 candidate sequences as queries against *S. salmonicida* and *S. vortens* genomic sequences. In both cases, we were unable to identify orthologs specifying either RNA in the corresponding species.

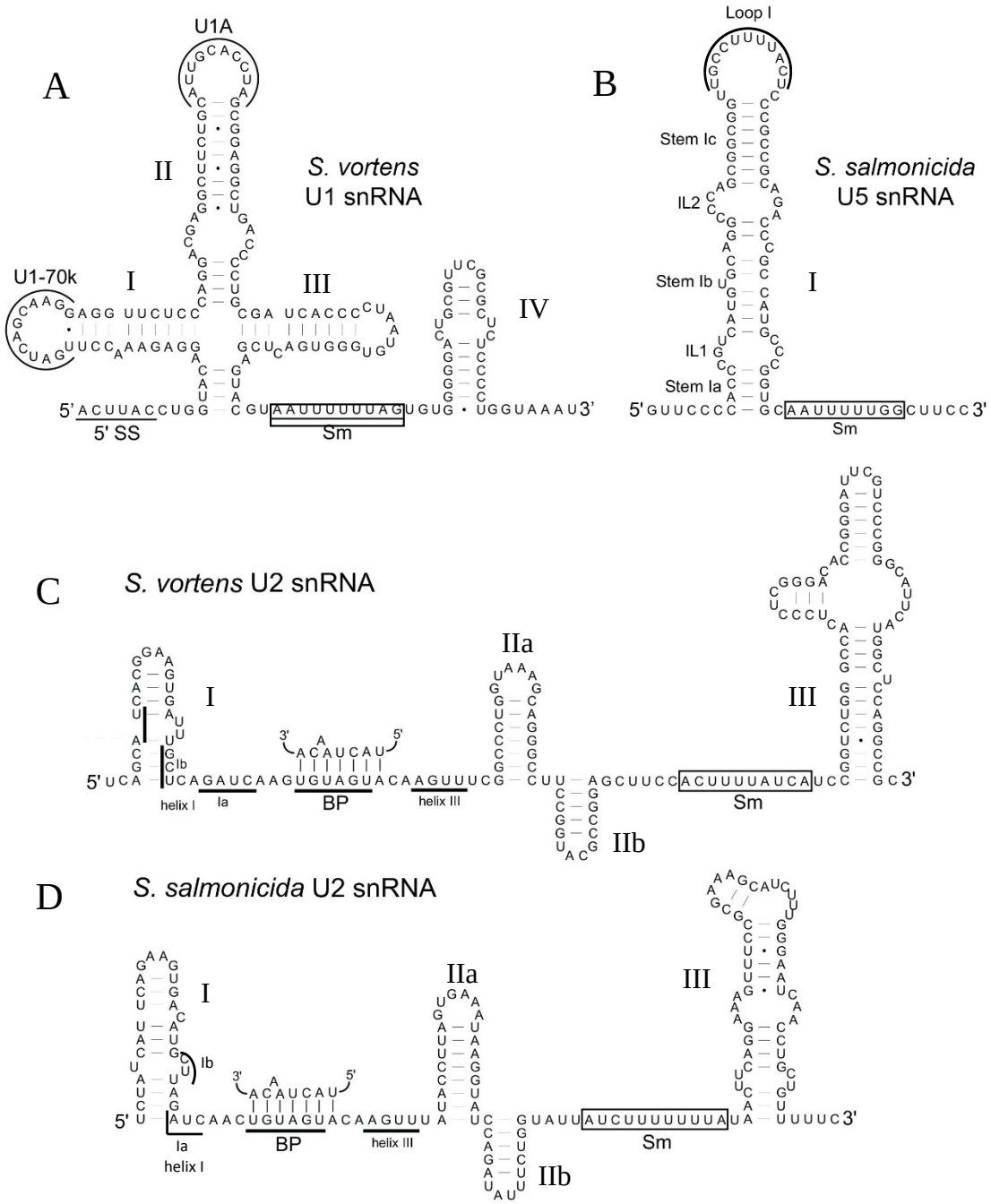


Figure 4.3. Spliceosomal snRNAs from *S. vortens* and *S. salmonicida*.
(**A**-**D**) Predicted secondary structures (MFOLD) for *Spironucleus* snRNA sequences are shown with conserved sequence and structural elements indicated. Sm = Sm protein binding site. (**A**) U1 snRNA, 5′ SS = 5′ splice site interacting sequence. Binding sites for U1-70k and U1A proteins are indicated. (**C,D**) U2 snRNAs in *Spironucleus*. BP = branch point interacting sequence. Helix I and III are regions of U2 snRNA predicted to form intermolecular base pairs with U6 snRNA. Accession numbers of genomic contigs containing snRNA sequences are: (**A**) *S. vortens* U1 (NCBI trace archive - ti|2141736653: nucleotide positions 512-345), (**B**) *S. salmonicida* U5 (GenBank AUWU01000115:5304-5391), (**C**) *S. vortens* U2 (ti|2141663608:84-246) and (**D**) *S. salmonicida* U2 (AUWU01000434:68649-68502).

4.3.5 The phylogenetic distribution of the *Rps4* and *Rps24* introns indicates they are ancient

The previous examination of the *G. lamblia Rpl7a* intron revealed intron conservation at the identical position within *Rpl7a* orthologs from representative organisms of two of the five currently accepted eukaryotic supergroups (Russell *et al.* 2005) (Figure 4.5). This phylogenetic conservation of the *Rpl7a* intron indicated an early 'appearance' of this intron in eukaryotic evolution. Polymerase chain reaction (PCR) analysis of *Spironucleus barkhanus* genomic DNA indicated a lack of the *Rpl7a* intron in this species (Russell *et al.* 2005). We have now discovered that *Spironucleus vortens* contains the *Rpl7a* intron (Figure 4.1A) indicating recent loss of the *Rpl7a* intron in some *Spironucleus* species.

We next analyzed the conservation patterns of the other *S. vortens* RP gene introns. We examined more than 80 eukaryotes representing all five proposed eukaryotic supergroups (Burki 2014) and identified spliceosomal introns in identical position and phase in the *Rps4* and *Rps24* genes in several other distantly-related organisms (Figure 4.4B and C). For both genes, at least one representative organism within each of the five supergroups contains an intron at the same position as *S. vortens* with some organisms conserving both introns; the *Rps24* intron displays a somewhat wider distribution (Figure 4.5 and Table A.3.3 and A.3.4). The introns are nearly always inserted at the same relative coding position and phase within the ORF; however, we also found some evidence of 'intron sliding' in which organisms had an *Rps4* or *Rsp24* intron in an adjacent codon to the conserved intron insertion position (data not shown). Collectively, our analyses reveal

that the *Rps4* and *Rps24* introns may be even more widespread in eukaryotes than the *Rpl7a* intron (Figure 4.5).

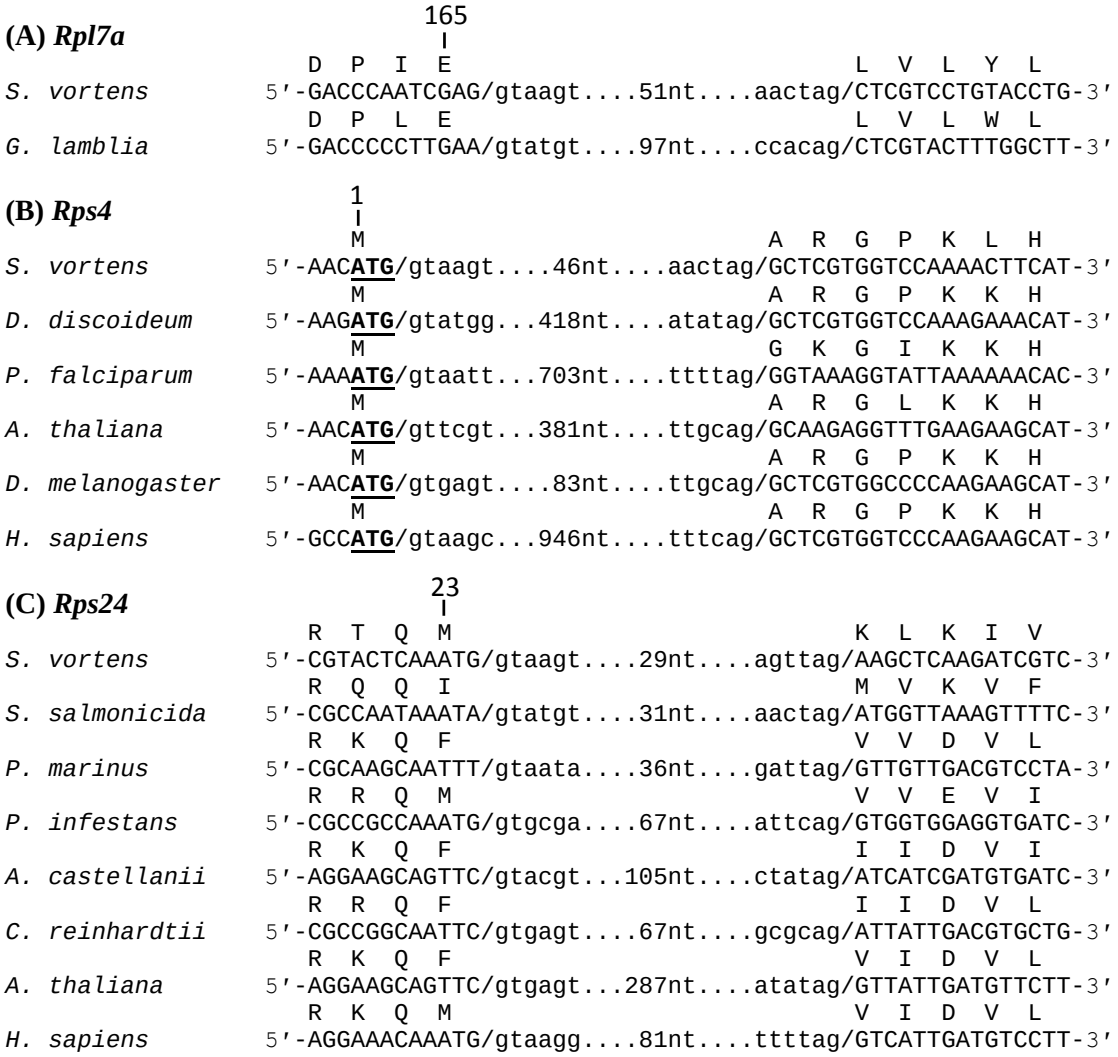


Figure 4.4. Conservation of *Rpl7a, Rps4* and *Rps24* intron insertion sites.
Gene sequences from representative eukaryotes containing *Rpl7a* (**A**), *Rps4* (**B**) and *Rps24* (**C**) spliceosomal introns are aligned with slashes (/) representing intron-exon boundaries, intronic sequences in lower case and exonic sequences in uppercase. The number of nucleotides between splice site sequences is indicated. Translated amino acid sequences are shown above the first nucleotide of each codon and the start 'ATG' codons for the *Rps4* coding sequences are underlined. Amino acid positions for each protein are indicated based on the *H. sapiens* orthologs. NCBI accession numbers for (**A**) *Rpl7a* - *S. vortens* [NCBI Trace Archive:ti|2141515448], *G. lamblia* [GenBank:NW_002477099], (**B**) *Rps4* - *S. vortens* [ti|2141550682], *S. salmonicida* [gb|AUWU01000316] *D. discoideum* [NC_007088], *P. falciparum* [NC_004315], *A. thaliana* [NC_003071], *D. melanogaster* [NT_037436] and *H. sapiens* [NC_000023] and (**C**) *Rps24 – S. vortens* [ti|2141541737], *P. marinus* [NW_003201404], *P. infestans* [NW_003303749], *A. castellanii* [NW_004457654], *C. reinhardtii* [NW_001843791], *A. thaliana* [NC_003074] and *H. sapiens* [NC_000010].

We also found spliceosomal introns in the *S. vortens Rpl30* and *Rps12* genes in the same relative positions as those in humans. These are found in less well-conserved regions of RP amino acid sequence and therefore it is more difficult to ascertain whether these represent ancient intron insertion events or more recent independent intron acquisitions in nearby sites.

An alternative (but less parsimonious) explanation for the observed evolutionary distribution of the *Rps4* and *Rps24* introns was the occurrence of numerous independent (and widespread) intron gain events at proto-splice sites in these genes. Nucleotide sequence in the flanking exon portions adjacent to either the *Rps4* or *Rps24* introns shows conservation amongst distantly-related eukaryotes, consistent with conserved RP amino acid sequence encoded by these regions. For the *Rps24* intron, these sequences do not conform to the proto-splice site consensus (A/C)AG/G (Sverdlov *et al.* 2005). However, the exonic sequences flanking the *Rps4* intron are a better match (typically 3 out of 4 nt). Because exonic sequence encodes the invariant 'ATG' start codon (proto-splice site nucleotides underlined) and conserved alanine ('GCN') or glycine ('GGN') resides, we cannot refute the possibility that the widespread distribution of the *Rps4* intron is the result of multiple independent intron gains. Thus, we conclude that the observed distribution of the *Rpl7a* and *Rps24* introns are not likely due to independent intron gains at proto-splice sites and the phylogenetic distribution of *Rps24* introns may be explained by single ancient intron gain events in the last common ancestor of the examined taxa.

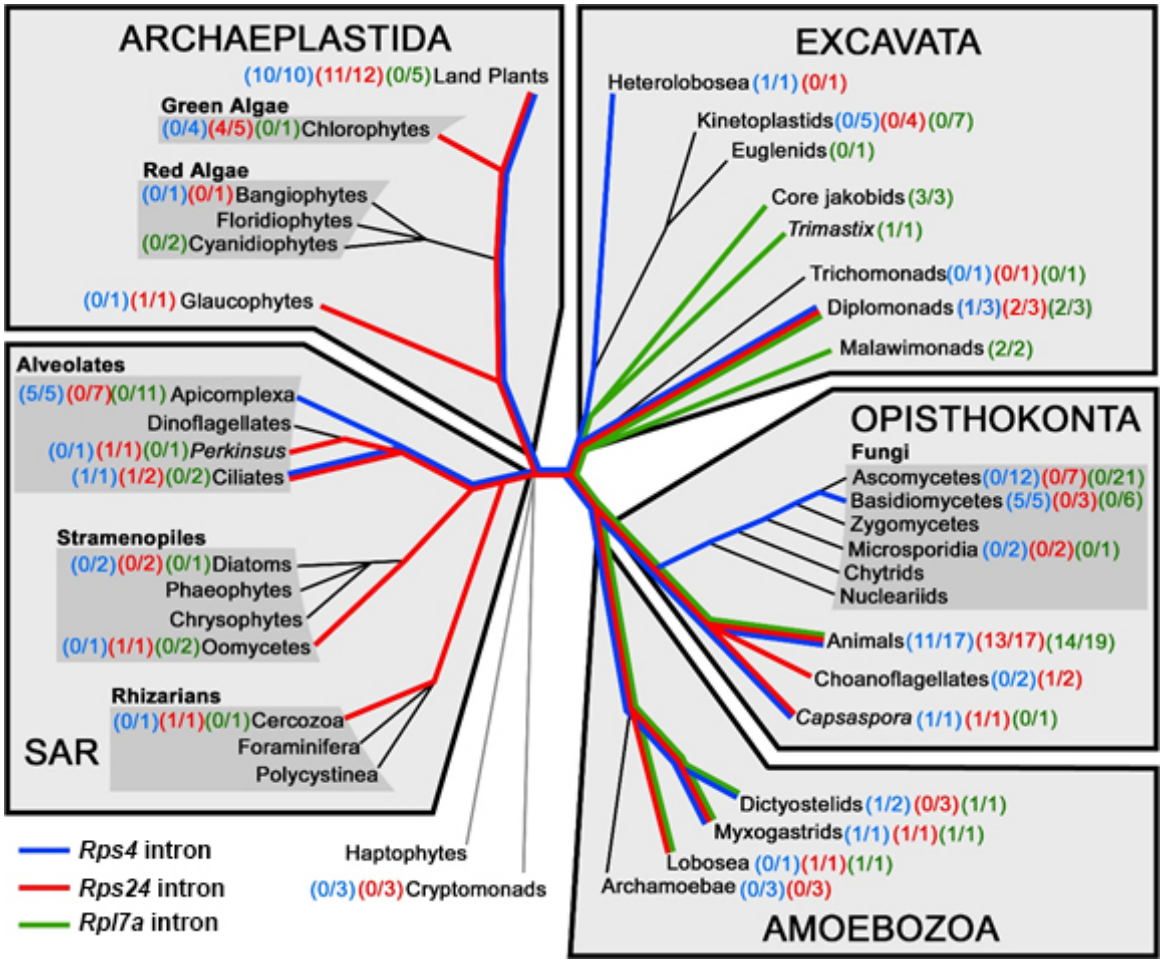


Figure 4.5. Phylogenetic distribution of RP gene introns in eukaryotes.
Representative eukaryotes from each eukaryotic supergroup were examined for intron insertion at the same conserved position within *Rps4* (blue), *Rps24* (red) and *Rpl7a* (green) (Russell *et al.* 2005) genes and the distribution for each intron was mapped onto a recent eukaryotic tree by Burki (2014). The number of species containing an intron (numerator) and the number sampled (denominator) are indicated for each eukaryotic group (See Tables A.3.3 and A.3.4 for organism names). Coloured lines indicate extant eukaryotic groups which contain each intron to the predicted last common ancestor to contain each respective RP gene intron.

4.4 Discussion

4.4.1 Intron conservation in diplomonads and parabasalids

Identification of the first spliceosomal introns in *G. lamblia* and the parabasalid *T. vaginalis* revealed an unexpected level of intron structure and sequence conservation, with near-identical 5′ SS and fused BP + 3′ SS consensus sequences between the two species

(Russell *et al.* 2005, Vanacova *et al.* 2005) (Figure 4.1C). Indeed, the *G. lamblia* ferredoxin intron was readily spliced from an expressed reporter gene construct in *T. vaginalis* (following 5′ SS 'CT' dinucleotide substitution to 'GT') (Vanacova *et al.* 2005) emphasizing the similarities in intron structure and splicing mechanism in these two organisms. While the *Spironucleus* introns display the fused BP + 3′ SS shared property, they also show some differences in splice site sequence preference and intron element spacing relative to *G. lamblia* or *T. vaginalis* introns, such as variation in 5′ and 3′ SS nucleotide sequences and an additional nucleotide insertion between the BP and 3′ SS sequences in the *Spironucleus* introns (Figure 4.1C).

The structural properties of the identified *G. lamblia* introns suggest that intron elements may have particular importance in the spatial positioning of splice sites and the branch point 'A' during the splicing pathway, relative to other eukaryotes. These intron properties include an invariant distance between the branch point and 3′ SS, and extensive base-pairing potential present not only between *trans*-spliced intron halves but also in the larger *cis*-spliced introns, such as the *Rpl7a* intron. We now provide evidence for intron structural potential in the longer *cis*-introns from *Spironucleus* and *T. vaginalis* which, similar to *G. lamblia*, may reduce the single-stranded distance between intron elements to lengths comparable to the shorter and uniformly-sized spliceosomal introns in these organisms. The conservation of extensive intron base-pairing potential in a parabasalid (*T. vaginalis*) and diplomonads (*G. lamblia* and *Spironucleus spp.*) further indicates a shared requirement to maintain a specific spatial positioning of intron elements for efficient splicing and suggests that this property of spliceosomal introns may be much more phylogenetically wide spread than previously thought.

4.4.2 *Spironucleus* snRNAs reveal spliceosome structure divergence in diplomonads

Characterization of the U1, U2, U4 and U6 spliceosomal snRNAs from *G. lamblia* revealed that they are evolutionarily divergent and possess secondary structures and sequence motifs characteristic of both major (U2-dependent) and minor (U12-dependent) spliceosomal snRNAs (see Chapter 3). In contrast, our bioinformatic searches for snRNAs in *Spironucleus spp.* yielded a more conventional U1 snRNA candidate in *S. vortens*, containing all expected secondary structures and sequences motifs (*G. lamblia* U1 snRNA lacks a putative U1-70k binding site and SL IV). The *S. vortens* and *S. salmonicida* U2 snRNA candidates also appear to be more major/U2 snRNA-like, based on primary sequence comparisons with other representative U2 and U12 snRNAs (Figure A.3.5). However, the *Spironucleus* U2 snRNAs appear to lack SL IV (Figure 4.3C and D) and instead they are predicted to form an extended long SL III – a characteristic of minor U12 snRNAs (Patel and Steitz 2003, Russell *et al.* 2006). Similarly, the *G. lamblia* U2 snRNA (Hudson *et al.* 2012) 3′ half may also fold into a similar conformation with a single long SL III (Figure 3.3). However, we find that neither the *G. lamblia* U2 snRNA nor the *Spironucleus* U2 candidates contain the conserved SL III loop sequence ‘CUACUUU’ that is bound by the U12 snRNP-specific 65kDa protein (Benecke *et al.* 2005). Therefore, we conclude that the *Spironucleus* and *G. lamblia* U2 snRNAs are more likely *bona fide* U2-dependent/major spliceosomal components.

Taken together, the identification of both canonical and less-canonical snRNAs with different divergent patterns from several diplomonads indicates that at least some of the unusual features of *G. lamblia* and *Spironucleus* snRNAs are likely to be recent

alterations, occurring after branching of the *Giardia* and *Spironucleus* lineages. However, it remains to be determined whether the peculiar hybrid U2/U12-dependent spliceosomal nature of the *G. lamblia* snRNAs represent fusion events involving ancestral major and minor spliceosomal snRNAs or rather convergent evolution of U2-dependent spliceosomal snRNAs to adopt U12-dependent snRNA-like features. The only other spliceosomal snRNA candidate we have been able to identify so far is the U5 snRNA in *S. salmonicida* (Figure 4.3B). U5 snRNA is a shared component of both U2-dependent and U12-dependent spliceosomes and thus, does not help explain the major/minor duality of spliceosomal snRNAs in diplomonads. Finally, some snRNAs (particularly U4 and U6 snRNAs) escaped detection in our searches. Although we acknowledge other possibilities to explain this, such as incomplete coverage during genomic sequence determination or inherent search strategy biases, it is reasonable to speculate that these missing snRNAs may be sufficiently divergent to escape 'easy' detection. Identification of the remaining snRNAs in *S. vortens* and *S. salmonicida* (and other diplomonads) should provide additional insight into the evolutionary history of their respective spliceosomes.

4.4.3 A high frequency of ancient RP gene spliceosomal introns in diplomonads

Ancient spliceosomal introns are often maintained in intron poor eukaryotes and particularly within RP genes. Consistent with this, we find that several (2 out of 6) of the EST-confirmed *S. vortens* spliceosomal introns are ancient RP gene introns, with the *Rps4* and *Rps24* introns representing some of the most evolutionarily-conserved introns discovered to date. However, no single spliceosomal intron has been maintained in all of the diplomonad species studied thus far, indicating that spliceosomal intron loss is still ongoing and may eventually reach completion in members of this group.

4.5 Conclusions

Intron-poor eukaryotes are marked by constrained and extended intron splicing signals and reduced splicing machinery. Here we find a remarkable level of sequence conservation of spliceosomal introns in *Spironucleus* species and evidence for additional structural constraints to position intron elements for efficient splicing – a feature apparently conserved in both diplomonads and parabasalid introns. The requirement for such positioning of intron elements is intriguing and points to a more simplified splicing mechanism(s) in these organisms. This coincides with drastic changes in typically conserved spliceosomal structures including loss (or modification) of snRNA domains as observed in *G. lamblia* and *Spironucleus spp*. Such changes in snRNA structure may be concurrent with the loss of auxiliary spliceosomal proteins involved in splicing regulation (and alternative splicing). Indeed, searches for spliceosomal proteins in *G. lamblia* revealed divergent homologs with several core spliceosomal components seemingly absent (Nixon *et al.* 2002, Collins and Penny 2005). It will be interesting to determine whether snRNAs from other diplomonads and other divergent eukaryotes share these unusual features.

Finally, intron base pairing is proposed to mediate association of the known *G. lamblia trans*-introns and thus may be a required first step towards intron fragmentation and gene fission (Roy *et al.* 2012). The conservation of intron base pairing in *cis*-introns in diplomonads and a parabasalid and the large proportion of *trans*-spliced introns in *G. lamblia*, suggests that additional *trans*-spliced introns may await discovery in members of these groups.

Chapter 5: Conclusions and future perspectives

In this work, I examined spliceosomal introns and spliceosomal snRNAs in *G. lamblia* and two other diplomonads, *S. vortens* and *S. salmonicida* which revealed a number of surprising phenomena in these species: i) a high occurrence of pre-mRNA *trans*-splicing; ii) a prevalent ncRNA 3′ end processing motif in *G. lamblia*; iii) both conventional and highly evolutionarily-divergent snRNAs in *Spironucleus spp.* and *G. lamblia*; and iv) a large proportion of ancient spliceosomal introns in intron-poor diplomonads.

In Chapter 2, four cases of pre-mRNA *trans*-splicing were identified in *G. lamblia*. A closer examination revealed corresponding *trans*-intron 5′ and 3′ halves contain complementary sequences which may position intron elements correctly to permit exon *trans*-splicing by the *G. lamblia* spliceosome. However, specific requirements for intron base-pairing stability and factors required for *trans*-splicing are currently unknown. *G. lamblia* ncRNA 3′ end processing motifs were also identified downstream of base-pairing regions of *trans*-spliced intron 5′ halves, suggesting that the motif may play an important role in *trans*-splicing. My preliminary experiments indicate that motif cleavage occurs at these sites, but cleavage is not a strict requirement for the first step of splicing. Whether motif cleavage is required for the second step of splicing (or product release) is currently uncertain. One possibility is that motif cleavage has a (non-essential) beneficial function in facilitating *trans*-intron base-pairing by generating discrete RNA 3′ ends for intron 5′ halves which lack additional downstream sequences. Further elucidating the *G. lamblia* *trans*-splicing pathway will require development of *in vitro* and *in vivo* splicing assays. Several expression vectors have now been developed for use in *G. lamblia* (Jerlstrom-Hultqvist et al. 2012) and *Spironucleus spp.* (Dawson et al. 2008, Jerlstrom-Hultqvist et al.

2012) which will permit establishment of *in vivo trans*-splicing assays capable of determining intron substrate requirements for *trans*-splicing in these species. It will be interesting to learn whether *trans*-splicing plays a role in the expression of virulence factors in *G. lamblia*. Relevantly, Hsp90 (whose pre-mRNAs are *trans*-spliced) is enriched in *G. lamblia* exosomes that may contribute to disease during infection (S. G. Svärd, personal communication). Thus, elucidating details of the *trans*-splicing pathway in *G. lamblia* may also identify new therapeutic strategies that target this unusual process.

The identified *G. lamblia* 3′ end RNA motif is predicted to be involved in the processing of ~40 putative and confirmed ncRNAs including numerous box C/D and box H/ACA snoRNAs, RNase MRP RNA, spliceosomal snRNAs, telomerase RNA and several ncRNAs of unknown function (Chapter 3). While alternative pathways to generate ncRNA 3′ ends may (and likely do) exist in *G. lamblia*, it is evident that the 3′ RNA motif is an integral RNA processing signal in this organism. As of yet, nothing is known about the cellular components involved in motif recognition and cleavage. I have made preliminary attempts to purify motif-binding factors from *G. lamblia* cellular extracts; however, these experiments have not been successful in identifying motif-specific binding proteins (data not shown). Determining optimal conditions for the purification of motif binding factors will be an important next step for elucidating the *G. lamblia* motif cleavage pathway. Also, despite the prevalence of the 3′ end motif in *G. lamblia*, I was not able to identify obvious motif sequences downstream of genes encoding the predicted *S. vortens* and *S. salmonicida* snRNAs. With only a handful of putative ncRNAs identified in either *Spironucleus* species it is difficult to ascertain whether an equivalent 3′ end ncRNA processing motif is present in members of this diplomonad genus. Deep-sequencing of *Spironucleus* ncRNA cDNA

libraries and examination of corresponding ncRNA gene sequences will likely be required to identify alternate 3′ end processing signals in these organisms. Nonetheless, further elucidation of the *G. lamblia* ncRNA motif cleavage pathway will be an important area of further study which may identify novel drug targets for treatment of *Giardia* infection.

My use of the *G. lamblia* motif as a bioinformatic tool to identify additional ncRNAs revealed several novel evolutionarily-divergent snRNAs, two box H/ACA snoRNAs, and a telomerase RNA candidate. However, the U5 snRNA is still unconfirmed in *G. lamblia*. I have also attempted to purify the U5 snRNP by immunoprecipitation from *G. lamblia* cellular extracts using rabbit polyclonal antibodies raised against peptide fragments of *G. lamblia* Prp8p, followed by RT-PCR amplification of co-purified RNAs (data not shown). However, these attempts have not yet been successful in identifying an authentic U5 snRNA. Purification of tagged *G. lamblia* U5 snRNP proteins (such as Prp8p) using *in vivo G. lamblia* expression vectors may prove more successful in verifying the identity of this missing snRNA. Nevertheless, identification of 4 out of the 5 spliceosomal snRNAs provides an excellent starting point for elucidating additional components of the *G. lamblia* spliceosome. Biochemical purification and analysis of *G. lamblia* spliceosomes will help in determining minimal core protein components and snRNA structures required for spliceosome activity.

Finally, my analysis of *S. vortens Rps4* and *Rps24* intron conservation revealed that these introns are some of the most highly-conserved spliceosomal introns in eukaryotes (Chapter 4). It is interesting that certain spliceosomal introns are particularly resistant to loss and this may suggest these introns provide some beneficial function to the host. A recent study in *S. cerevisiae* revealed that nearly all ribosomal protein (RP) gene introns

have important roles in regulating the abundance of RP gene paralogs and intron deletion from RP genes has a significant effect on cell growth under stress conditions (Parenteau *et al.* 2011). Thus, it is plausible that ancient RP gene introns (such as the *S. vortens Rps4* and *Rps24* introns) have resisted loss in evolution due to providing important functions in fine-tuning RP gene expression. To test this, it would be interesting to determine if the remaining RP gene introns in *G. lamblia* and *Spironucleus spp.* have a significant effect on transcription/translation of their host gene. One might envision using *in vivo* expression vectors to express intron-containing or intron-less versions of a reporter gene (such as green fluorescent protein) to determine if the presence of these particular introns affect transcription or translation of the gene. The finding of common gene regulatory functions for certain ancient spliceosomal introns in different eukaryotes would indicate that these functions arose very early in eukaryotic evolution and perhaps predated LECA.

References

Adl, S. M., Simpson, A. G., Lane, C. E., Lukes, J., Bass, D., Bowser, S. S., Brown, M. W., Burki, F., Dunthorn, M., Hampl, V., *et al.* (2012). The revised classification of eukaryotes. *J Eukaryot Microbiol* **59**(5): 429-493.

Alioto, T. S. (2007). U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res* **35**(Database issue): D110-115.

Ambrosio, D. L., Silva, M. T. and Cicarelli, R. M. (2007). Cloning and molecular characterization of *Trypanosoma cruzi* U2, U4, U5, and U6 small nuclear RNAs. *Mem Inst Oswaldo Cruz* **102**(1): 97-105.

Andersson, J. O., Sjogren, A. M., Davis, L. A., Embley, T. M. and Roger, A. J. (2003). Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. *Curr Biol* **13**(2): 94-104.

Andersson, J. O., Sjogren, A. M., Horner, D. S., Murphy, C. A., Dyal, P. L., Svard, S. G., Logsdon, J. M., Jr., Ragan, M. A., Hirt, R. P. and Roger, A. J. (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.

Archibald, J. M., O'Kelly, C. J. and Doolittle, W. F. (2002). The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution. *Mol Biol Evol* **19**(4): 422-431.

Asai, D. J. and Koonce, M. P. (2001). The dynein heavy chain: structure, mechanics and evolution. *Trends Cell Biol* **11**(5): 196-202.

Bartschat, S. and Samuelsson, T. (2010). U12 type introns were lost at multiple occasions during evolution. *BMC Genomics* **11**: 106.

Benecke, H., Luhrmann, R. and Will, C. L. (2005). The U11/U12 snRNP 65K protein acts as a molecular bridge, binding the U12 snRNA and U11-59K protein. *EMBO J* **24**(17): 3057-3069.

Berget, S. M., Moore, C. and Sharp, P. A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A* **74**(8): 3171-3175.

Biswas, S., Buhrman, G., Gagnon, K., Mattos, C., Brown, B. A., 2nd and Maxwell, E. S. (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**(14): 2907-2918.

Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72**: 291-336.

Blackburn, E. H. and Collins, K. (2011). Telomerase: an RNP enzyme synthesizes DNA. *Cold Spring Harb Perspect Biol* **3**(5).

Blocker, F. J., Mohr, G., Conlan, L. H., Qi, L., Belfort, M. and Lambowitz, A. M. (2005). Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. *RNA* **11**(1): 14-28.

Bon, E., Casaregola, S., Blandin, G., Llorente, B., Neuveglise, C., Munsterkotter, M., Guldener, U., Mewes, H. W., Van Helden, J., Dujon, B., *et al.* (2003). Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Res* **31**(4): 1121-1135.

Bonen, L. (1993). Trans-splicing of pre-mRNA in plants, animals, and protists. *FASEB J* **7**(1): 40-46.

Brow, D. A. and Guthrie, C. (1988). Spliceosomal RNA U6 is remarkably conserved from yeast to mammals. *Nature* **334**(6179): 213-218.

Brow, D. A. and Vidaver, R. M. (1995). An element in human U6 RNA destabilizes the U4/U6 spliceosomal RNA complex. *RNA* **1**(2): 122-131.

Burge, C. B., Padgett, R. A. and Sharp, P. A. (1998). Evolutionary fates and origins of U12-type introns. *Mol Cell* **2**(6): 773-785.

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R., Gardner, P. P. and Bateman, A. (2013). Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* **41**(Database issue): D226-232.

Burger, G., Yan, Y., Javadi, P. and Lang, B. F. (2009). Group I-intron trans-splicing and mRNA editing in the mitochondria of placozoan animals. *Trends Genet* **25**(9): 381-386.

Burke, J. E., Sashital, D. G., Zuo, X., Wang, Y. X. and Butcher, S. E. (2012). Structure of the yeast U2/U6 snRNA complex. *RNA* **18**(4): 673-683.

Burki, F. (2014). The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harb Perspect Biol* **6**(5): a016147.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. and Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**(18): 1915-1927.

Canaday, J., Tessier, L. H., Imbault, P. and Paulus, F. (2001). Analysis of *Euglena gracilis* alpha-, beta- and gamma-tubulin genes: introns and pre-mRNA maturation. *Mol Genet Genomics* **265**(1): 153-160.

Cavalier-Smith, T. (1991). Intron phylogeny: a new hypothesis. *Trends Genet* **7**(5): 145-148.

Chanfreau, G., Rotondo, G., Legrain, P. and Jacquier, A. (1998). Processing of a dicistronic small nucleolar RNA precursor by the RNA endonuclease Rnt1. *EMBO J* **17**(13): 3726-3737.

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

Chen, X. S., Rozhdestvensky, T. S., Collins, L. J., Schmitz, J. and 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**(14): 4619-4628.

Chen, X. S., White, W. T., Collins, L. J. and Penny, D. (2008). Computational identification of four spliceosomal snRNAs from the deep-branching eukaryote *Giardia intestinalis*. *PLoS One* **3**(8): e3106.

Chow, C. S., Lamichhane, T. N. and Mahto, S. K. (2007). Expanding the nucleotide repertoire of the ribosome with post-transcriptional modifications. *ACS Chem Biol* **2**(9): 610-619.

Chow, L. T., Gelinas, R. E., Broker, T. R. and Roberts, R. J. (1977). An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**(1): 1-8.

Collins, L. and Penny, D. (2005). Complex spliceosomal organization ancestral to extant eukaryotes. *Mol Biol Evol* **22**(4): 1053-1066.

Crick, F. (1970). Central dogma of molecular biology. *Nature* **227**(5258): 561-563.

Crooks, G. E., Hon, G., Chandonia, J. M. and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res* **14**(6): 1188-1190.

Csuros, M., Rogozin, I. B. and Koonin, E. V. (2008). Extremely intron-rich genes in the alveolate ancestors inferred with a flexible maximum-likelihood approach. *Mol Biol Evol* **25**(5): 903-911.

Csuros, M., Rogozin, I. B. and Koonin, E. V. (2011). A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput Biol* **7**(9): e1002150.

Davids, B. J. and Gillin, F. D. (2011). Methods for Giardia Culture, Cryopreservation, Encystation, and Excystation In Vitro. *Giardia: A Model Organism*: 381-394.

Dawson, S. C., Pham, J. K., House, S. A., Slawson, E. E., Cronembold, D. and Cande, W. Z. (2008). Stable transformation of an episomal protein-tagging shuttle vector in the piscine diplomonad *Spironucleus vortens*. *BMC Microbiol* **8**: 71.

de Souza, S. J., Long, M., Schoenbach, L., Roy, S. W. and Gilbert, W. (1996). Intron positions correlate with module boundaries in ancient proteins. *Proc Natl Acad Sci U S A* **93**(25): 14632-14636.

Dibb, N. J. (1991). Proto-splice site model of intron origin. *J Theor Biol* **151**(3): 405-416.

Dibb, N. J. and Newman, A. J. (1989). Evidence that introns arose at proto-splice sites. *EMBO J* **8**(7): 2015-2021.

Dietrich, R. C., Fuller, J. D. and Padgett, R. A. (2005). A mutational analysis of U12-dependent splice site dinucleotides. *RNA* **11**(9): 1430-1440.

Dietrich, R. C., Incorvaia, R. and Padgett, R. A. (1997). Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. *Mol Cell* **1**(1): 151-160.

Dietrich, R. C., Padgett, R. A. and Shukla, G. C. (2009). The conserved 3' end domain of U6atac snRNA can direct U6 snRNA to the minor spliceosome. *RNA* **15**(6): 1198-1207.

Dietrich, R. C., Peris, M. J., Seyboldt, A. S. and Padgett, R. A. (2001). Role of the 3' splice site in U12-dependent intron splicing. *Mol Cell Biol* **21**(6): 1942-1952.

Doolittle, W. F. (1987). The Origin and Function of Intervening Sequences in DNA - a Review. *American Naturalist* **130**(6): 915-928.

Dorn, R., Reuter, G. and Loewendorf, A. (2001). Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in *Drosophila*. *Proc Natl Acad Sci U S A* **98**(17): 9724-9729.

Dsouza, M., Larsen, N. and Overbeek, R. (1997). Searching for patterns in genomic data. *Trends Genet* **13**(12): 497-498.

Elmendorf, H. G., Singer, S. M., Pierce, J., Cowan, J. and Nash, T. E. (2001). Initiator and upstream elements in the alpha2-tubulin promoter of *Giardia lamblia*. *Mol Biochem Parasitol* **113**(1): 157-169.

Fabrizio, P., Dannenberg, J., Dube, P., Kastner, B., Stark, H., Urlaub, H. and Luhrmann, R. (2009). The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome. *Mol Cell* **36**(4): 593-608.

Fast, N. M., Roger, A. J., Richardson, C. A. and Doolittle, W. F. (1998). U2 and U6 snRNA genes in the microsporidian *Nosema locustae*: evidence for a functional spliceosome. *Nucleic Acids Res* **26**(13): 3202-3207.

Fedorov, A., Merican, A. F. and Gilbert, W. (2002). Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc Natl Acad Sci U S A* **99**(25): 16128-16133.

Fedorov, A., Suboch, G., Bujakov, M. and Fedorova, L. (1992). Analysis of nonuniformity in intron phase distribution. *Nucleic Acids Res* **20**(10): 2553-2557.

Fica, S. M., Tuttle, N., Novak, T., Li, N. S., Lu, J., Koodathingal, P., Dai, Q., Staley, J. P. and Piccirilli, J. A. (2013). RNA catalyses nuclear pre-mRNA splicing. *Nature* **503**(7475): 229-234.

Fischer, S. E., Butler, M. D., Pan, Q. and Ruvkun, G. (2008). Trans-splicing in C. elegans generates the negative RNAi regulator ERI-6/7. *Nature* **455**(7212): 491-496.

Franzen, O., Jerlstrom-Hultqvist, J., Castro, E., Sherwood, E., Ankarklev, J., Reiner, D. S., Palm, D., Andersson, J. O., Andersson, B. and Svard, S. G. (2009). Draft genome sequencing of *Giardia intestinalis* assemblage B isolate GS: is human giardiasis caused by two different species? *PLoS Pathog* **5**(8): e1000560.

Franzen, O., Jerlstrom-Hultqvist, J., Einarsson, E., Ankarklev, J., Ferella, M., Andersson, B. and Svard, S. G. (2013). Transcriptome profiling of *Giardia intestinalis* using strand-specific RNA-seq. *PLoS Comput Biol* **9**(3): e1003000.

Friesen, M. J. and Dreyfuss, G. (2000). Specific sequences of the Sm and Sm-like (Lsm) proteins mediate their interaction with the spinal muscular atrophy disease gene product (SMN). *Journal of Biological Chemistry* **275**(34): 26370-26375.

Frilander, M. J. and Steitz, J. A. (1999). Initial recognition of U12-dependent introns requires both U11/5 ' splice-side and U12/branchpoint interactions. *Genes & Development* **13**(7): 851-863.

Galej, W. P., Oubridge, C., Newman, A. J. and Nagai, K. (2013). Crystal structure of Prp8 reveals active site cavity of the spliceosome. *Nature* **493**(7434): 638-643.

Gao, K. P., Masuda, A., Matsuura, T. and Ohno, K. (2008). Human branch point consensus sequence is yUnAy. *Nucleic Acids Research* **36**(7): 2257-2267.

Gelfman, S., Burstein, D., Penn, O., Savchenko, A., Amit, M., Schwartz, S., Pupko, T. and Ast, G. (2012). Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. *Genome Res* **22**(1): 35-50.

Gilbert, W. (1978). Why genes in pieces? *Nature* **271**(5645): 501.

Gilbert, W. (1987). The Exon Theory of Genes. *Cold Spring Harbor Symposia on Quantitative Biology* **52**: 901-905.

Gilson, P. R., Su, V., Slamovits, C. H., Reith, M. E., Keeling, P. J. and McFadden, G. I. (2006). Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. *Proc Natl Acad Sci U S A* **103**(25): 9566-9571.

Goldschmidt-clermont, M., Choquet, Y., Girardbascou, J., Michel, F., Schirmerrahire, M. and Rochaix, J. D. (1991). A small chloroplast RNA may be required for *trans*-splicing in *Chlamydomonas reinhardtii*. *Cell* **65**(1): 135-143.

Guthrie, C. and Patterson, B. (1988). Spliceosomal snRNAs. *Annu Rev Genet* **22**: 387-419.

Hall, S. L. and Padgett, R. A. (1994). Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *J Mol Biol* **239**(3): 357-365.

Hashimoto, T., Sanchez, L. B., Shirakura, T., Muller, M. and Hasegawa, M. (1998). Secondary absence of mitochondria in *Giardia lamblia* and *Trichomonas vaginalis* revealed by valyl-tRNA synthetase phylogeny. *Proc Natl Acad Sci U S A* **95**(12): 6860-6865.

Haugen, P., Simon, D. M. and Bhattacharya, D. (2005). The natural history of group I introns. *Trends Genet* **21**(2): 111-119.

Henras, A. K., Soudet, J., Gerus, M., Lebaron, S., Caizergues-Ferrer, M., Mougin, A. and Henry, Y. (2008). The post-transcriptional steps of eukaryotic ribosome biogenesis. *Cell Mol Life Sci* **65**(15): 2334-2359.

Hetzer, M., Wurzer, G., Schweyen, R. J. and Mueller, M. W. (1997). Trans-activation of group II intron splicing by nuclear U5 snRNA. *Nature* **386**(6623): 417-420.

Hirt, R. P., Logsdon, J. M., Jr., Healy, B., Dorey, M. W., Doolittle, W. F. and Embley, T. M. (1999). Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proc Natl Acad Sci U S A* **96**(2): 580-585.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res* **31**(13): 3429-3431.

Hook, P. and Vallee, R. B. (2006). The dynein family at a glance. *J Cell Sci* **119**(Pt 21): 4369-4371.

Hudson, A. J., Moore, A. N., Elniski, D., Joseph, J., Yee, J. and Russell, A. G. (2012). Evolutionarily divergent spliceosomal snRNAs and a conserved non-coding RNA processing motif in *Giardia lamblia*. *Nucleic Acids Res* **40**(21): 10995-11008.

Huppler, A., Nikstad, L. J., Allmann, A. M., Brow, D. A. and Butcher, S. E. (2002). Metal binding and base ionization in the U6 RNA intramolecular stem-loop structure. *Nat Struct Biol* **9**(6): 431-435.

Irimia, M., Penny, D. and Roy, S. W. (2007). Coevolution of genomic intron number and splice sites. *Trends Genet* **23**(7): 321-325.

Irimia, M. and Roy, S. W. (2008). Evolutionary convergence on highly-conserved 3' intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. *PLoS Genet* **4**(8): e1000148.

Jackson, I. J. (1991). A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Res* **19**(14): 3795-3798.

Jarrell, K. A., Dietrich, R. C. and Perlman, P. S. (1988). Group II intron domain 5 facilitates a trans-splicing reaction. *Mol Cell Biol* **8**(6): 2361-2366.

Jarrell, K. A., Peebles, C. L., Dietrich, R. C., Romiti, S. L. and Perlman, P. S. (1988). Group II intron self-splicing. Alternative reaction conditions yield novel products. *J Biol Chem* **263**(7): 3432-3439.

Jerlstrom-Hultqvist, J., Einarsson, E. and Svard, S. G. (2012). Stable transfection of the diplomonad parasite *Spironucleus salmonicida*. *Eukaryot Cell* **11**(11): 1353-1361.

Jerlstrom-Hultqvist, J., Einarsson, E., Xu, F., Hjort, K., Ek, B., Steinhauf, D., Hultenby, K., Bergquist, J., Andersson, J. O. and Svard, S. G. (2013). Hydrogenosomes in the diplomonad *Spironucleus salmonicida*. *Nat Commun* **4**: 2493.

Jerlstrom-Hultqvist, J., Franzen, O., Ankarklev, J., Xu, F., Nohynkova, E., Andersson, J. O., Svard, S. G. and Andersson, B. (2010). Genome analysis and comparative genomics of a *Giardia intestinalis* assemblage E isolate. *BMC Genomics* **11**: 543.

Jerlstrom-Hultqvist, J., Stadelmann, B., Birkestedt, S., Hellman, U. and Svard, S. G. (2012). Plasmid vectors for proteomic analyses in *Giardia*: purification of virulence factors and analysis of the proteasome. *Eukaryot Cell* **11**(7): 864-873.

Juneau, K., Miranda, M., Hillenmeyer, M. E., Nislow, C. and Davis, R. W. (2006). Introns regulate RNA and protein abundance in yeast. *Genetics* **174**(1): 511-518.

Juneau, K., Nislow, C. and Davis, R. W. (2009). Alternative splicing of PTC7 in Saccharomyces cerevisiae determines protein localization. *Genetics* **183**(1): 185-194.

Jurica, M. S. and Moore, M. J. (2003). Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell* **12**(1): 5-14.

Kabnick, K. S. and Peattie, D. A. (1990). *In situ* analyses reveal that the two nuclei of *Giardia lamblia* are equivalent. *J of Cell Sci* **95**: 353-360.

Kamikawa, R., Inagaki, Y. and Hashimoto, T. (2014). Secondary loss of a cis-spliced intron during the divergence of Giardia intestinalis assemblages. *BMC Res Notes* **7**: 413.

Kamikawa, R., Inagaki, Y., Tokoro, M., Roger, A. J. and Hashimoto, T. (2011). Split introns in the genome of *Giardia intestinalis* are excised by spliceosome-mediated *trans*-splicing. *Curr Biol* **21**(4): 311-315.

Karijolich, J. and Yu, Y. T. (2010). Spliceosomal snRNA modifications and their function. *RNA Biol* **7**(2): 192-204.

Keeling, P. J., Corradi, N., Morrison, H. G., Haag, K. L., Ebert, D., Weiss, L. M., Akiyoshi, D. E. and Tzipori, S. (2010). The reduced genome of the parasitic microsporidian Enterocytozoon bieneusi lacks genes for core carbon metabolism. *Genome Biol Evol* **2**: 304-309.

Keeling, P. J. and Doolittle, W. F. (1997). Widespread and ancient distribution of a noncanonical genetic code in diplomonads. *Mol Biol Evol* **14**(9): 895-901.

Kiss, T. (2004). Biogenesis of small nuclear RNPs. *J Cell Sci* **117**(Pt 25): 5949-5951.

Klein, D. J., Schmeing, T. M., Moore, P. B. and Steitz, T. A. (2001). The kink-turn: a new RNA secondary structure motif. *EMBO J* **20**(15): 4214-4221.

Kol, G., Lev-Maor, G. and Ast, G. (2005). Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. *Hum Mol Genet* **14**(11): 1559-1568.

Koonin, E. V. (2006). The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol Direct* **1**: 22.

Korneta, I., Magnus, M. and Bujnicki, J. M. (2012). Structural bioinformatics of the human spliceosomal proteome. *Nucleic Acids Res* **40**(15): 7046-7065.

Laggerbauer, B., Achsel, T. and Luhrmann, R. (1998). The human U5-200kD DEXH-box protein unwinds U4/U6 RNA helices *in vitro*. *Proc Natl Acad Sci U S A* **95**(8): 4188-4192.

Lambowitz, A. M. and Zimmerly, S. (2011). Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol* **3**(8): a003616.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Lane, C. E., van den Heuvel, K., Kozera, C., Curtis, B. A., Parsons, B. J., Bowman, S. and Archibald, J. M. (2007). Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. *Proc Natl Acad Sci U S A* **104**(50): 19908-19913.

Lane, S. and Lloyd, D. (2002). Current trends in research into the waterborne parasite *Giardia*. *Crit Rev Microbiol* **28**(2): 123-147.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**(21): 2947-2948.

Le Blancq, S. M., Kase, R. S. and Van der Ploeg, L. H. (1991). Analysis of a *Giardia lamblia* rRNA encoding telomere with [TAGGG]n as the telomere repeat. *Nucleic Acids Res* **19**(20): 5790.

Lee, R. C., Gill, E. E., Roy, S. W. and Fast, N. M. (2010). Constrained intron structures in a microsporidian. *Mol Biol Evol* **27**(9): 1979-1982.

Lewandowska, D., Simpson, C. G., Clark, G. P., Jennings, N. S., Barciszewska-Pacak, M., Lin, C. F., Makalowski, W., Brown, J. W. and Jarmolowski, A. (2004). Determinants of plant U12-dependent intron splicing efficiency. *Plant Cell* **16**(5): 1340-1352.

Li, B. L., Li, X. L., Duan, Z. J., Lee, O., Lin, S., Ma, Z. M., Chang, C. C., Yang, X. Y., Park, J. P., Mohandas, T. K., *et al.* (1999). Human acyl-CoA:cholesterol acyltransferase-1 (ACAT-1) gene organization and evidence that the 4.3-kilobase ACAT-1 mRNA is produced from two different chromosomes. *J Biol Chem* **274**(16): 11060-11071.

Liang, X. H., Haritan, A., Uliel, S. and Michaeli, S. (2003). Trans and cis splicing in trypanosomatids: mechanism, factors, and regulation. *Eukaryot Cell* **2**(5): 830-840.

Lim, L. P. and Burge, C. B. (2001). A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A* **98**(20): 11193-11198.

Liu, S., Rauhut, R., Vornlocher, H. P. and Luhrmann, R. (2006). The network of protein-protein interactions within the human U4/U6.U5 tri-snRNP. *RNA* **12**(7): 1418-1430.

Logsdon, J. M., Jr. (1998). The recent origins of spliceosomal introns revisited. *Curr Opin Genet Dev* **8**(6): 637-648.

Long, M., de Souza, S. J. and Gilbert, W. (1995). Evolution of the intron-exon structure of eukaryotic genes. *Curr Opin Genet Dev* **5**(6): 774-778.

Long, M., Rosenberg, C. and Gilbert, W. (1995). Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc Natl Acad Sci U S A* **92**(26): 12495-12499.

Lopez, M., Rosenblad, M. A. and Samuelsson, T. (2008). Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. *Nucleic Acids Res* **36**(9): 3001-3010.

Lopez, P. J. and Seraphin, B. (2000). YIDB: the Yeast Intron DataBase. *Nucleic Acids Res* **28**(1): 85-86.

Luo, J., Zhou, H., Chen, C., Li, Y., Chen, Y. and Qu, L. (2006). Identification and evolutionary implication of four novel box H/ACA snoRNAs from *Giardia lamblia*. *Chinese Science Bulletin* **51**: 2451–2456.

Luz Ambrosio, D., Lee, J. H., Panigrahi, A. K., Nguyen, T. N., Cicarelli, R. M. and Gunzl, A. (2009). Spliceosomal proteomics in *Trypanosoma brucei* reveal new RNA splicing factors. *Eukaryot Cell* **8**(7): 990-1000.

Lynch, M. and Richardson, A. O. (2002). The evolution of spliceosomal introns. *Curr Opin Genet Dev* **12**(6): 701-710.

Madhani, H. D. and 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**(5): 803-817.

Maki, R., Roeder, W., Traunecker, A., Sidman, C., Wabl, M., Raschke, W. and Tonegawa, S. (1981). The role of DNA rearrangement and alternative RNA processing in the expression of immunoglobulin delta genes. *Cell* **24**(2): 353-365.

Massenet, S., Motorin, Y., Lafontaine, D. L., Hurt, E. C., Grosjean, H. and Branlant, C. (1999). Pseudouridine mapping in the Saccharomyces cerevisiae spliceosomal U small nuclear RNAs (snRNAs) reveals that pseudouridine synthase pus1p exhibits a dual substrate specificity for U2 snRNA and tRNA. *Mol Cell Biol* **19**(3): 2142-2154.

Matlin, A. J. and Moore, M. J. (2007). Spliceosome assembly and composition. *Adv Exp Med Biol* **623**: 14-35.

Matsuura, M., Saldanha, R., Ma, H., Wank, H., Yang, J., Mohr, G., Cavanagh, S., Dunny, G. M., Belfort, M. and Lambowitz, A. M. (1997). A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes Dev* **11**(21): 2910-2924.

McManus, C. J. and Graveley, B. R. (2011). RNA structure and the mechanisms of alternative splicing. *Curr Opin Genet Dev* **21**(4): 373-379.

Meyer, M., Plass, M., Perez-Valle, J., Eyras, E. and Vilardell, J. (2011). Deciphering 3'ss selection in the yeast genome reveals an RNA thermosensor that mediates alternative splicing. *Mol Cell* **43**(6): 1033-1039.

Michel, F., Costa, M. and Westhof, E. (2009). The ribozyme core of group II introns: a structure in want of partners. *Trends Biochem Sci* **34**(4): 189-199.

Mitrovich, Q. M. and Guthrie, C. (2007). Evolution of small nuclear RNAs in *S. cerevisiae*, *C. albicans*, and other hemiascomycetous yeasts. *RNA* **13**(12): 2066-2080.

Montzka, K. A. and Steitz, J. A. (1988). Additional low-abundance human small nuclear ribonucleoproteins: U11, U12, etc. *Proc Natl Acad Sci U S A* **85**(23): 8885-8889.

Morrison, H. G., McArthur, A. G., Gillin, F. D., Aley, S. B., Adam, R. D., Olsen, G. J., Best, A. A., Cande, W. Z., Chen, F., Cipriano, M. J., *et al.* (2007). Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. *Science* **317**(5846): 1921-1926.

Mouaikel, J., Verheggen, C., Bertrand, E., Tazi, J. and Bordonne, R. (2002). Hypermethylation of the cap structure of both yeast snRNAs and snoRNAs requires a conserved methyltransferase that is localized to the nucleolus. *Molecular Cell* **9**(4): 891-901.

Muller, S., Wolpensinger, B., Angenitzki, M., Engel, A., Sperling, J. and Sperling, R. (1998). A supraspliceosome model for large nuclear ribonucleoprotein particles based on mass determinations by scanning transmission electron microscopy. *J Mol Biol* **283**(2): 383-394.

Nadimi, M., Beaudet, D., Forget, L., Hijri, M. and Lang, B. F. (2012). Group I intron-mediated trans-splicing in mitochondria of *Gigaspora rosea* and a robust phylogenetic affiliation of arbuscular mycorrhizal fungi with Mortierellales. *Mol Biol Evol* **29**(9): 2199-2210.

Nageshan, R. K., Roy, N., Hehl, A. B. and Tatu, U. (2011). Post-transcriptional repair of a split heat shock protein 90 gene by mRNA trans-splicing. *J Biol Chem* **286**(9): 7116-7122.

Nakao, A., Yoshihama, M. and Kenmochi, N. (2004). RPG: the Ribosomal Protein Gene database. *Nucleic Acids Res* **32**(Database issue): D168-170.

Nawrocki, E. P. and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**(22): 2933-2935.

Newman, A. J. and Norman, C. (1992). U5 snRNA interacts with exon sequences at 5' and 3' splice sites. *Cell* **68**(4): 743-754.

Nielsen, H. and Wernersson, R. (2006). An overabundance of phase 0 introns immediately after the start codon in eukaryotic genes. *BMC Genomics* **7**: 256.

Nilsen, T. W. and Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**(7280): 457-463.

Nixon, J. E., Wang, A., Morrison, H. G., McArthur, A. G., Sogin, M. L., Loftus, B. J. and Samuelson, J. (2002). A spliceosomal intron in *Giardia lamblia*. *Proc Natl Acad Sci U S A* **99**(6): 3701-3705.

Ohno, M., Segref, A., Bachi, A., Wilm, M. and Mattaj, I. W. (2000). PHAX, a mediator of U snRNA nuclear export whose activity is regulated by phosphorylation. *Cell* **101**(2): 187-198.

Padgett, R. A. and Shukla, G. C. (2002). A revised model for U4atac/U6atac snRNA base pairing. *RNA* **8**(2): 125-128.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J. and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**(12): 1413-1415.

Parenteau, J., Durand, M., Morin, G., Gagnon, J., Lucier, J. F., Wellinger, R. J., Chabot, B. and Abou Elela, S. (2011). Introns within Ribosomal Protein Genes Regulate the Production and Function of Yeast Ribosomes. *Cell* **147**(2): 320-331.

Patel, A. A., McCarthy, M. and Steitz, J. A. (2002). The splicing of U12-type introns can be a rate-limiting step in gene expression. *EMBO J* **21**(14): 3804-3815.

Patel, A. A. and Steitz, J. A. (2003). Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol* **4**(12): 960-970.

Phizicky, E. M. and Hopper, A. K. (2010). tRNA biology charges to the front. *Genes Dev* **24**(17): 1832-1860.

Podlevsky, J. D., Bley, C. J., Omana, R. V., Qi, X. and Chen, J. J. (2008). The telomerase database. *Nucleic Acids Res* **36**(Database issue): D339-343.

Pomeranz Krummel, D. A., Oubridge, C., Leung, A. K., Li, J. and Nagai, K. (2009). Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution. *Nature* **458**(7237): 475-480.

Qu, G., Dong, X., Piazza, C. L., Chalamcharla, V. R., Lutz, S., Curcio, M. J. and Belfort, M. (2014). RNA-RNA interactions and pre-mRNA mislocalization as drivers of group II intron loss from nuclear genomes. *Proc Natl Acad Sci U S A* **111**(18): 6612-6617.

Randau, L., Munch, R., Hohn, M. J., Jahn, D. and Soll, D. (2005). *Nanoarchaeum equitans* creates functional tRNAs from separate genes for their 5'- and 3'-halves. *Nature* **433**(7025): 537-541.

Reddy, R., Henning, D. and Busch, H. (1981). Pseudouridine residues in the 5'-terminus of uridine-rich nuclear RNA I (U1 RNA). *Biochem Biophys Res Commun* **98**(4): 1076-1083.

Rest, J. S. and Mindell, D. P. (2003). Retroids in archaea: phylogeny and lateral origins. *Mol Biol Evol* **20**(7): 1134-1142.

Robertson, H. M., Navik, J. A., Walden, K. K. and Honegger, H. W. (2007). The bursicon gene in mosquitoes: an unusual example of mRNA trans-splicing. *Genetics* **176**(2): 1351-1353.

Roca, X. and Krainer, A. R. (2009). Recognition of atypical 5' splice sites by shifted base-pairing to U1 snRNA. *Nat Struct Mol Biol* **16**(2): 176-182.

Roger, A. J., Svard, S. G., Tovar, J., Clark, C. G., Smith, M. W., Gillin, F. D. and Sogin, M. L. (1998). A mitochondrial-like chaperonin 60 gene in *Giardia lamblia*: evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria. *Proc Natl Acad Sci U S A* **95**(1): 229-234.

Rogozin, I. B., Carmel, L., Csuros, M. and Koonin, E. V. (2012). Origin and evolution of spliceosomal introns. *Biol Direct* **7**: 11.

Rogozin, I. B., Wolf, Y. I., Sorokin, A. V., Mirkin, B. G. and Koonin, E. V. (2003). Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol* **13**(17): 1512-1517.

Roxstrom-Lindquist, K., Jerlstrom-Hultqvist, J., Jorgensen, A., Troell, K., Svard, S. G. and Andersson, J. O. (2010). Large genomic differences between the morphologically indistinguishable diplomonads *Spironucleus barkhanus* and *Spironucleus salmonicida*. *BMC Genomics* **11**: 258.

Roy, S. W. and Gilbert, W. (2006). The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet* **7**(3): 211-221.

Roy, S. W., Hudson, A. J., Joseph, J., Yee, J. and Russell, A. G. (2012). Numerous fragmented spliceosomal introns, AT-AC splicing, and an unusual dynein gene expression pathway in *Giardia lamblia*. *Mol Biol Evol* **29**(1): 43-49.

Russell, A. G., Charette, J. M., Spencer, D. F. and Gray, M. W. (2006). An early evolutionary origin for the minor spliceosome. *Nature* **443**(7113): 863-866.

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

Russell, A. G., Watanabe, Y., Charette, J. M. and Gray, M. W. (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**(9): 2781-2791.

Sakharkar, M. K., Chow, V. T. and Kangueane, P. (2004). Distributions of exons and introns in the human genome. *In Silico Biol* **4**(4): 387-393.

Sashital, D. G., Cornilescu, G., McManus, C. J., Brow, D. A. and Butcher, S. E. (2004). U2-U6 RNA folding reveals a group II intron-like domain and a four-helix junction. *Nat Struct Mol Biol* **11**(12): 1237-1242.

Sawa, H. and Abelson, J. (1992). Evidence for a base-pairing interaction between U6 small nuclear RNA and 5' splice site during the splicing reaction in yeast. *Proc Natl Acad Sci U S A* **89**(23): 11269-11273.

Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., Dixon, J. E. and Zipursky, S. L. (2000). *Drosophila Dscam* is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**(6): 671-684.

Schneider, C., Will, C. L., Makarova, O. V., Makarov, E. M. and Luhrmann, R. (2002). Human U4/U6.U5 and U4atac/U6atac.U5 tri-snRNPs exhibit similar protein compositions. *Mol Cell Biol* **22**(10): 3219-3229.

Scofield, D. G. and Lynch, M. (2008). Evolutionary diversification of the Sm family of RNA-associated proteins. *Mol Biol Evol* **25**(11): 2255-2267.

Seetharaman, M., Eldho, N. V., Padgett, R. A. and Dayie, K. T. (2006). Structure of a self-splicing group II intron catalytic effector domain 5: parallels with spliceosomal U6 RNA. *RNA* **12**(2): 235-247.

Seraphin, B., Kretzner, L. and Rosbash, M. (1988). A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5' cleavage site. *EMBO J* **7**(8): 2533-2538.

Shukla, G. C. and Padgett, R. A. (1999). Conservation of functional features of U6atac and U12 snRNAs between vertebrates and higher plants. *RNA* **5**(4): 525-538.

Shukla, G. C. and Padgett, R. A. (2001). The intramolecular stem-loop structure of U6 snRNA can functionally replace the U6atac snRNA stem-loop. *RNA* **7**(1): 94-105.

Shukla, G. C. and Padgett, R. A. (2002). A catalytically active group II intron domain 5 can function in the U12-dependent spliceosome. *Mol Cell* **9**(5): 1145-1150.

Sikand, K. and Shukla, G. C. (2011). Functionally important structural elements of U12 snRNA. *Nucleic Acids Res* **39**(19): 8531-8543.

Siliciano, P. G. and Guthrie, C. (1988). 5' splice site selection in yeast: genetic alterations in base-pairing with U1 reveal additional requirements. *Genes Dev* **2**(10): 1258-1267.

Simoes-Barbosa, A., Meloni, D., Wohlschlegel, J. A., Konarska, M. M. and Johnson, P. J. (2008). Spliceosomal snRNAs in the unicellular eukaryote *Trichomonas vaginalis* are structurally conserved but lack a 5'-cap structure. *RNA* **14**(8): 1617-1631.

Sogin, M. L. (1991). Early evolution and the origin of eukaryotes. *Curr Opin Genet Dev* **1**(4): 457-463.

Sogin, M. L., Gunderson, J. H., Elwood, H. J., Alonso, R. A. and Peattie, D. A. (1989). Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia*. *Science* **243**(4887): 75-77.

Spingola, M., Grate, L., Haussler, D. and Ares, M., Jr. (1999). Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA* **5**(2): 221-234.

Straube, A., Enard, W., Berner, A., Wedlich-Soldner, R., Kahmann, R. and Steinberg, G. (2001). A split motor domain in a cytoplasmic dynein. *EMBO J* **20**(18): 5091-5100.

Suchy, M. and Schmelzer, C. (1991). Restoration of the self-splicing activity of a defective group II intron by a small *trans*-acting RNA. *J Mol Biol* **222**(2): 179-187.

Surowy, C. S., van Santen, V. L., Scheib-Wixted, S. M. and Spritz, R. A. (1989). Direct, sequence-specific binding of the human U1-70K ribonucleoprotein antigen protein to loop I of U1 small nuclear RNA. *Mol Cell Biol* **9**(10): 4179-4186.

Sverdlov, A. V., Rogozin, I. B., Babenko, V. N. and Koonin, E. V. (2005). Conservation versus parallel gains in intron evolution. *Nucleic Acids Res* **33**(6): 1741-1748.

Takahara, T., Kanazu, S. I., Yanagisawa, S. and Akanuma, H. (2000). Heterogeneous Sp1 mRNAs in human HepG2 cells include a product of homotypic *trans*-splicing. *J Biol Chem* **275**(48): 38067-38072.

Tang, J. and Rosbash, M. (1996). Characterization of yeast U1 snRNP A protein: identification of the N-terminal RNA binding domain (RBD) binding site and evidence that the C-terminal RBD functions in splicing. *RNA* **2**(10): 1058-1070.

Tarn, W. Y. and Steitz, J. A. (1996). Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. *Science* **273**(5283): 1824-1832.

Tarn, W. Y. and Steitz, J. A. (1996). A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell* **84**(5): 801-811.

Tessier, L. H., Chan, R. L., Keller, M., Weil, J. H. and Imbault, P. (1992). The *Euglena gracilis* rbcS gene contains introns with unusual borders. *FEBS Lett* **304**(2-3): 252-255.

Toor, N., Keating, K. S., Taylor, S. D. and Pyle, A. M. (2008). Crystal structure of a self-spliced group II intron. *Science* **320**(5872): 77-82.

Toro, N., Jimenez-Zurdo, J. I. and Garcia-Rodriguez, F. M. (2007). Bacterial group II introns: not just splicing. *FEMS Microbiol Rev* **31**(3): 342-358.

Tovar, J., Leon-Avila, G., Sanchez, L. B., Sutak, R., Tachezy, J., van der Giezen, M., Hernandez, M., Muller, M. and Lucocq, J. M. (2003). Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation. *Nature* **426**(6963): 172-176.

Upcroft, J. A., Krauer, K. G. and Upcroft, P. (2010). Chromosome sequence maps of the *Giardia lamblia* assemblage A isolate WB. *Trends Parasitol* **26**(10): 484-491.

Valadkhan, S. and Jaladat, Y. (2010). The spliceosomal proteome: at the heart of the largest cellular ribonucleoprotein machine. *Proteomics* **10**(22): 4128-4141.

Vanacova, S., Yan, W., Carlton, J. M. and Johnson, P. J. (2005). Spliceosomal introns in the deep-branching eukaryote *Trichomonas vaginalis*. *Proc Natl Acad Sci U S A* **102**(12): 4430-4435.

Vankan, P., McGuigan, C. and Mattaj, I. W. (1992). Roles of U4 and U6 snRNAs in the assembly of splicing complexes. *EMBO J* **11**(1): 335-343.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001). The sequence of the human genome. *Science* **291**(5507): 1304-1351.

Vicens, Q. and Cech, T. R. (2006). Atomic level architecture of group I introns revealed. *Trends Biochem Sci* **31**(1): 41-51.

Vidovic, I., Nottrott, S., Hartmuth, K., Luhrmann, R. and Ficner, R. (2000). Crystal structure of the spliceosomal 15.5kD protein bound to a U4 snRNA fragment. *Mol Cell* **6**(6): 1331-1342.

Wahl, M. C., Will, C. L. and Luhrmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**(4): 701-718.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P. and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**(7221): 470-476.

Wang, Z. and Burge, C. B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**(5): 802-813.

Watanabe, K. and Lambowitz, A. M. (2004). High-affinity binding site for a group II intron-encoded reverse transcriptase/maturase within a stem-loop structure in the intron RNA. *RNA* **10**(9): 1433-1443.

Watkins, N. J. and Bohnsack, M. T. (2012). The box C/D and H/ACA snoRNPs: key players in the modification, processing and the dynamic folding of ribosomal RNA. *Wiley Interdiscip Rev RNA* **3**(3): 397-414.

Will, C. L. and Luhrmann, R. (2001). Spliceosomal UsnRNP biogenesis, structure and function. *Curr Opin Cell Biol* **13**(3): 290-301.

Will, C. L. and Luhrmann, R. (2005). Splicing of a rare class of introns by the U12-dependent spliceosome. *Biol Chem* **386**(8): 713-724.

Will, C. L. and Luhrmann, R. (2011). Spliceosome structure and function. *Cold Spring Harb Perspect Biol* **3**(7).

Will, C. L., Schneider, C., Reed, R. and Luhrmann, R. (1999). Identification of both shared and distinct proteins in the major and minor spliceosomes. *Science* **284**(5422): 2003-2005.

Williams, C. F., David, L., Poynton, S. L., Anders, J., Millet, C. O. and Joanne, C. (2011). *Spironucleus* species: Economically-important fish pathogens and enigmatic single-celled eukaryotes. *Journal of Aquaculture Research & Development* **S2**(002).

Wolff, T. and Bindereif, A. (1993). Conformational changes of U6 RNA during the spliceosome cycle: an intramolecular helix is essential both for initiating the U4-U6 interaction and for the first step of slicing. *Genes Dev* **7**(7B): 1377-1389.

Xu, F., Jerlstrom-Hultqvist, J., Einarsson, E., Astvaldsson, A., Svard, S. G. and Andersson, J. O. (2014). The genome of Spironucleus salmonicida highlights a fish pathogen adapted to fluctuating environments. *PLoS Genet* **10**(2): e1004053.

Yang, C. Y., Zhou, H., Luo, J. and Qu, L. H. (2005). Identification of 20 snoRNA-like RNAs from the primitive eukaryote, *Giardia lamblia*. *Biochem Biophys Res Commun* **328**(4): 1224-1231.

Yee, J., Mowatt, M. R., Dennis, P. P. and Nash, T. E. (2000). Transcriptional analysis of the glutamate dehydrogenase gene in the primitive eukaryote, *Giardia lamblia*. Identification of a primordial gene promoter. *J Biol Chem* **275**(15): 11432-11439.

Yee, J., Tang, A., Lau, W. L., Ritter, H., Delport, D., Page, M., Adam, R. D., Muller, M. and Wu, G. (2007). Core histone genes of *Giardia intestinalis*: genomic organization, promoter structure, and expression. *BMC Mol Biol* **8**: 26.

Yoshihisa, T. (2014). Handling tRNA introns, archaeal way and eukaryotic way. *Front Genet* **5**: 213.

Yu, L. Z., Birky, C. W. and Adam, R. D. (2002). The two nuclei of *Giardia* each have complete copies of the genome and are partitioned equationally at cytokinesis. *Eukaryotic Cell* **1**(2): 191-199.

Yuan, F., Griffin, L., Phelps, L., Buschmann, V., Weston, K. and Greenbaum, N. L. (2007). Use of a novel Forster resonance energy transfer method to identify locations of site-bound metal ions in the U2-U6 snRNA complex. *Nucleic Acids Res* **35**(9): 2833-2845.

Zamore, P. D. and Green, M. R. (1989). Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. *Proc Natl Acad Sci U S A* **86**(23): 9243-9247.

Zhang, L. and Doudna, J. A. (2002). Structural insights into group II intron catalysis and branch-site selection. *Science* **295**(5562): 2084-2088.

Zhang, X., Tolzmann, C. A., Melcher, M., Haas, B. J., Gardner, M. J., Smith, J. D. and Feagin, J. E. (2011). Branch point identification and sequence requirements for intron splicing in *Plasmodium falciparum*. *Eukaryot Cell* **10**(11): 1422-1428.

Zhuang, Y. A., Goldstein, A. M. and Weiner, A. M. (1989). UACUAAC is the preferred branch site for mammalian mRNA splicing. *Proc Natl Acad Sci U S A* **86**(8): 2752-2756.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**(13): 3406-3415.

Appendix 1 - Supplementary Material for Chapter 2:

Numerous Fragmented Spliceosomal Introns, AT–AC Splicing, and an Unusual Dynein Gene Expression Pathway in *Giardia lamblia*

A. *Cis*-Spliced Introns

5′-A/ctatgtt...(13nt)..acaactaacacacag/C-3′ **Ferredoxin**

5′-G/gtatgtt...(10nt)..taacctaacacacag/A-3′ **Dynein-like**

5′-A/gtatgtt...(87nt)..acaactgacccacag/C-3′ ***Rpl7a***

5′-A/gtatgtt..(198nt)..ccaactgacacacag/C-3′ **Unassigned ORF on contig [AACB01000025]**

5′-T/gtatgtt...(7nt)...ccatctaacacccag/C-3′ **26S proteasome non-ATPase regulatory subunit 4**

B. *Trans*-Spliced Introns

5′-T/gtatgtt...// ***Hsp90* N-half (ORF98054)**
//...ataactaacacgcag/G-3′ ***Hsp90* C-half (ORF13864)**

5′-G/gtatgtt...// ***DHC β* Outer-Arm Intron 1 5′ half (ORF17243)**
//...aatactaacacacag/C-3′ ***DHC β* Outer-Arm Intron 1 3′ half (ORF10538)**

5′-G/gtatgtt...// ***DHC β* Outer-Arm Intron 2 5′ half (ORF10538)**
//...catactaacacacag/C-3′ ***DHC β* Outer-Arm Intron 2 3′ half (ORF8172)**

5′-G/atatgtt...// ***DHC γ* Outer-Arm N-half (ORF16804)**
//...gctaacacacagcac/C-3′ ***DHC γ* Outer-Arm C-half (ORF17265)**

Figure A.1.1. Conserved splice site boundaries in *G. lamblia* spliceosomal introns.
A. *cis*-spliced *G. lamblia* introns in genes coding for ferredoxin (Nixon et al., 2002), a dynein-like protein (Morrison et al., 2007), *Rpl7a* (Russell et al., 2005), an unassigned ORF (Russell et al., 2005) and 26S proteasome non-ATPase regulatory subunit 4 (identified in this study) show strong sequence conservation at both their 5′ and 3′ splice sites. **B.** Fragmented *trans*-spliced introns also show similar sequence conservation. These introns interrupt open-reading frames (ORFs) numbered according to the *Giardia* genome database v1.3 at http://giardiadb.org/giardiadb/. Intronic sequences are in lower case letters and their predicted branch point sequences are underlined. Single slashes indicate exon-intron junctions whereas double slashes indicate discontinuity due to genomic intron fragmentation.

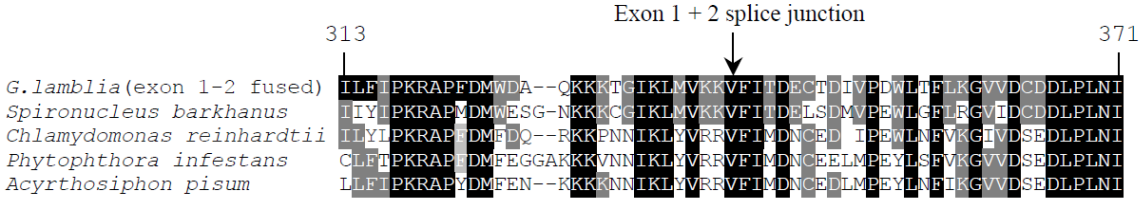


Figure A.1.2. Hsp90 protein sequence alignment.
ClustalW2 generated amino acid sequence alignment of the region corresponding to positions 313 to 371 of the *C. reinhardtii* protein. The arrow indicates the location of the *trans*-spliced *G. lamblia* intron. The *G. lamblia* sequence shown is the translated sequence following trans-splicing of the two mRNAs. Sources of the sequences are: *Spironucleus barkhanus* (GenBank ABC54647), *Chlamydomonas reinhardtii* (GenBank XP_001695264), *Phythophthora infestans* (EEY69894.1), *Acyrthosiphon pisum* (XP_001943172), and *G. lamblia* (exon 1 from nucleotide positions 28161 – 29200 on contig AACB02000010; exon 2 from nucleotide 47951 – 49025 on contig AACB02000034 at the *Giardia* genome database http://giardiadb.org/giardiadb/).

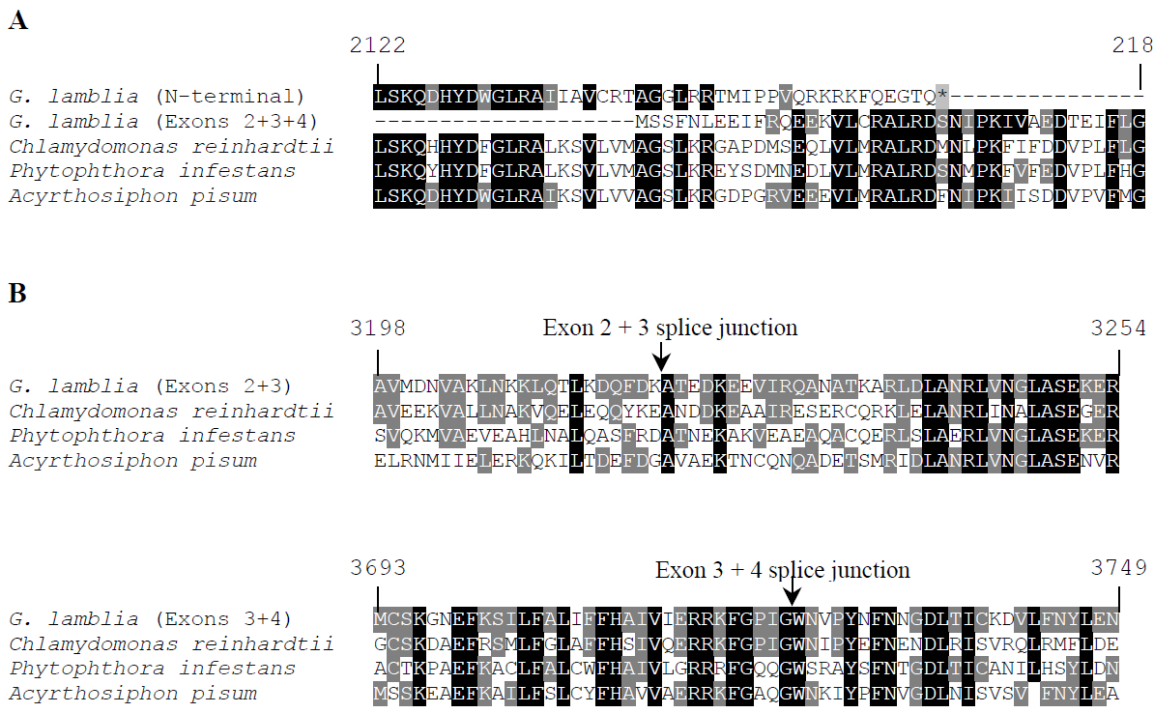


Figure A.1.3. Dynein heavy chain β outer-arm protein sequence alignment.
ClustalW2 generated amino acid sequence alignments near the break between the N- and C-terminal half polypeptides of the *G. lamblia* protein **(A)** and in the regions flanking the boundaries of the two *trans*-spliced *G. lamblia* introns **(B)**. Amino acid numbering corresponds to the *C. reinhardtii* sequence. *G. lamblia DHC β* "exon" 1 encodes a complete conserved N-terminal half of this protein and the *trans*- spliced exons 2-4 encode the complete conserved C-terminal portion. Arrows indicate the splicing junctions for each of the *G. lamblia DHC β* exons. The asterisk (*) on grey background denotes an in- frame stop codon for *DHC β* "exon" 1. Sources of the sequences are: *Chlamydomonas reinhardtii* (GenBank XP_001703170), *Phytophthora infestans* (EEY61420), *Acyrthosiphon pisum* (XP_001949713), and *G. lamblia* (exon 1 is encoded between nucleotide positions 46134 – 53354 on contig AACB02000053; exon 2 from 125582 – 129746 on contig AACB02000024; exon 3 from 174113 – 176532 on contig AACB02000007; exon 4 from 30741 – 31995 on contig AACB02000034 at the *Giardia* genome database http://giardiadb.org/giardiadb/).

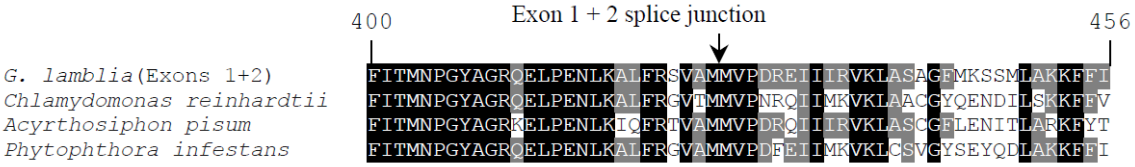


Figure A.1.4. Dynein heavy chain γ outer-arm protein sequence alignment.
ClustalW2 generated amino acid sequence alignment of the region corresponding to positions 400-456 of the *C. reinhardtii* protein. An arrow indicates the location of the *trans*-spliced fragmented *G. lamblia* intron. Sources of the sequences are: *Chlamydomonas reinhardtii* (GenBank XP_001702026), *Phytophthora infestans* (EEY70506), *Acyrthosiphon pisum* (XP_001943595), and *G. lamblia* (exon 1 is located between nt positions 193841 – 200650 on contig AACB02000001; exon 2 between nt 147472 – 155499 on contig AACB02000023 at the *Giardia* genome database http://giardiadb.org/giardiadb/).

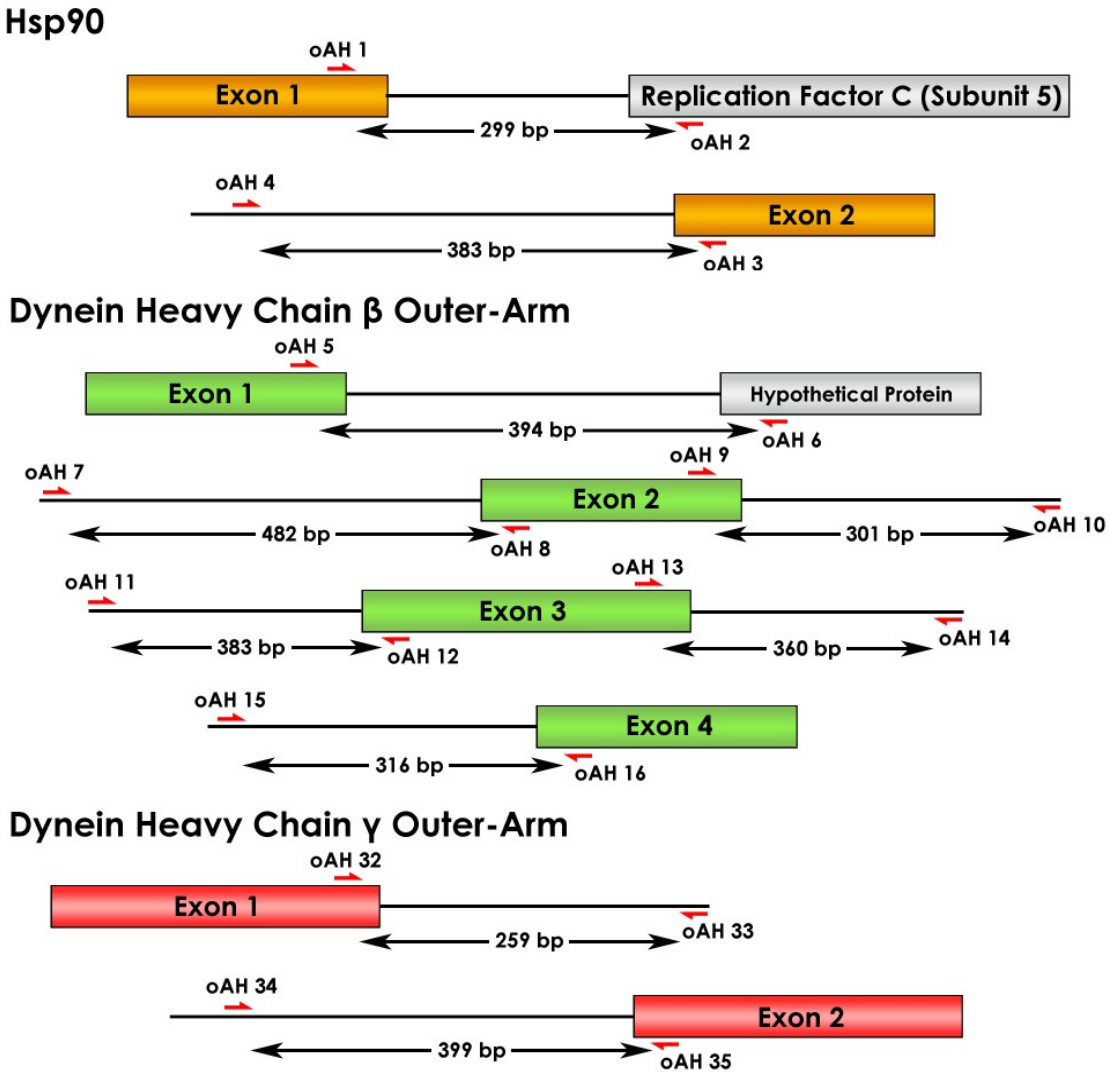


Figure A.1.5. Schematic representation of the annealing positions of oligonucleotides used for *Giardia* mRNA analysis and genomic DNA amplification.
Relative annealing positions for primers used for amplification of *Hsp90* and dynein isoform gene fragments are indicated by red arrows with arrowheads indicating the direction of amplification. Exons or open reading frames (ORFs) are illustrated as colored boxes, introns or non-coding sequences are represented as lines. Grey boxes denote adjacent protein-coding genes as annotated by the *Giardia* genome database http://giardiadb.org/giardiadb/. Refer to Appendix IV for primer DNA sequences.

Table A.1.1. EST verification of the production and processing of *G. lamblia* dynein heavy chain and Hsp90 protein-coding mRNAs.

Espressed Sequence Tags (EST) identification numbers are given as annotated by the *G. lamblia* genome database (http://giardiadb.org/giardiadb/).

Gene of Interest	Description	Confirming EST(s)
Dynein Heavy Chain (DHC) *β* Outer Arm	Confirmation of expression and polyadenylation of DHC β Exon 1 mRNA.	EV519056, EV504297, EV501365, EV500635, EV510597, EV505838, EV512357, EV502134, EV519093, EV507057, EV507624, EV499363.
	Confirmation of expression and polyadenylation of DHC β Exon 4 mRNA.	EV507166, EV511588, EV501206, EV517662, EV507092, EV512968, EV515901, EV504972, EV511853.
	Verification of splicing at DHC β Exon 3-Exon 4 junction	EV501205, EV504971, EV519392, EV511587.
Dynein Heavy Chain (DHC) *γ* Outer Arm	Evidence for expression and polyadenylation of DHC γ Exon 2	EV499943, EV510793, EV514693, EV508969, EV515499, EV507849,
Hsp90	Confirmation of expression and polyadenylation of *Hsp90* Exon 2 mRNA.	EV507984, EV515374, EV501277, EV499960, EV500047, EV502455, EV507480, EV512110, EV509377, EV512171, EV518598, EV518598. (Note: confirmed by ~50 ESTs – others not shown)
	Verification of splicing at Hsp90 Exon 1-2 junction	EV510245, EV500048, EV499718, EV517261, EV519052, EV500404, EV519475, EV503441, EV501278, EV501772, EV510008, EV514223. (Note: Confirmed by ~16 ESTs – others not shown)

Figure A.1.6. Sequences of *cis*- and *trans*-spliced introns from available *Giardia* genome assemblies.
Intronic sequences are in lower-case red text and exonic sequences in upper-case black text. Complementary regions of intron halves are shown with green and blue highlighting. Scaffold and nucleotide positions are indicated.

A) Hsp90

>G. lamblia* – P15 *Hsp90

Exon 1 (contig421:16108-16564)

CCAGAGAAGAAGGACGACGAGAAGAAAGAAGACGAGAAGAAGGAAGATGAGGTGCATGAGGCCTCCGATGAC
GAAGAGAAGAAGGAGGAGAAGAAGAAGACTATCAAGAAGGAAGTCGAAGTCCACGAACACGTAAACAAGCAG
CCCGCGATCTGGACCAGGGATCCGAAGGACGTCACCGAGGACGAATACAAGGACTTCTACAAGCAGATCAAC
CCCTCCGACTACGAGGGCCACCTTGCTGTGTCGCACTTCCGCGTAGACGGCGCCGCTCAGTTCCGCGGTATC
CTCTTCATCCCCAAGCGCGCGCCCTTCGACATGTGGGACGCTCAAAAGAAGAAGACGGGCATCAAGCTCATG
GTCAAGAAAGTgtatgttatgttgtatgctgtatgtgtgcgagactcctttactcaaattcgcggcattat
gctgcgggctctccctatgaaaatt

Exon 2 (contig137:47124-47739)

aaataaaattactgatgccttgcaccgcatcgtgacgatgtccgcatgctgaggtatgtgggtgtgtcgcgc
ggtgtccggtctgggcggagttctcgcacacatacagtatacacgacttccttctaactaacacgcagATTT
ATCACAGACGAATGCACGGATATTGTTCCTGACTGGCTCACGTTCCTTAAGGGCGTGGTCGACTGCGACGAT
CTTCCGCTGAACATCTCCCGTGAGATGCTTCAGAAGAATCGCATCGTCAACACAATCCGCAAGAACCTCATC
AACAAGGCTCTCAAGCTCTTCAAGGACCTCGAGGATGACAAGGACAAGTATGAGACATTCCTCAAGCAATTT
GGAAAGTCCATCAAACTCGGCATCCATGAGGACAGCGAGAACCGCGGTAAGCTCGCCAAGCTCCTGCGCTAC
TACAGTACAAAGTCCAAGGACAAGCGCACATCTCTTGATGATTATGTTACTCGCATGCCTGAAAACCAAAAG
TCGATCTACTACATCACAGGCGACTCTCTCGAAAACTTGAAAGAGTCTCCATTCCTCGAGAGATTTAATAAG
AAGGGCATAGAGGTTCTGCTCATGGACGAAGCCATTGACG

>G. lamblia* – GS *Hsp90

Exon 1 (ACGJ01002286:15900-16326)

CCTTGTTGCCGAGAAGGTCGAGGTCATCACCAAGCACAACGACGACGGCTGCTTTAAGTGGAGCTCGACGGC
CGGTGGCACGTTCGAGATTGAGGAGTGCCCGCGCGACTACTTCGGCGAGGAGCTCCAGCGCGGTACCGAGAT
CATCCTCCACCTCAAGGAGGACCAGAAGGAGTACCTGAAGGCCGACCGCCTCGAGGAGCTGATCAAGAAGCA
CTCCATGTTCATCGGCTATCCGATCTACCTCTATAAGACGCGCGAGGAGGAGGTAGAGGTCGAGAGCGAGCC
GGAGAAGAAGGACGAGGAGACGAAGGAGGACGAGAAGAAAGAGGACGAGGTCCACGAGGCTTCTGATGATGA
CGAGAAGAAGGAGGAGAAGAAGAAGACCGTCAAGAAGGAGGTCGAGGTCCATGAGCACGTGAACAAGCAGCC
TGCGATCTGGACCAGGGACCCGAAGGATGTCACTGAGGACGAGTACAAGGATTTCTATAAGCAGATCAACCC
GTCCGACTACGAGGGCCACCTTGCCGTGTCGCACTTCCGCGTCGATGGCGCCGCACAGTTCCGCGGCATCCT
CTTCATCCCTAAGCGCGCGCCCTTCGACATGTGGGACGCTCAGAAGAAGAAGACGGGCATCAAGCTCATGGT
CAAGAAAGTgtatgtttatgtatgtgtgttgtatgtgtgcgagactcctttactcaattcgtggcttcgtgc
tgcgggtcctctttagaaaattcccgtaccagtctcgccggtatgtggtgcaacaagtatacccccacgcag
ctagatgctatggattaccacagggatgctactcggctcttcattagcgttgcccgagcaaaaaacccgcct
caccttcttgtccatgggcctcgaggatcaggccgtcatacccgaattctagcatatatcagaagtgtttat
ggcgtgcaaacgctagattacattccgtcaacaatgctatacgaaacaaatacggatacttgcg

Exon 2 (ACGJ01002298:20179-20792)

aaataaaattaccagtgccttgtaccacatcgtggtgatgcccccatgctgagttatgcgggcatagcgtgc
gatgttcggtcagggtaggattctcgcacacatgcagcatacacgttcctcacaactaacacacagATTTAT
CACAGACGAATGCACAGATATCGTTCCCGACTGGCTCACGTTCCTCAAGGGTGTGGTTGACTGCGACGACCT
TCCTCTAAACATCTCCCGTGAGATGCTCCAGAAGAACCGCATTGTCAACACAATCCGCAAGAACCTCATCAA
CAAGGCTCTCAAGCTCTTCAAGGACCTCGAGGAGGACAAGGATAAGTATGAGACATTCCTCAAGCAGTTCGG
AAAGTCCATCAAGCTCGGCATCCACGAGGATAGCGAAAACCGTGGCAAGCTCGCTAAGCTCCTGCGTTACTA
CAGCACCAAGTCCAAGGACAAGCGCACGTCCCTCGATGACTACGTCACACGCATGCCCGAGAACCAGAAGTC
GATCTATTACATCACGGGAGACTCTCTCGAAAATCTAAAGGAGTCTCCGTTCCTTGAGAGGTTCAACAAGAA
GGGCATCGAGGTACTGCTTATGGACGAAGCCATCGATG

B) Giardia DHC Beta - Intron 1 (joining ‘exons’ 2+3)

>*G. lamblia* – P15 *DHC β* Intron 1

Exon 2 (contig30:114200-114765)

GTGTCTTCTCGACACACCCGGAAAGCTCCCAAAGGACAGAAGCTGGAAAGCAGCCAAAAATGTTATGGGCTC
TATTGACACGTTCCTGAATAGGCTACAAAACTATGATAAGGATAATATTCATGAGGTCAATTTTGCGGCGGC
AAAGAAATACACGTCTGATCCTAATTTCACGGGCGAATTCATCCGCTCCAAGTCGGTCGCTGCCGCTGGTAT
TTGCGAGTGGGCGCGGAACATTGTTTCCTACAACGAAATCTACAAAATTGTTTTGCCGCTACGTGAGGCCGC
CGCAGAGGCCGCAATGCAGCTCGAAGCTGCCCGTAAGAAGTACAAAGCTGTCATGGACAACGTTGCAAAGCT
GAACAAGAAGCTTCAAACTCTCAAGGATCAATTTGACAAGGgtatgttactaggtgaaacgctacttatgtg
tgtatgtctatatgtctcgcgctcgggcgctcctttactcaattatcaacggattcgtttacagaccagcga
gaagtcaacccgatgatagggtggaattcaagcggctgctat

Exon 3 (contig39:69453-70092)

cccggcccgtcacggcccccacgtccgccgtttcgatagtctgtacccattcaaataattgcggtgttttt
gcagttttttggagtaaaagtaaaatccaaaaaataagtagcgtttcaccgacagaaacctttccagaaga
tactaacacacagCCACCGAAGACAAGGAAGAAGTCATCCGGCAGGCAAATGCCACGAAAGCACGTCTAGAT
CTTGCTAATCGACTCGTTAACGGTCTTGCGTCTGAAAAGGAACGTTGGAAGCAGTCGGTCACCGATCTTCAG
AGTCGCGACGGAACTCTGGTTGGAGACGTTCTTATAACAGCAGCCTTTATATCATATTCTGGGCCGTTTAAT
CGCCAATTCAGAACGGAGCTTCTCTCTAAGTGGATCAGCAGGGCGAAGGAACTTAAGATTCCCATGCAGGAA
AACTTTGATCCGTTGCAGCTTTTAACCAATGATGCGCTGATTGCGCGCTGGAACAATGACAAGCTTCCTACC
GATCGAGTTTCACTCGAGAATGCTTCCATCTTTTCAACCGCAGAGCGTTGGCCACTTATCATCGACCCTCAG
CTGCAGGGGATTGCGTGGATAAAGGCCAGAGAGGAGCGCCGAAAAGAAGAAGAGCGCCGCTTCA

>*G. lamblia* – GS *DHC β* Intron 1

Exon 2 (ACGJ01002314:963-1456)

GGAAAGCTTCCAAAGGATAGAAGCTGGAAGGCAGCCAAGAATGTTATGGGTTCCATTGATACCTTCTTAAAT
AGACTTCAGAACTATGACAAGGACAATATTCACGACGTCAATTTTGCGGCAGCCAAGAAGTACACGTCTGAT
CCTAATTTCACAGGTGAGTTTATCCGCTCTAAATCGGTTGCTGCTGCCGGTATCTGTGAGTGGGCAAGAAAC
ATCGTTTCATACAATGAAATCTATAAGATTGTCCTACCATTACGTGAGGCCGCCGCAGAGGCCGCAACGCAG
CTCGAAGCTGCCCGTAAGAAGTACAAAGCCGTCATGGACAACGTTTCGAAGCTCAACAAGAAGCTCCAAACT
CTCAAAGACCAATTCGACAAGGgtatgttaccggtgaaacgctacttatgtgtgtatgtctatatgttcttc
gcgcttaggcgctcctttactcaaaacaccaacagattggttcactaattgctggagacaaagccgtcaaga
atgcagaattcagacagctgttat

Exon 3 (ACGJ01002906:1897-2394)

cccggcccatcacggcctccgcgcccgctgtttcggtagtctgtgcccattcaaataattacggtatttttt
gaagcttttggagcaaaagcgaaatccaaaaaatagagtagcgtttcaccgacagaagcttttccagacaa
tactaacacacagCTACTGAGGACAAAGAAGAAGTTATTCGGCAGGCAAATGCCACCAAAGCTCGTTTGGAT
CTTGCTAATAGACTCGTTAACGGTCTTGCATCCGAAAAGGAGCGCTGGAAGCAATCTGTTACCGATCTCCAG
AGCCGCGATGGGACTCTGGTCGGGGACGTCCTTATAACAGCCGCCTTTATATCGTATTCTGGGCCATTCAAC
CGTCAGTTTAGAACCGAGCTCCTCTCCAAGTGGATCAGTAGAGCAAAGGAACTGAAGATACCCATGCAGGAG
AACTTCGATCCACTACAGCTTTTAACCAATGATGCTCTAATTGCTCGTTGGAATAATGACAAACTT

C) Giardia DHC Beta - Intron 2

>*G. lamblia* – P15 *DHC β* Intron 2

Exon 3 (contig39:71655-72150)

GCGCAGAAAGGTGGTTGGGTACTCCTCCAGAACATTCACCTTATGAAGATTTGGCAGGTCAAGTTGGAAAAG
ATGATGGAGCAGTACTGTTCTGAAACAGCACACGACAACTTCCGTCTCTTCCTCTCGGGAGAACCGGATTCC
GACCCTGCTGTAGCGTCTGTTCTCCCTGGAATTGTGCAGATGTGTATTAAGGTGACCAACGAGCCCCCACGA
GGAATAAAAGCCAACATGAACAGGGCTATCGGACTATTTACACCAGATACGTTTGAAATGTGCTCTAAAGGT
AACGAGTTCAAGAGTATTCTGTTTGCTCTGATTTTCTTCCATGCAATTGTCATCGAGAGACGAAAGTTCGGT
CCTATAGGgtatgttcgtaatctgtgtagttgcagtatgccattgttttataacgtgtatgtcactatgtca
gtatgtcagtacgccagtacaagtaattttcctttactcaaatattttataactgctccattct

Exon 4 (contig137:29037-29533)
ccatcacagttttggattagaaagcctcggaacgcagatctacagcaagggatgcgaagaaggaaagatag
catttaaactacaaattcctaatgacagcaaaaatgacaacgaaaatttaaaatggcatactgttactcaca
gctttactgacctgatacatactaacacacagTTGGAACGTCCCCTACAACTTCAACAATGGCGACCTGACC
ATATGCAAGGATGTCTTGTTCAACTATCTGGAAAATAACACTAAAATCCCCTGGGACGACCTCAAATACATG
TTTTGCGATATCTTTTACGGAGGACACGTCGGCGATGATCTGGATCGCCGGCTCATGCGCTCATTTATGGAC
TCTTTAATGTCGGATGCTCTCTTTGAAGATGGGAAATTCTTCGCCCCGGATTTCCCTGTGCCGTCCCCGATG
AGCTATGACGGTTACAAAGCGTACATTGCTGAAGCACTTCCTGAGGAATCACCGAAGATGTATGG

>*G. lamblia* – GS *DHC β* Intron 2
Exon 3 (ACGJ01002906:4170-4540)
GATGATGGAGCAGTACTGCTCTGAGACAGCACATGACAACTTCCGTCTCTTTCTATCTGGAGAACCAGACTC
TGATCCTGCTGTTGCATCTGTCCTTCCTGGGATTGTGCAGATGTGTATTAAGGTGACCAACGAGCCTCCACG
AGGAATAAAGGCTAATATGAACAGGGCCATCGGGCTGTTTACGCCAGACACATTTGAAATGTGCTCAAAAGG
TAACGAGTTTAAGAGCATTCTGTTTGCTTTAATATTCTTCCATGCAATCGTCATCGAGAGACGAAAATTCGG
TCCCATAGGgtatgttcacgctgtgtagtgcagtatgccatttttaaatatgttagtatgttaatatgtcag
tacgctagtac

Exon 4 (ACGJ01002298:2173-2668)
acttaaaatattgagaatcaggggcggcaaaaatttaaaaaatgaaatttaaaatggcatactgctactcac
agctttactgaccgaatgcatactaacacacagCTGGAACGTCCCCTACAATTTCAACAATGGTGACCTGAC
CATATGCAAAGATGTTCTGTTCAACTATCTGGAGAACAACACCAAGATCCCCTGGGATGACCTCAAATACAT
GTTTTGCGATATTTTCTATGGAGGGCACGTTGGTGATGACTTGGATCGCCGGCTTATGCGCTCATTTATGGA
CTCTCTGATGTCGGACGCCCTATTTGAGGACGGGAAGTTTTTCGCCCCGGACTTCCCCGTTCCGTCTCCGAT
GAGCTATGATGGCTATAAGGCATATATTGCTGAAGCGCTCCCCGAGGAGTCACCGAAGATGTACGGTTTACA
CCCCAACGCAGAGATCATGTTCCTCACTACCCAGTCAAACACGCTCTTCATGATGCTGATGCAG

D) Giardia DHC Gamma

>*G. lamblia* – P15 *DHC γ* Intron
Exon 3 (contig38:36625-37189)
AACCCTTGCCCAGGCCATCGGAATGTACCTAGGTGGTGCTCCCGCAGGCCCTGCCGGGACGGGCAAGACTGA
GACTGTCAAGGATCTTGGCAAGACCCTTGGAATGTACGTCGTTGTCTTCAACTGCTCTGATCAGATGGACTA
TAAGGGCCTCGGCAAGATCTATCGCGGGATTGCCCAAACCGGATCGTTCGGTGACTTTGACGAGTTCAACCG
AATCGACCTTCCAGTTCTCTCTGTCTCTGCTCAGCAGATCCAGTGTGTCCTCTCCGCTGTCAAGGAGCGGAA
GAAGACGTTTCTCTACACAGATGGCTGTGTTATCACTCTCATCCCCTCTTGCGGTATCTTCATCACGATGAA
CCCCGGTTACGCCGGGCGTCAGGAACTTCCAGAGAATCTCAAAGCCTTATTCCGTAGCGTTGCAATGatatg
tttacaggtggttcggtgtgtatgcttggcgtgtatgtgtgtatgtccttcctttactcaatgcttgggccg
aaaatgaaaaccggcctgggccaaatggatctctacgtggtccaggagctctttatgtcca

Exon 4 (contig5:23495-24062)
gtaatcattgcgggaggagggcagaacgccatctgtttttgtaaattcctggtaaaaacaaattctcaataa
aaaacgccaagctttacccccgatgccgaggaacacacacccgggccacctctccctacagtcagctaaca
cacagcacATGGTACCAGATCGCGAGATCATCATCCGAGTCAAGCTCGCCTCCGCGGGCTTCATGAAGAGCA
GCATGCTAGCGAAGAAGTTCTTTATTCTCTACCAACTCTGCGAGGAACAGCTTTCCAAGCAACGCCACTATG
ACTTTGGACTTAGAAACATCCTCTCCGTCCTGCGTATTTGCGGGTCGCGTCGCCGCAGCAACCCAGACCTGT
CTGAAGAGAACATACTTTTGCGCGTCCTCACAGACATGAACCGCTCGAAGTTAGTTGACGAGGATGCTCCTC
TTTTCATGTCTTTGACCGAAGACCTCTTTCCCGGCCTTCGGGTCGAGGATAACAGTTATCCAGACCTCGACG
CTGCCCTCTCTGTCGTCTGCGGTGAGCTTTTTCTTCAGCAGCACCCTGACTGGCTCAAGGCCGT

>*G. lamblia* – GS *DHC γ* Intron
Exon 3 (ACGJ01002918: 40640-41203)
Aaccctcgcccaggctattgggatgtacttgggcggggcccctgcaggtcctgcgggaacaggcaagaccga
gacggttaaggatcttggcaagaccctcggcatgtacgtcgtcgtcttcaactgctctgaccagatggacta
caagggtcttggtaagatttatcgtggaattgcccagacagggtcgttcggtgactttgacgagttcaaccg
gatcgatcttccagttctttccgtttctgcgcagcagatccagtgcgtcctttctgccgtcaaggagcggaa
gaagacgtttctctatacagatggctgtgtcattactctcataccctcttgcggtatcttcatcacgatgaa

ccccggttacgccgggcgtcaagaacttccagagaatctcaaagctttattccgtagtgttgcaatgatatg
tttacaggtggttcggtgtgtatgcttggcgtgtatgtgtgtatgtccttcctttactcaatgcccaggcca
aaaatgaaatctggcttggttcgtgatggatctctatgtggtccaggagctgttcacatc

Exon 4 (ACGJ01002422:30156-30724)
gtaatcattgcggggaaggaggcagaacgccgtccgtttttgtaaattcttgctgaaattagattctcaata
aaaaacgccaagctttacccccgatgccgagggacacacaccccggaccacctctccctacagtcagctaac
acacagcacATGGTACCAGATCGTGAAATCATCATCCGAGTTAAGCTCGCCTCCGCGGGCTTCATGAAGAGC
AGCATGCTAGCAAAGAAGTTCTTTATCCTCTATCAGCTCTGCGAGGAACAACTCTCCAAGCAGCGCCACTAT
GACTTTGGGCTCAGAAACATCCTCTCGGTCCTCCGTATATGCGGATCGCGTCGCCGCAGCAACCCCGATCTC
TCTGAAGAGAACATCCTTTTGCGGGTCCTCACAGACATGAACCGCTCGAAGCTGGTCGATGAGGATGCTCCT
CTCTTTATGTCTCTAACAGAAGACCTTTTCCCCGGCCTCCGAGTCGAGGATAATAGCTATCCAGACCTTGAC
GCCGCTCTCTCCGTCGTCTGCGGCGAGCTCTTCCTCCAGCAACACCCCGACTGGCTCAAGGCTGT

E) *Rpl7A Cis*-Intron

>*G. lamblia* – WB *Rpl7a* Intron (GL50803_17244)
ATGTCCAAGGTTTCTGGCAGCGACATTAAGAGGGCCCTCGCCGTACCCGAGAACAAGAGCCGCAGCAAGTGC
GACTTCGACCTGACTCCGTTCGTCAGGTGGCCTCGCCAGGTCCGCATCCAGAGACAGAAGGCAGTCCTCCAG
AGGCGCCTCAAGGTTCCCCCAACTGTCAATCAGTTCATGAATCCGATCTCGAGGAACCTCACAAACGAGATT
TTCAACCTTGCTCGCAAGTACTCCCCCGAATCGAAGGAGGAGCACAAGGCGCGCCTGCTCCAGATCGCCGAC
GCAAAGGCCAACGGGAAGCCCCTTCCGGAGAAGTCTGACAAGCTTGTCATCGCGTCTGGTATCAGACGCATA
ACATCCCTCGTCGAGAGCAAGCGCGCGAAGCTTGTCCTGATTGCAAATGATGTCGACCCCCTTGAAgtatgt
tcttatgcgcgaggagccgtccgctgaccgcacacacctctgattgcgggttgtgtgttgtcagcgggtgga
cttcgctgttcacctgacaactgacccacagCTCGTACTTTGGCTTCCCACACTCTGTCACAAGATGGGCGT
CCCGTACGCCATCGTTCGCACTAAGGGCGATCTGGGCAAGCTCGTCCATCTGAAGAAGACGACCAGCGTCTG
CTTCACCGACGTGAACCCAGAGGACAAGCCCACCTTTGATAAGATCCTCGCGGCCGTGGCCCATGAAGTTGA
TTATGCAAAGGCCATGAAGACGTACGGAGGCGGCGTTCGCCGCGAGGATGAGGCCCAGCAGATGTAA

>*G. lamblia* – P15 *Rpl7a* Intron (GLP15_934)
ATGTCCAAGGTTTCTGGCAGCGACATCAAGAGGGCCCTCGCCGTACCCGAAAACAAAAGCCGCAGCAAATGC
GACTTCGATCTGACCCCATTTGTCAGATGGCCTCGCCAGGTTCGTATCCAGAGACAGAAGGCAGTCCTCCAG
AGGCGCCTCAAGGTTCCCCCTACTGTCAATCAGTTCATGAATCCAATCTCGAGGAACCTCACAAATGAGATC
TTCAACCTCGCTCGTAAGTACTCCCCCGAGTCGAAGGAGGAGCACAAGGCCCGTCTGCTCCAGATCGCCGAC
GCAAAGGCCAACGGCAAGCCCCTCCCGGAGAAGTCTAACAAGCTTGTCATCGCATCCGGCATCAGGCGCATA
ACGTCTCTTGTCGAGAGCAAACGCGCGAAGCTTGTCCTAATTGCAAATGACGTCGATCCCCTTGAAgtatgt
tcttatgcgcgaggaaccatccgctgaccgcacgtgcctctaacagcgggctgtgtgttgtcggcgggtggg
cttcgctgttcacctgacaactgacccacagCTCGTACTTTGGCTTCCCACGCTCTGTCACAAGATGGGCGT
CCCGTACGCCATCGTTCGCACTAAGGGCGATCTAGGCAAGCTCGTTCATTTGAAAAAGACAACTAGCGTCTG
CTTCACCGATGTGAACCCGGAAGACAAGCCCACCTTCGATAAGATTCTCGCGGCCGTGGCTCACGAAGTTGA
TTATGCAAAGGCTATGAAGACGTACGGAGGCGGTGTCCGCCGCGAGGACGAGGCCCAGTAA

>*G. lamblia* – GS *Rpl7a* Intron (GL50581_195)
ATGTCCAAGGTTTCTGGCAGCGATATRAAGAGGGCCCTTGCTGTACCCGAGAATCAGCGCCGCAGCAAGTGC
GACTTTGACCTGACCCCGTTCGTCAGATGGCCGCGCCAGGTTCGCGTCCAGAGACAGAAGGCAGTCCTGCAG
AGGCGTCTCAAGGTTCCCCCTACCGTCAATCAGTTCATGAACCCGATCTCGAGGAACCTTACGAATGAAATA
TTCAATCTCGCTCGTAAGTACTCCCCTGAGTCGAAGGAAGAGCACAAGGCGCGCTTGCTCCAAATCGCTGAC
GCAAAGGCCAACGGGAAGCCTCTCCCAGAGAAGTCCAACAAGCTCGTCATCGCATCTGGCATTAGGCGCATA
ACGTCCCTTGTCGAGAGCAAACGTGCGAAGCTTGTTCTTATTGCAAACGACGTAGATCCCCTTGAAgtatgt
tcctatgcgcgaggagccatccgctgaccgcacacgttactcgtccacgaatcgcgtgctgtcagcgggtgg
ccttcgctgtccacctgacaactgacccacagCTCGTACTTTGGCTTCCCACACTCTGCCACAAGATGAACG
TTCCGTATGCTATTGTCCGCACCAAGGGAGACCTGGGCAAGCTCGTCCACTTGAAGAAGACGACTAGTGTTT
GCTTCACCGACGTGAACCCAGAAGACAAGCCAACCTTTGACAAGATCCTCGCAGCAGTGGCCCAGGAGGTCG
ACTATGCAAAGGCCATGAAGACGTATGGAGGCGGCGTTCGCCGTGAGGACGAGTCCCAGTAA

F) Unassigned ORF on contig [AACB01000025]

>*G. lamblia* – WB Unassigned ORF Intron (GL50803_35332)

ATGACTTTCTTTGACACGAGTAACGGTGTTTTCCTTCCAGGCGAATTTGTCGTTATCCATTACTCGAACAAA
GAACTCCATTTAGGAAGAgtatgtttgtagctcggcggcactatacttcaagattactggaaactagcccag
cggatcgaaggtagaacaatttcctctcctatcacgctctacgaaactgccaaaagggtacgcattcctgcc
aactattcaacttcttacctcttttggctttctattaacgggcttttagacgagggattgaccgccgagcat
ttaccatccaactgacacacagATAATTCCAATAGGCAATATACCTAAGGAAGCCATTACGGGTCAACGTTC
GTTGGACAAATTAACATATTTTGTCTATATACTTCGAACCAAAGAGTCTCTAGCTGTACCTTACTTTAAGAT
ACAGCATCTCAGCACTCCGCTACTAATTCAGTTATCGCAGAATCCAGCCACTAAGGATCTTCTTCGTGACAT
TTGTGCTCATATTTTCGATATTGTTCTTGTTGGCTC

>*G. lamblia* – P15 Unassigned ORF Intron (contig16:7495-8034)

ATGACTTTCTTTGACACGAGTAACGGTGTTTTCCTTCCAGGCGAATTTGTCGTTATCCATTACTCGAACAAA
GAACTACATTTAGGAAGAgtatgtttgtagctcggtggcactatacttcaagattactggaaactagcccag
tggatcgaaggtagaacaatttcctcttctattgcgctctacaagaccgtcaaaagagtacacattcctgcc
aactattcaacttcttacttcttttgacttccttttaatgggcttttagacgagggattgaccgccgagcgc
ttaccattcaactgacacacagATAATTCCACTAGGCAATATACCTAAGGAAGCCATTACTGGTCAACGTTC
ATTGGACAAATTAACATATTTTGTTTATATACTTCGAACCAAAGAATCTCTAGCTATACCTTACTTTAAAAT
ACAGCATCTTAGTGCTCCGCTACTAATTCAGTTATCACAGAATACAACTAATAGGGATCTTCTTCGTGATAT
TTGCACTCATATTTTTGATATCGTTCTCGTTGGTTC

>*G. lamblia* – GS Unassigned ORF Intron (ACGJ01000100:677-1215)

ATGACTTTCTTTAACACGAGTAACGGCGTTTTCCTCCCAGGCGAGTTTGTCATTGTTCAATACACAGATAAG
GAACTGCATTTAGGAAGAgtatgtttgtagctcggtggcactatactttgagtttactggaaattagcccag
tggaccgaaggtaggacaatctccttttctatcacgctctgcgggactgtcaaaaaagcatacccccatcac
gactgttcgatttcttgcttctttgacttctgcttaacgggcgtttaggcaagagattgaccaccgagtatt
taccatcaaactgacacacagATAATACCAATGGAAAAGATCCCCAAGGAGGCCATTACGGCCCAGCGCTCC
CTCGATAAGTTGATGTATTTTGTATATATACTTAGAACCAAGGAGTCTCTAGCTGTGCCTTACTTCAAAATA
CAGTGCCTTAGCGCTCCATTATTAGTCCAGCTGTCGAGGAATGCAACTAACAGGGACCTTCTTCGCGACATC
TGTAGCCACATTTTCGACATCGTTCTTGTTGGGTC

***Note: The next two sequences are predicted gene duplicates of the sequence above but have A->G mutations at the branch-point A and may be "pseudo-introns"**

>*G. lamblia* – GS Unassigned ORF Intron (ACGJ01001160:1-521)

AGTAACGGCGTTTTCCTCCAAGGCGAGCTTGTCATTGTTCAATACACAGATAAGGAACTGCATTTAGGAAGA
gtatgtttgtagctcggtggcactatgctttgagtttactggaaattagcccagtggaccgaggacaggaca
atttccttttctatcacgctctgcgggactgtcraaaaagcatacccccatcacgactgttcgatttcttgc
ttctttgacttctgcttaacgggcgtttaggcaagagattgaccaccgagtacttaccatcaacctggcaca
cagATAATACCAATGGAAAGGATCCCCAAGGAGGCCATTACGGCCCAGCGCTCCCTCGATAAGTTGATGTAT
TTTGTATATATACTTAGAAYCAAGGAGTCTCTAGCTGTGCCTTACTTCAAAATACAGTGCCTTAGCGCTCCA
TTATTAGTCCAGCTGTCGAGGAATGCAACTAACAGGGACCTTCTTCGCGACATCTGTAGCCACATTTTCGAC
ATCGTTCTTGTTGGGTC

>*G. lamblia* – GS Unassigned ORF Intron (ACGJ01001154:186-667) (RC)

CAATACACAGATAAGGAACTGCATTTAGGAAGAgtatgtttgtagctcggtggcactatactttgagtttac
tggaaattagcccagtggaccgaggacaggacaatttccttttctatcacgctctgcgggactgtcaaaaaa
gcatacccccatcacgactgttcgatttcttgcttctttgacttctgcttaacgggcgtttaggcaagagat
tgaccaccgagtacttaccatcaacctggcacacagATAATACCAATGGAAAGGATCCCCAAGGAGGCCATT
ACGGCCCAGCGCTCCCTCGATAAGTTGATGTATTTTGTATATATACTTAGAACCAAGGAGTCTCTAGCTGTG
CCTTACTTCAAAATACAGTGCCTTAGCGCTCCATTATTAACCCAGCTGTCGAGGAATGCAACTAACAGGGAC
CTTCTTCGCGACATCTGTAGCCACATTTTCGACATCGTTCTTGTTGGGTC

G) Ferredoxin *cis*-intron

>*G. lamblia* – WB Ferredoxin Intron (GL50803_27266)

ATGTCTCTACTATCGTCAATAActatgttgagaaccacccaaacaactaacacacagGGCGCTTCATAACGT
TCCGAGTGGTCCAACAGGGCGTGGAACACACAGTTTCAGGTGCTGTCGGCCAGAGCTTACTAGATGCCATCA
AGGCTGCGCATATCCCCATTCAGGACGCGTGCGAAGGACACCTTGGCTGTGGTACCTGCGGTGTTTATTTGG
ACAAAAAGACGTACAAGCGTATTCCGCGAGCGACAAAGGAAGAAGCGGTTCTCCTAGATCAGGTACCCAACC
CCAAGCCCACATCGCGGCTTTCGTGTGCAGTAAAACTCAGTAGCATGCTGGAGGGAGCGACAGTACGCATAC
CCTCCTTTAACAAGAACGTCCTTAGCGAAAGTGACATTCTTGCAAGCGAAGAGAAGAAAAGGCACGGGCAGC
ACTGA

>*G. lamblia* – P15 Ferredoxin Intron (GLP15_2567)

ATGTCTCTACTATCGTCAATAActatgttaaataccactcaaacaactaacacacagGGCGCTTCATAACAT
TCCGAGTGGTTCAGCAGGGTGTAGAACACACAGTCTCAGGTGCTGTTGGCCAGAGCTTACTGGATGCCATCA
AGGCTGCACGTATCCCCATTCAGGATGCGTGCGAAGGACACCTTGGCTGTGGTACCTGTGGTGTTTACCTGG
ATAAAAAGACGTACAAACGTATTCCGCGAGCGACAAAGGAGGAAGCGATTCTCCTAGACCAGGTACCTAATC
CCAAGCCCACATCACGGCTTTCCTGTGCAGTAAAGCTCAGTAATATGCTGGAAGGAGCGACAGTATGTATAC
CATCCTTCAACAAGAATGTCCTTAGTGAAAGTGACATTCTTGCAAACGAAGAGAAAAAAAGGCACGGGCAGC
ACTAA

>*G. lamblia* – GS Ferredoxin Intron (GL50581_3971)

ATGTCACTAATATCGTCAATAActatgtatgggcaccatctgaacaactaacacgcagGGCGTTTCATAACA
TTTCGAGTAGTCCAGCAGGGTATAGAACACACTGTGTCAGGCGCTGCTGGCCAGAGTTTGTTGGACGCCATC
AAGGCGGCACACATCCCTATTCAAGACGCATGTGAGGGTCATCTTGGTTGTGGAACATGCGGTGTGTATCTG
GATAAGAAAACGTACAAGCGTATTCCGCGTGCGACGAAGGCAGAGGAGACCCTCCTGGATCAAGTTCCCAAT
CCTAAGCCTACGTCACGGCTTTCGTGTGCAGTGAAGCTCAGCACTATGCTCGAAGGTGCAACCGTGCGTATA
CCTCCTTTTAATAAGAACGTCCTTAGTGAAAGCGATATCCTTGCAAATGAGGAAAAGAAGAAGTGCGGACAG
CGTTAA

H) Dynein-like *cis*-intron

>*G. lamblia* – WB Dynein-like Intron (GL50803_15124)

ATGgtatgttatctcccgcataacctaacacacagAATGACGTAGAGGCTACAATAAAGAGAATTAGCACCC
ATCCGGGTGTGCAGGGTCTTCTGGTCCTCATGGACGACGGCACAGTTATCCGGAGCACATTCGATGACGAAA
TGACTGCCAGATACGTGAAACTTGTCTCATCCTATACTGTATTGTCTCGATCTGCTATACGCGACATAGACC
CCACAAACGAGCTTAATTACGTCCGTATCAGGTCTGACAAGCGTGAAATTATTTGTATCCCCGAAGGCAAGT
ACTTCATCATCGCCGTGACGTCCATGGACGCCGAGTTCAAGCCCTAG

>*G. lamblia* – P15 Dynein-like Intron (GLP15_676)

ATGgtatgttatttcctgcttaacctaacacacagAATGACGTAGAGGCTACAATAAAGAGAATTAGCACCC
ATCCAGGTGTGCAGGGTCTTCTGGTCCTCATGGACGACGGCACAGTTATCCGGAGCACATTCGATGACGAAA
TGACTGCCAGATACGTAAAGCTTGTTTCATCTTATACCGTACTATCCCGATCTGCTATACGTGACATAGATC
CCACGAATGAACTCAATTACGTCCGTATTAGATCTGACAAGCGCGAAATTATTTGTATCCCCGAAGGCAAGT
ACTTCATAATTGCTGTAACGTCTATGGATGCTGAGTTCAAGCCCTAG

>*G. lamblia* – GS Dynein-like Intron (GL50581_2800)

ATGgtatgttatcttccatttaacctaacacgcagAATGACGTAGAGGCAACAATAAAGAGAATTAGTACCC
ACCCAGGTGTACAAGGTCTCCTAGTTCTCATGGACGATGGCACAGTCATCCGGAGCACATTCGACGACGAAA
TGACCGCTAGATACGTAAAATTGGTCTCCTCCTACACCGTTCTGTCCCGATCCGCCATACGTGATATAGATC
CTACGAATGAGCTCAACTATGTCCGTATTAGATCAGACAAACGTGAAATTATTTGCATTCCCGAGGGTAAGT
ACTTCATAATTGCCGTAACATCCATGGACGCCGAGTTTAAGCCCTAG

Figure A.1.7. Comparison of *cis*- and *trans*-spliced introns from available *Giardia* genome assemblies.

ORFs containing *cis*- and *trans*-spliced introns from the *Giardia* WB isolate assembly (used for our original data) were compared with the more recently released GS and P15 isolate genome assemblies. For *trans*-spliced introns (**A-D**), four items are shown: (i) alignment of three isolates for the 5′ and 3′ intron boundaries, from the splice site to the region corresponding to the complementarity for the secondary structure in reference (WB) isolate; (ii) intron basepairing in the complementary region for all three isolates (differences from WB are indicated by grey boxes; (iii) alignment of three isolates in the basepairing region of the 5′ molecule; and (iv) alignment of three isolates basepairing region of the 3′ molecule. For long introns (**E,F**), intron alignment and basepairing are shown. For short introns (**G,H**), only an intron alignment is shown. Central regions of complementarity of intron halves are shown in red text.

A) Hsp90

Intron Boundary Alignment

```
WB_HSP90    5′-GTATGTT-ATGTT...   ...CGACTTCCCTATAACTAACACGCAG-3′ (DIST = 36NT)
P15_HSP90   5′-GTATGTT-ATGTT...   ...CGACTTCCTTCTAACTAACACGCAG-3′ (DIST = 36NT)
GS_HSP90    5′-GTATGTTTATGTA...     ...CGTTCCTCACAACTAACACACAG-3′ (DIST = 35NT)
               ******* ****            ****    ********* ***
```

Intron Basepairing

```
WB_HSP90    5′-TGTATGCTGTATGTGTGCGAGA-CTCC-3′
               |||||.|||||||||||||||| ||||
            3′-ACATATGACATACACACGCTCTTGAGG-5′

P15_HSP90   5′-TGTATGCTGTATGTGTGCGAGA-CTCC-3′
               |||||.|||||||||||||||| ||||
            3′-ACATATGACATACACACGCTCTTGAGG-5′

GS_HSP90    5′-TGTGTGTTGTATGTGTGCGAGA-CTCC-3′
               |||.||.||.|||||||||||  |||
            3′-ACATACGACGTACACACGCTCTT-AGG-5′
```

5′ Basepairing Region Alignment

```
WB_HSP90    GTATGTT-ATGTTTGTATGCTGTATGTGTGCGAGACTCCTTTACTCAAATTTGCGGCGTC
P15_HSP90   GTATGTT-ATGTTTGTATGCTGTATGTGTGCGAGACTCCTTTACTCAAATTCGCGGCATT
GS_HSP90    GTATGTTTATGTATGTGTGTTGTATGTGTGCGAGACTCCTTTACTCAA-TTCGTGGCTTC
            ******* **** *** ** ***************************** ** * *** *
```

3′ Basepairing Region Alignment

```
WB_HSP90    --CGAGGCGGAGTTCTCGCACACATACAGTATACACGACTTCCCTATAACTAACACGCAG
P15_HSP90   --CTGGGCGGAGTTCTCGCACACATACAGTATACACGACTTCCTTCTAACTAACACGCAG
GS_HSP90    GTCAGGGTAGGATTCTCGCACACATGCAGCATACACG--TTCCTCACAACTAACACACAG
              *  **  *  ************ *** *******  ****    ********* ***
```

B) DHC Beta - Intron 1 (joining ‘exons’ 2+3)

Intron Boundary Alignment

```
WB_DHCB1    5′-GTATGTTACTG...  ...GACAGAAGCCTTTCCAGACAATACTAACACACAG-3′ (D=45NT)
P15_DHCB1   5′-GTATGTTACTA...  ...GACAGAAACCTTTCCAGAAGATACTAACACACAG—3′ (D=45NT)
GS_DHCB1    5′-GTATGTTAC...    ...GACAGAAGCTTTTCCAGACAATACTAACACACAG-3′ (D=43NT)
               *********          ******* * ********  **************
```

Intron Basepairing

```
WB_DHCB1    5′-GGTGAAACGCTACTTAT-3′
               ||||||||||||||||||
            3′-CCACTTTGCGATGAATA-5′

P15_DHCB1   5′-GGTGAAACGCTACTTAT-3′
               ||||||||||||||||||
            3′-CCACTTTGCGATGAATA-5′

GS_DHCB1    5′-GGTGAAACGCTACT-TAT-3′
               |||||||||||||| |||
            3′-CCACTTTGCGATGAGATA-5′
```

5′ Basepairing Region Alignment

```
WB_DHCB1    GTATGTTACTGGGTGAAACGCTACTTATGTATGTATGCTTATATGTCTTC
P15_DHCB1   GTATGTTACTAGGTGAAACGCTACTTATGTGTGTATGTCTATATGTCTCC
GS_DHCB1    GTATGTTACC-GGTGAAACGCTACTTATGTGTGTATGTCTATATGTTCTT
            *********  ****************** ******  *******
```

3′ Basepairing Region Alignment

```
WB_DHCB1    -----------CAAAAAATA-AGTAGCGTTTCACCGACAGAAGCCTTTCC
P15_DHCB1   AAAGTAAAATCCAAAAAATA-AGTAGCGTTTCACCGACAGAAACCTTTCC
GS_DHCB1    -AAGCGAAATCCAAAAAATAGAGTAGCGTTTCACCGACAGAAGCTTTTCC
                     ********* ********************** * *****
```

C) DHC Beta - Intron 2

Intron Boundary Alignment

```
WB_DHCB2    5′-GTATGTTTGTAAT...  ...CTTTACTGACCAGACACATACTAACACACAG-3′ (D=45NT)
P15_DHCB2   5′-GTATGTTCGTAAT...  ...CTTTACTGACCTGATACATACTAACACACAG—3′ (D=45NT)
GS_DHCB2    5′-GTATGTTCACG...     ..CTTTACTGACCGAATGCATACTAACACACAG-3′ (D=43NT)
               *******              ***********  *  ***************
```

Intron Basepairing

```
WB_DHCB2    5′-CTGTGTAGTCGCAGTATGCCATT-3′
               ||||| ||| ||||||||||||
            3′-GACAC-TCATCGTCATACGGTAA-5′

P15_DHCB2   5′-CTGTGTAGTTGCAGTATGCCATT-3′
               ||||| ||| .||||||||||||
            3′-GACAC-TCATTGTCATACGGTAA-5′

GS_DHCB2    5′-CTGTGTAGT-GCAGTATGCCATT-3′
               ||||| ||| ||||||||||||
            3′-GACAC-TCATCGTCATACGGTAA-5′
```

5′ Basepairing Region Alignment

```
WB_DHCB2    GTATGTTTGTAATCTGTGTAGTCGCAGTATGCCATTATTTTATAACGT-G
P15_DHCB2   GTATGTTCGTAATCTGTGTAGTTGCAGTATGCCATTGTTTTATAACGT-G
GS_DHCB2    GTATGTTCAC--GCTGTGTAGT-GCAGTATGCCATTTTTAAATATGTTAG
            *******      ********* ************* **  ***   * *
```

3′ Basepairing Region Alignment

```
WB_DHCB2    ----ATTTAAAATGGCATACTGCTACTCACAGCTTTACTGACCAGACACA
P15_DHCB2   GAAAATTTAAAATGGCATACTGTTACTCACAGCTTTACTGACCTGATACA
GS_DHCB2    TGAAATTTAAAATGGCATACTGCTACTCACAGCTTTACTGACCGAATGCA
                ****************** ********************  *  **
```

D) DHC Gamma

Intron Boundary Alignment

```
WB_DHCG     5′-ATATGTTCAC...  ...CTCCCTACAGCCAGCTAACACACAGCAC-3′ (DIST = 38NT)
P15_DHCG    5′-ATATGTTTAC...  ...CTCCCTACAGTCAGCTAACACACAGCAC—3′ (DIST = 38NT)
GS_DHCG     5′-ATATGTTTAC...  ...CTCCCTACAGTCAGCTAACACACAGCAC-3′ (DIST = 38NT)
               ******* **        ********** *****************
```

Intron Basepairing

```
WB_DHCG     5′-AGGTGGTTT--GGTGTGTATG-CTTGGCGT-3′
                ||||||...  ||||||| |  ||.|||.|
            3′-TCCACCGGGCTCCACACACAGGGAGCCGTA-5′

P15_DHCG    5′-AGGTGGTTC--GGTGTGTATG-CTTGGCGT-3′
                ||||||..|  ||||||| |  ||.|||.|
            3′-TCCACCGGGCCCCACACACAAGGAGCCGTA-5′

GS_DHCG     5′-AGGTGGTTC--GGTGTGTATG-CTTGGCGT-3′
                |||||||.|  ||||||| |  ||.|||.|
            3′-TCCACCAGGCCCCACACACAGGGAGCCGTA-5′
```

5′ Basepairing Region Alignment

```
WB_DHCG     ATATGTTCACAGGTGGTTTGGTGTGTATGCTTGGCGTGTATGTGTGTATGTTCCTCCTTT 60
P15_DHCG    ATATGTTTACAGGTGGTTCGGTGTGTATGCTTGGCGTGTATGTGTGTATGTCCTTCCTTT 60
GS_DHCG     ATATGTTTACAGGTGGTTCGGTGTGTATGCTTGGCGTGTATGTGTGTATGTCCTTCCTTT 60
            ******* ********** ********************************* * ******
```

3′ Basepairing Region Alignment

```
WB_DHCG     CCCCGATGCCGAGGGACACACACCTCGGGCCACCTCTCCCTACAGCCAGCTAACACACAG 60
P15_DHCG    CCCCGATGCCGAGGAACACACACCCCGGGCCACCTCTCCCTACAGTCAGCTAACACACAG 60
GS_DHCG     CCCCGATGCCGAGGGACACACACCCCGGACCACCTCTCCCTACAGTCAGCTAACACACAG 60
            ************** ********* *** **************** **************
```

E) *Rpl7A Cis*-Intron

Intron Alignment

```
WB_Rpl7a    GTATGTTCTTATGCGCGAGGAGCCGTCCGCTGACCGCACACACCTCT-GATTGCGGGTTG
P15_Rpl7a   GTATGTTCTTATGCGCGAGGAACCATCCGCTGACCGCACGTGCCTCT-AACAGCGGGCTG
GS_Rpl7a    GTATGTTCCTATGCGCGAGGAGCCATCCGCTGACCGCACACGTTACTCGTCCACGAATCG
            ******** ************ ** **************      **      **    *

WB_Rpl7a    TGTGTTGTCAGCGGGTGGACTTCGCTGTTCACCTGACAACTGACCCACAG (D=39 nt.)
P15_Rpl7a   TGTGTTGTCGGCGGGTGGGCTTCGCTGTTCACCTGACAACTGACCCACAG (D=39 nt.)
GS_Rpl7a    CGTGCTGTCAGCGGGTGGCCTTCGCTGTCCACCTGACAACTGACCCACAG (D=39 nt.)
            *** **** ******** ********* ********************
```

Intron Basepairing

```
WB_Rpl7a    5'-GC-GCGAGGAGCCGTCCGCTGACCGCACACA-CC-3'
               .| ||||.|  ||..|||||||| ||||||| ||
            3'-TGTCGCTTC-AGGTGGGCGACTGTTGTGTGTTGG-5'

P15_Rpl7a   5'-GCGAGGAACCATCCGCTGAC--CGCACGTGCC-3'
               ||||.|  |||.||||.|||  |.|||. |||
            3'-CGCTTC-GGGTGGGCGGCTGTTGTGTGT-CGG-5'

GS_Rpl7a    5'-GCGAGGAGCCATCCGCTGACCGCACACG-3'
               ||||.| ||||.|||||||| |||| ||
            3'-CGCTTC-CGGTGGGCGACTGTCGTGCGC-5'
```

F) Unassigned ORF on contig [AACB01000025]

Intron Alignment

```
WB_ORF      GTATGTTTGTAGCTCGGCGGCACTATACTTCAAGATTACTGGAAACTAGCCCAGCGGATC 60
P15_ORF     GTATGTTTGTAGCTCGGTGGCACTATACTTCAAGATTACTGGAAACTAGCCCAGTGGATC 60
GS_ORF      GTATGTTTGTAGCTCGGTGGCACTATACTTTGAGTTTACTGGAAATTAGCCCAGTGGACC 60
            ***************** ************  ** ********** ******** *** *

WB_ORF      GAAGGTAGAACAATTTCCTCTCCTATCACGCTCTACGAAACTGCCAAAAGGGTACGCATT 120
P15_ORF     GAAGGTAGAACAATTTCCTCTTCTATTGCGCTCTACAAGACCGTCAAAAGAGTACACATT 120
GS_ORF      GAAGGTAGGACAATCTCCTTTTCTATCACGCTCTGCGGGACTGTCAAAAAAGCATACCCC 120
            ******** ***** **** * ****  ****** *   ** * *****  * *  *

WB_ORF      CCTGCCAACTATTCAACTTCTTACCTCTTTTGGCTTTCTATTAACGGGCTTTTAGACGAG  180
P15_ORF     CCTGCCAACTATTCAACTTCTTACTTCTTTTGACTTCCTTTTAATGGGCTTTTAGACGAG  180
GS_ORF      CATCACGACTGTTCGATTTCTTGCTTCTTT-GACTTCTGCTTAACGGGCGTTTAGGCAAG  179
            * *  * *** *** * ***** * ***** * ***    **** **** ***** * **

WB_ORF      GGATTGACCGCCGAGCATTTACCATCCAACTGACACACAG  220 (35 nt.)
P15_ORF     GGATTGACCGCCGAGCGCTTACCATTCAACTGACACACAG  220 (35 nt.)
GS_ORF      AGATTGACCACCGAGTATTTACCATCAAACTGACACACAG  219 (35 nt.)
             ******** *****   *******  *************
```

Intron Basepairing

```
WB_ORF      5'-GCTCGGCGG-3'
               |||||||||
            3'-CGAGCCGCC-5'

P15_ORF     5'-GCTCGGTGG-3'
               ||||||.||
            3'-CGAGCCGCC-5'
```

```
GS_ORF      5′-GCTCGGTGG-3′
               .||||||||
            3′-TGAGCCACC-5′
```

G) Ferredoxin *cis*-intron

Intron Alignment

```
WB_Ferredoxin       CTATGTTGAGAACCACCCAAA-CAACTAACACACAG (36 nt)
P15_Ferredoxin      CTATGTTAAATACCACTCAAA-CAACTAACACACAG (36 nt)
GS_Ferredoxin       CTATGTATGGGCACCATCTGAACAACTAACACGCAG (37 nt)
                    ******       *   *  * ********** ***
```

H) Dynein-like *cis*-intron

Intron Alignment

```
WB_Dyn-like         GTATGTTATCTCCCGCATAACCTAACACACAG (32 nt.)
P15_Dyn-like        GTATGTTATTTCCTGCTTAACCTAACACACAG (32 nt.)
GS_Dyn-like         GTATGTTATCTTCCATTTAACCTAACACGCAG (32 nt.)
                    ********* * *    ********** ***
```

Appendix 2 – Supplementary Figures for Chapter 3:

Evolutionarily divergent spliceosomal snRNAs and a conserved non-coding RNA processing motif in *Giardia lamblia*

Figure A.2.1. Motif sequences are conserved between *Giardia* isolates.
ClustalW alignment of ncRNA genes and *trans*-spliced intron 5′ halves reveals the conservation of ncRNAs and motif sequences within the *Giardia* WB, P15 and GS genomes. Predicted mature RNAs for *G. lamblia* WB isolate are in bold letters with motif sequences highlighted in green. Conserved genomic regions encoding ncRNAs and *trans*-spliced intron 5′ halves are aligned.

Box C/D RNAs

1. GlsR1

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR03	412091	412256	+	166
G. lamblia GS	ACGJ01002434	489	654	+	166
G. lamblia P15	contig30	137430	137594	-	165

```
G. lamblia WB     CAATGTAAATCATATGTTCA--AAAAAGCAAATTAATTCCGCTTCTGATTTCATATAAAT 58
G. lamblia P15    CAATGTAAACCATG-GTTCA--TAAAAGTAAATTAATTCCGTTTTCGATTTTACATAAAT 57
G. lamblia GS     CAATATGGGCTATA--CTCAGGAAAAAATAAATTAATTTCGCTTCCGATTTTATATAAAT 58
                  **** *     **    ***   ****  ********* ** **  ***** * ******

G. lamblia WB     TTCAAGTCCACTGGCCTCTCCTGAGGCAGATGATGACTTTGCGACGGGCGGACGGAGGGA 118
G. lamblia P15    TTCAAGTCCACTGGCCTCTTCTGAGGCAGATGATGACTTTGCGACGGGCGGACGGAGGGA 117
G. lamblia GS     GTCAAATCCACCAGTCTCTCTGGAGACAGATGATGACTTTGCGACGGGCGGACGGAGGGA 118
                   **** *****  * ****   *** **********************************

G. lamblia WB     CGCGTGACGAAGTTTGTCGTATTCTGAATTCCTTCATTTAAAATTGGG 166
G. lamblia P15    CGCGTGACGAAGTTTGTCGTATTCTGAATTCCTTCATTTAAAATTGGG 165
G. lamblia GS     CACGTGACGAATTCTGTCGTATTCTGAATTCCTTCATTTAAAATTGGG 166
                  * ********* * **********************************
```

2. GlsR2

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR05	2883739	2883904	+	166
G. lamblia GS	ACGJ01002924	13357	13522	-	166
G. lamblia P15	contig204	1336	1501	+	166

```
G. lamblia WB     ATGAAACAAAGCTCACGCATACACAGGCTCCGGAAAAATAAATGTAGCGAACCCACGCGC 60
G. lamblia P15    ATGAAACAAAGTTCTAGCGTACACAGGCCCCGGGAAAATAAATGTAGCGAACCCACGCGC 60
G. lamblia GS     ATGAAACGAAGATCAGGTGTGCACAGGTCTCGGGAAAATAAATGTAGCGAACCCACGCGC 60
                  ******* *** **  *  * ******   *** *************************

G. lamblia WB     AAGCGTTGCTACGAGGCGATGGAGACAAAAGCAGTTACGTTCGCAACTCTCTGAGGGTTC 120
G. lamblia P15    AAGCGTTGCTACGAGGCGATGGAGACAAAAGCAGTTACGTTCGCAACTCTCTGAGGGTTC 120
G. lamblia GS     AAGCGTTGCTACGAGGCGATGGAGATAAAAGCAGTTACGTTCGCAACTCTCTGAGGGTTC 120
                  ************************ **********************************

G. lamblia WB     CTGATGCTTCCTTGGATGTCCGAGCCTTCCTTTACTTAATCGACCG 166
G. lamblia P15    CTGATGCTTCCTTGGATGTCCGAGCCTTCCTTTACTCAATCGACCG 166
G. lamblia GS     CTGATGCTTCCTTGGATGTCCGAGCCTTCCTTTACTCAATTGACCG 166
                  ********************************** *** *****
```

3. GlsR4

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR05	2609094	2609259	-	166
G. lamblia GS	ACGJ01001410	70458	70589	+	132
G. lamblia P15	contig19	6558	6725	-	168

```
G. lamblia WB     ATAATGCATGCGCACACTGGTCCCAAATTTTACAATAAAATCAAAAGTATTGTAAATTCA 60
G. lamblia P15    ATAATACGTGCGCATACTGGTCCCAAATTTTGTAATAAAATCTAAAGTATTTTAAATTCA 60
G. lamblia GS     ----------------------------TTT--AAGAAAGTTTGAAGTATTTTAAATTCA 30
                                              ***  ** *** *   ******* ********

G. lamblia WB     ATTTGGGAAGAAAAAAAGTGAGG--CAGGCAGTCTCCATGACGAGAATTACGCCGCCCCA 118
G. lamblia P15    ATTTAGTAAGAAAAAAAGTGGGG--CAGGCAGTCTCCATGACGAGAATTACGCCGCCCCA 118
G. lamblia GS     ATTCG--AAGGAGAAAAATGGAGGCTGGGCAGTCTCCATGATGAAAGTTACGCCGCCCCA 88
                  ***    *** * **** **  *    ************** ** * *************

G. lamblia WB     GTCTGACCCCTGACGAACGGCTTCTCTGATCATTCACTCAACCC-GCCG- 166
G. lamblia P15    GTCTGAACCCTGACGAACGGCTTTTCTGATCATTCACTCAATCCTGCCCG 168
G. lamblia GS     GTCTGACGTCTGACGAACGGCTTTTCTGATCATTTACTCAA--CAG---- 132
                  ******   ************** ********** ******  * *
```

4. GlsR5

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR03	1407731	1407896	-	166
G. lamblia GS	ACGJ01002910	40807	40981	-	175
G. lamblia P15	contig377	136179	136345	-	167

```
G. lamblia WB     GAAATTTCACTTGAATTTCCAATTTAATATCGTTTT----------TTGCCGACTTCCAG 50
G. lamblia P15    GAAATTTCACTTGAATTTCTAATTTAATATCGTTTTC---------TTGCAAGCTTCCAG 51
G. lamblia GS     GAAATTCAAGTAAAACTTCTAATTTAATGTTGTTTTGCCATTCCTTTTGCCACTTCTCAG 60
                  ******  * *  ** *** ******** * *****          ****    *  ***

G. lamblia WB     AATCGCACTAAATTAAAAGCTGTGATGACAGGTTCTTGCCCCGTATGACCCTGCGATGAG 110
G. lamblia P15    AATCGCATCAAATTAAAAGCTATGATGACAGGTTCTTGCCCCGTATGACCCTGCGATGAG 111
G. lamblia GS     AACC-CACTAATTTAAAAGCTGTGATGACAGGTTCTTGCCCCGTATGACCCTGCGATGAG 119
                  ** * **  ** ********* ************************************

G. lamblia WB     TTATACAAAAGAACGCATCCAAGCCAACCGGCTGAGCTCCTTCACTCAAATCCTGC 166
G. lamblia P15    TTATACAAAAGAACACATCCAAGCCAACCAGCTGAGCTCCTTTACTAAAACATTAC 167
G. lamblia GS     TTATACAAAAGAACGCACTCAAGCCAACCGACTGAGCTCCTTTACTCAAATTCTGT 175
                  ************** **  **********  *********** *** ***   *
```

5. GlsR6

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR03	957087	957252	-	166
G. lamblia GS	ACGJ01002918	30612	30786	-	175
G. lamblia P15	contig38	47067	47231	+	165

```
G. lamblia WB     CTACCGTGTCCTACACTCTGACACTCAGCAGCTAA----------AGGCATTCCACCAC 49
G. lamblia P15    CTGCCGTGTCTTGCACTCTAACACTCAGCAGTTAA----------AGGTATTCCGCCAC 49
G. lamblia GS     CTGGCACGTCTTGCACTCTGACGTTTAGTAATTAGCTTAGTACTAGAAGCGCCTTGCCGT 60
                  **  *  *** * ****** **  * ** *  **           * *       **

G. lamblia WB     AAGTAAGTGAATAAATCGAAGTGAAACGTAAAAAATGCAATGATGGCTTGTTATCCCTGT 109
G. lamblia P15    AAGTAAGTGAATAAATCAAAGTGAAACATAAAAA-TGCAATGATGGCTTGTTATCCCTGT 108
G. lamblia GS     AAACGAGTGAATAAATTAAAGCGCGA--TAAAAAATGCAATGATGGCTTGTTATCCCTGT 118
                  **   ***********  *** *  *  ****** ************************

G. lamblia WB     CTGAGGTCAATACCTTGATTAGACGATTTGACAGAGCATCCTTCACTCAACACCCTT 166
G. lamblia P15    CTGAGGCTAACCTCCTGATTAGACGATTTGGCAGAGCATCCTTCACTCAATCCCCTC 165
G. lamblia GS     CTGAGGCCAATGGCTTGATTAGATGATTTGACAGAGCATCCTTCACTCAACCTCTTC 175
                  ******  **   * ******** ****** ******************   * *
```

6. GlsR7

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR01	808568	808733	+	166
G. lamblia GS	ACGJ01002422	30090	30256	-	167
G. lamblia P15	contig5	23963	24128	+	166

```
G. lamblia GS     CGGCATCGGGGGTAAAGCTTGGCGTTTTTTATTGAGAATCTAATTTCAGCAAGAATTTAC 60
G. lamblia P15    CGGCATCGGGGGTAAAGCTTGGCGTTTTTTATTGAGAATTTGTTTTTACCAGGAATTTAC 60
G. lamblia WB     CGGCATCGGGGGTGAAGCTTGGCGTTTTTTATTGAGAATTTGTTTTTAACGGGAATTTAC 60
                  ************* ************************ *  *** * *  ********

G. lamblia GS     AAAAACGGACGGCGTTCTGCCTCCTTCCCCGCAATGATTACTACATCACAGCGATATAGA 120
G. lamblia P15    AAAAACAGATGGCGTTCTGCCCTCCTCCC-GCAATGATTACTGAATCACAGCGACACATG 119
G. lamblia WB     AAAAACGGACGGCGTCTTGCCTTCCTCCC-GCGATGATTACCGAATCACAGCGATACACG 119
                  ****** ** *****  ****  * **** ** ********   ********** * *

G. lamblia GS     ATGAAGCGTTCATAGTTACTCTGAGCGGTCCTTTACTCAACAGGTAA 167
G. lamblia P15    ATGAAGCGTTCATAGTTACTCTGAGCGGTCCTTTACTCAACAGATAA 166
G. lamblia WB     ATGAAGCACTCATAGTTACTCTGAGCGGTCCTTTACTCAACAAGCAA 166
                  *******  ********************************   **
```

7. GlsR8

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR03	136649	136814	-	166
G. lamblia GS	ACGJ01002439	4398	4564	+	167
G. lamblia P15	contig30	12902	13068	+	167

```
G. lamblia WB     TACGCCATAATGTGATGAAAAGATACTTTAAAAA-TAGATTGTATTTTAAATTCACTTTG 59
G. lamblia P15    TAAGCCATAATGTAATGAAAAGATACGTTAAAAAATAGTTTGTATTTGAAATTCACTTTC 60
G. lamblia GS     TAAGCCATAGTGTGATGAAAATGTACTCCAAAAAATAATTTGCATTTGAAATTTACTTTT 60
                  ** ****** *** *******  ***   ***** **  *** **** ***** *****

G. lamblia WB     CAGCTCACAGAAAAGGGGTTCGTAGATGAAGAGAGATAAATCAGCTACCGCTGAGCCCAA 119
G. lamblia P15    CGGTCCACGGAAAAGGAGCTCGTAGATGAAGAGAGATAAATCAGCTACCGCTGAGCCCAA 120
G. lamblia GS     AATTATATGGAAAAGGGCCTCTTAGATGAAGAGAGATAAATCAGCTACCGCTGAGCCCAA 120
                        *  *******   ** ************************************

G. lamblia WB     CGTGAGGAAGAAACCGCCTTTCGTCTGACCCTTCACTCAACAGCCCC 166
G. lamblia P15    CGTGAGGAAGAAACCGCCTTTCGTCTGACCCTTCACTCAACAGCCCC 167
G. lamblia GS     CGTGAGGAAGAAACCGCCTTTCGTCTGATCCTTCACTCAACAGCTCC 167
                  *************************** *************** **
```

8. GlsR9

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR01	636600	636765	-	166
G. lamblia GS	ACGJ01002928	11774	11939	-	166
G. lamblia P15	contig348	14517	14692	-	176

```
G. lamblia WB     CAAAATCCATACTAAA----------AAATGGATGACAGTAATCATATTAAATTGCATTG 50
G. lamblia P15    CAAAGTTCATGCTAAATCATGCCAAAAAATGAATGATAGTAACCATATTAAATTGCATTA 60
G. lamblia GS     CAAAGTTCATGCTCAA----------AAATGAATGACAGTAACTATATTAAATTACATTG 50
                  **** * *** ** **          ***** **** *****  ********** ****

G. lamblia WB     CATGTGGGGAAAAAAGTCCTAGCAACCCGTGATTTGCAACGCTTAGTCCGTGTTTCGGAG 110
G. lamblia P15    CTGACAAGGAAAAAAGTCTTAGCAACCCGTGATTTGCAACGCTTAGTCCGTGTTTCGGAG 120
G. lamblia GS     CTGGCAGGGAAAAAAGACTAGATAACCCATGATTTGCAATGCTTAGTCCGTGTTTCGAAG 110
                  *      ********* *     ***** ********** ***************** **

G. lamblia WB     TGTCTTGCACGCTGATGAGTGAAAGCACACATGAGGTTCCTTTAATAAAATGCAGA 166
G. lamblia P15    TGTTTTGCACACTGATGAGTGAAAGCACACATGAGGTTCCTTTAATAAAATGCAAA 176
G. lamblia GS     TGTTTTGCACGCTGGTGAGTGAAAGCACACATGAGGTTCCTTTAATAAAATGCAGA 166
                  *** ****** *** ************************************** *
```

9. GlsR10

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR05	3181737	3181902	+	166
G. lamblia GS	ACGJ01002075	5843	6008	+	166
G. lamblia P15	contig161	35239	35404	+	166

```
G. lamblia WB     GTTCGTTTCCGGATTGCTCGTTTCCAGTTGTAAATTAATTTAAAGTGAATTCTTCTCAAG 60
G. lamblia P15    GTTCGTTTCCGGATTGCTCGTTTCCAGTTGTAAATTAATTTAAAATGAATTCTTCCCGAA 60
G. lamblia GS     TTTCGTTTCCAGATTGTCTGTTTCCAATTATGAACCAATTTAAAATGAATTCTCTCAGGC 60
                   ********* *****   ******* ** * **  ******** ********

G. lamblia WB     TTTTTACGCGGTGTCGAGAATGATGAGACGTGTTCCTCTCTCCTACAGACTCCCTGGGGA 120
G. lamblia P15    TTTTCGTGTGATGCTGAGAATGATGAGACGTGTTCCTCTCTCCTACAGACCTCCTGGGGA 120
G. lamblia GS     TTTTTGCGCCAGGCCAAAGATGATGAGACGTGTTCCTCTCTCCTACGGACACCCTGGGGA 120
                  ****   *    *   *  *************************** ***  ********

G. lamblia WB     TGCTATGTACACCTTACTGATTTACTTTCCTTTTCTCAAGGGCCAT 166
G. lamblia P15    TGCTATGTACACCTTACTGATTTACTTTCCTTTTCTCAAAGGCTAT 166
G. lamblia GS     TGCCATGTACACCTTACTGACTTACTTTCCTTCTCTCAAGATCTAG 166
                  *** ***************** ************ ******   * *
```

10. GlsR13

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR05	4314721	4314886	-	166
G. lamblia GS	ACGJ01002906	52793	52956	-	164
G. lamblia P15	contig39	120508	120671	-	164

```
G. lamblia WB     AGTCATCTATTTAGAATTGGAATTAGACTTTGAAATTCATCGCCCTCCGATCCATTCGTA 60
G. lamblia P15    --TCATCTATTTAGAATTGGAATTAGACTTTGAAATTCATCGCCTCCCGATCCATTCGTA 58
G. lamblia GS     AATCATCTATTTAGAATCAGAATCGGGTTTTAAAACTCTTCCCTCTCCAATCCATTCGTG 60
                    ***************  ****  *  *** *** ** ** *   ** **********

G. lamblia WB     TGAGATATGATGATTGGGAGCGACCTATCTTGAGGACGACGGCCGCCCGTCTTACCTTGT 120
G. lamblia P15    TGAGATATGATGATTGGGAGCAACCTATCTTGAGGATGGCGGCCGCCCGTTTTACCTTGT 118
G. lamblia GS     TGAGATATGATGATTGGGAGCGACCTATGTTGAGGATGGCGGCTGCCCGTCTTACCTTGT 120
                  ******************** ****** ******* * **** ****** *********

G. lamblia WB     GACGTTTGCCGTCTTACAATGCTCTGACCCTTTACTTAAGCTGCCG 166
G. lamblia P15    AACGTTTGCCGTCTTACAATGCTCTGACCCTTTACTTAAGCTGCTG 164
G. lamblia GS     GACTTCTGCCGTCTTGCAATGCTCTGACCCTTTACTCAAGCCAC-- 164
                   ** * ********* ******************** ****  *
```

11. GlsR14

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR01	1279750	1279915	-	166
G. lamblia GS	ACGJ01002915	13366	13500	-	135
G. lamblia P15	contig173	79729	79895	+	167

```
G. lamblia WB     CACAGACAAA-ACTAATCCACCAGTAGAATGACGAGGGGGTACACCGACAGCGGTTGATC 59
G. lamblia P15    --TATAAAAA-ACTAATCCGTCAGTAGAATGACGAGAAAATACACTGACAGCAGCTGACT 57
G. lamblia GS     TATAGCTAAATGCTATTTAGTTGATAG--CTGCGA-----------ACAGC-------- 38
                     *   ***  *** *       ***     ***           *****

G. lamblia WB     TCCAC---GGGAACCGGAAATAAATAAAATGATGACAATGCGCATTTGTCAGAAGGCTCA 116
G. lamblia P15    TCCACCACGGGGGCTGAAAATAAATAAAATGATGACAATGCGCATTTGTCGGAAGGCTTA 117
G. lamblia GS     ------------CTAGGAATAAATAAAATGATGATAATGCGCATTTGTCGGAAGGCTCA 85
                            *    ***************** ************** ******* *

G. lamblia WB     CTTCTGATGATTCCTCTGTCCATTCCCCTGATCCTTTCCTCAACAGGTAT 166
G. lamblia P15    CTTCTGATGATTCCTCTGTCCATTCCCCTGACCCTTTGCTCAATAGGTAT 167
G. lamblia GS     CTTCTGACGATTCCTTTGTCCATTCCCCTGACCCTTTATTCAAAGATCAA 135
                  ******* ******* *************** *****  ****     *
```

12. GlsR15

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR01	1329992	1330157	-	166
G. lamblia GS	ACGJ01002916	840	1001	-	162
G. lamblia P15	contig173	30316	30481	+	166

```
G. lamblia WB     AAGAGGCTGCGACGCGGGTTATTCAGTTCGATGCGCCCAGGCTGACGGTAGGACGCCTAA 60
G. lamblia P15    AAGAGGCCGCGACGTGGGTTGTTCAGTTCGATGCGCCCAGGCTGACGGTAGGACGCCTAA 60
G. lamblia GS     AAGGGGCTGCGACGCGGGTTGTTCAGTTCGATGCGCTCAGGCTGACAGTAGGACGCCTAA 60
                  *** *** ****** ***** *************** ********* *************

G. lamblia WB     CCCGATTCAGACTACTCCTTGGTTCCTCGCAGAATGATTATCTGTCTCCGAGCAAGCACG 120
G. lamblia P15    CTCGATTCAGACTACTCCTTGGTTCCTTGCAGAATGATTATCCGTCTCTGAGCAAGTGCG 120
G. lamblia GS     CTCAATTCAGACTACTCCTCGATCCTTCGCAGAATGATTATCTATCTCTGGGCAAGCGTG 120
                  * * *************** * * * * **************  **** * *****   *

G. lamblia WB     ACTATGAGCTTACTTATGAGATCTGACTCCTTTACTCAATGTTAGT 166
G. lamblia P15    ACTATGAGCTTACTTATGAGATCTGACTCCTTTACTCAATGTCAGA 166
G. lamblia GS     GCTATGAGCTTACTTATGAGATATGACTCCTTTACTCAATGA---- 162
                   ******************** ******************
```

13. Candidate-1 [as named in (Chen *et al.* 2007)]

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR01	1215994	1216159	-	166
G. lamblia GS	ACGJ01002331	42201	42365	+	165
G. lamblia P15	contig25	103482	103675	+	194

```
G. lamblia WB     CAGCAAGTTCAAGTCTGGGAACCGAGATCGTTTCAAAAACGGTTTTAAAAAGC----TCC 56
G. lamblia P15    TCGGAAGTCTAAATCCAGAGACCAAAGTCGTTTAAAAAATGATTTTAAAGAGCGAGCTTT 60
G. lamblia GS     CAAAAAATTCAACACCAGAGACAAAATCTGAGTCAAAAACAGTTTTAAA-AGC----CCC 55
                      ** *  **  *  *  **  *    *  * *****   ******* ***

G. lamblia WB     GAAGCAAATGAGAACAAAAGCA-GACGAAAAAATAAATGAAGACAGAACCACAGACCTGT 115
G. lamblia P15    GAAGCAAATGAGAAACAAAATATGAAAAACAAAAAAATGAAGACAGAACCACAGACCTGT 120
G. lamblia GS     GAAATAAAT------------------------AAATGAAGATAGAACCACAGACCTGT 90
                  ***  ****                         ********* ***************

G. lamblia WB     ACTGACCCTTGATGTTAGTTGTCGCTCTGATATCCTTTACTCAATCGTGT-C-------- 166
G. lamblia P15    ACTGACTCTTGATGTTAGTTGTCGCTCTGATATCCTTTACTCAATCATTT-CTGGGACTA 179
G. lamblia GS     ACTGACTATTGATGTTAGTTGTCGCTCTGATACCCTTTACTCAAGCTTTTTCAAGTGTCG 150
                  ******  ********************** *********** * * * *
```

14. Candidate-2

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR01	33235	33400	+	166
G. lamblia GS	ACGJ01002208	5884	6052	+	169
G. lamblia P15	contig52	39757	39923	-	167

```
G. lamblia WB     GTCTTT-TTCCAGAATTTGTTCCTTTCAGTGTTTAGTGTCTTT-GTCTTTTATCTTAGC- 57
G. lamblia P15    GTCTTT-TTCCAAAATTCGTCTCTTTCAACATTTAATGTTTTTTGTCTTTTACTCTAAC- 58
G. lamblia GS     GTCTGTGTTCCAAAATTTGCTCCCTTCCATTTTTAAGTTTTTTTGGCTTTTA--TTAATC 58
                  **** * ***** **** *   * ***    ****   * *** * ******   **

G. lamblia WB     -TTTTCTATTAATTGAAAGTCGA-AAATAAAGTGATGATTCGAATTACCGCCCGAGGGCC 115
G. lamblia P15    -TTTCTTATTAATTGAAGGATGA-AAATAAAGTGATGATCCGAATTACCGCCCGAGGGCC 116
G. lamblia GS     ATTTTCTATTAATTGAGAATCGCTAAATAAAGTGATGATCCGAATTACCGCCCGAGGGCC 118
                   ***  **********     *  *************** ********************

G. lamblia WB     CTCGGGCTCCGCTGAGGACATGCTGGTCTGACTCCTTTGCTCAACCTTTCC 166
G. lamblia P15    CTTGGGCTCTGCTGAAGACATGCTGGTCTGATTCCTTTGCTCAACCTTTCT 167
G. lamblia GS     TTCGGGCTCCGCTGAGGACATGCTGGTCTGACTCCTTTTCTTAATCTTTCT 169
                   * ****** ***** *************** ****** ** ** *****
```

15. Candidate-13

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR01	1010277	1010442	-	166
G. lamblia GS	ACGJ01002748	30795	30925	-	131
G. lamblia P15	contig59	121266	121421	+	156

```
G. lamblia WB     TAGGTATACTTTGTGCGGACTAGAAACGAACTAGAAAATCAGTAAA-AAGGTCTTGAGCA 59
G. lamblia P15    TAAATGTACTTTGTGAG----------AACTAGAAAGCCAGTAAATAAGGTCAAAAACA 49
G. lamblia GS     --CACGTGTTCTG-GCG---------------TGAAA-----TAAATAA---------- 26
                        *  * ** * *                ****     **** **

G. lamblia WB     AAACCAGTAAATTAAAAATGATTACTCCAACACGACGGTCTACTGAGAACCCAGTATCTT 119
G. lamblia P15    AAGCCAACAAATAAAAAATGATTACTTTAACACGACGGTCTGCTGAGAACCCAGTACCTT 109
G. lamblia GS     --ATCAGTAAATAAAAAATGATTACTCCAACACGACGGTCTGCTAAGAACCCAGTATCTT 84
                      **  **** *************  ************* ** *********** ***

G. lamblia WB     TAGACTGCTGAGACAGTGTTATATGATTTCCTTTACTTAAGGCTCTC 166
G. lamblia P15    TAGACTGCTGAGATAGTGTTATATGATTTCCTTTACTTAAAGTTCAC 156
G. lamblia GS     TAGACTGCTGAGACAGTGTTGTATGATTCCCTTTATTCAAACATCTC 131
                  ************* ****** ******* ****** * **   ** *
```

16. Candidate-23

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR01	449998	450163	-	166
G. lamblia GS	ACGJ01001805	1801	1965	+	165
G. lamblia P15	contig818	118027	118189	+	163

```
G. lamblia WB     GAGGCATGTATAATTATACCAAAATTAATTGCAGAGTTCTCCTTTTTCAAAAAGCCTCTC 60
G. lamblia P15    -GGGTGTGTGTAATTATACACAAATTAATTGCAAAATTCT--TTTTTTAAAAAGCCTCTC 57
G. lamblia GS     -GGGGATGTATAATTATACACAAATTAATTGTAAAGTCGACTTTTCTCAAAATGCCCCTC 59
                    **  *** *********  ********** * * *     *** * **** *** ***

G. lamblia WB     TGTAGGTAGGGCCGATGAGCTATTTTGTACCACTCTGACCGTGAGGCGTATGCCTAGGGC 120
G. lamblia P15    TGTAGGCAGGGCCAATGAACTGTTTTGTACCACTCTGACCGTGAGGCATGCGCCTAGGGC 117
G. lamblia GS     CGTGGGGGAGGCCGATGAGCTATTTTGTACCACTCTGACTGCAGGGCGTACGCCCAGGGT 119
                   ** **   **** **** ** ***************** *   *** *  *** ****

G. lamblia WB     ATGGAGAAGAGCAGACTTGAGGCCTGTTCCTTTACTCAATTTTGTG 166
G. lamblia P15    ATGGAGAAGAGCAGACTTGAGGCCCGTTCCTTTACTCAATTTTGTG 163
G. lamblia GS     ATGGAGAAGAGCAGACTTGAGGCCTATTCCTTTATTCAATTGTCCG 165
                  ***********************  ******** ****** *  *
```

Box H/ACA RNAs

17. GlsR17

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR01	149315	149480	+	166
G. lamblia GS	ACGJ01001410	76605	76772	+	168
G. lamblia P15	contig11	8260	8426	+	167

```
G. lamblia WB     GGTCAGTTTCTAG-ACCTCCTGGGATAATGCGCTTCTTTGAGCCGCGGGTTTACTCGTGG 59
G. lamblia P15    GGTCAGTTTCTAG-ACCTCCTGGGATAATGCGCTTCTTTGAGCCGCGGGTTTACTCGTGG 59
G. lamblia GS     AAGCAGTTTCTAGCACCTCCTGAGTTAATGCGCTTCTTTGAGCCGCGGGTTTACTCGTGG 60
                     ********** ******** * **********************************

G. lamblia WB     TGAGGATCCGGGGCACTGAGCAATCCCCAGGACACAGGCGGAGCGGAAGGCACGGCTGCG 119
G. lamblia P15    TGAGGATCCGGGGCACTGAGCAATCCCCAGGACATAGACGGAGCGGAAGGCACGGCTGCA 119
G. lamblia GS     TGAGGATCCGGGGCACCGAGCAATCCTCAGGACACAGACGGAGCGGAAGGCACGGTTGTG 120
                  *************** ********* ******* ** ***************** **

G. lamblia WB     CCACGCAGCCTAATCACCGCCCCTATAGTCCTTTTCTAAACGC-GTGG 166
G. lamblia P15    CTGTGCAGCCTAATCACCGCCCCTATAGTCCTTTTCTAAACCTTATGG 167
G. lamblia GS     TGACGCAACCTAATCACCGCCCCGACAGTCCTTCGCTTAATTATGCGG 168
                      *** *************** * *******  ** **      **
```

18. GlsR18

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR01	149436	149601	+	166
G. lamblia GS	ACGJ01001410	76727	76894	+	168
G. lamblia P15	contig11	8381	8548	+	168

```
G. lamblia WB     ACGCAGCCTAATCACCGCCCCTATAGTCCTTTTCTAAACGC-GTGGCCGGTGCAGCTGCC 59
G. lamblia P15    GTGCAGCCTAATCACCGCCCCTATAGTCCTTTTCTAAACCTTATGGCTGGTGCAGCCGCC 60
G. lamblia GS     ACGCAACCTAATCACCGCCCCGACAGTCCTTCGCTTAATTATGCGGCCGGTGCGGCTGCC 60
                    *** *************** * *******  ** **      *** ***** ** ***

G. lamblia WB     CGCTGGCGCTTGCGAGCGTGCACAGGCCTACATCCAGGGTCATAGGTGGGGAGCGGATCC 119
G. lamblia P15    CGCCAGTGCTTGCGGGTGTGCACAGGCCTACATCTAGGGTCATAGGTGGGGAGCGGATCC 120
G. lamblia GS     CACTGGTACCTTTGGGCGCGCACAGGCCCACAGCCGGGGTCATAGGTGGGGAGCGGATAC 120
                  * *  *  * *  * * * ********* *** *  ********************** *

G. lamblia WB     CGTCCATCCTCAATCCGGGCCCGCACA-GTCCTTTACTCAAGCTTACT 166
G. lamblia P15    TGTCCATCCTCAATCCGGGCCCGCACATGTCCTTTATTCAAGTTTACT 168
G. lamblia GS     TGTCCATCCTCAATCCGGGCCCTCACATGTCCTTTACTCAAATTTCAT 168
                   ******************** **** ******** ****  **  *
```

19. GlsR19

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR05	2415660	2415825	-	166
G. lamblia GS	ACGJ01001859	18642	18807	-	166
G. lamblia P15	contig10	39723	39888	-	166

```
G. lamblia WB     TGGAGGCTCGGCGTCTCGTTCTGGGAAAAGCAAGCAGAAGCCCAGTTTGGTCTCTACCGG 60
G. lamblia P15    TGGAGGCTTGGCATCCCGTTCTGGGAAGGGCAAGCGGAGATCCAGTCTGGTCTCTACCAG 60
G. lamblia GS     TGGAGGTCTGACGTCCCGTTCTGGGAGGGGAGGATGGAAGCCTAGCTTGGTCACTACCAG 60
                  ******   * * ** **********   *      **   * **  ***** ***** *

G. lamblia WB     CGTATGCATGTGCATAGGCTGGCCAAGCATCGTTGATAGAAGCTGCTCTTGGTCACCGGA 120
G. lamblia P15    CGTATGCATGTGCGTAAGCTGGTCAAGCATCGTTGATAGAAGCTGCTCTTGGTCACCGGA 120
G. lamblia GS     TGTGCACATGTGTGTAGACTAGTCAAGCATCGTTGATACAAGCTGCTCTTGGTCACCGGA 120
                   **   ******  **  ** * *************** ********************

G. lamblia WB     GGGTCTCCGGTTTCATACGCAGAGACATCCTTCAATTAAAAACTTT 166
G. lamblia P15    GGGTCTCCGGTTTCATACGCAGAGACATCCTTCAATTAAAAACTTT 166
G. lamblia GS     GGGCCTTCGGTTTCATATGCAGAGACATCCTTCAATTAAAAACTTT 166
                  *** ** ********** ****************************
```

20. GlsR20

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR04	893921	894086	-	166
G. lamblia GS	ACGJ01000491	21037	21190	+	154
G. lamblia P15	contig4	36319	36488	-	170

```
G. lamblia WB     CCATCCAGTTTGATA----GGGGGT-TCTTTTCTTTTTGCCAAGTTAAAAATGCCAGCTG 55
G. lamblia P15    CCATCCAGTGTTAAATGATGGGGGT-TCTTTTCTTTTTGCTAAGTTAAAAATGCCAGCTA 59
G. lamblia GS     CCATC-AATGTTGACGCTTGAAGTTATCTTTTTCTTTTG----GTTAAAAATGCCAGCTA 55
                  ***** * * *         *  * * ******  *****    ****************

G. lamblia WB     AGTTACGTCTGTGTGCACAGGCGCGTCAGAGGCCGGCTAGAGCGCGACTGGTTGAGTTCC 115
G. lamblia P15    AGTTACGTCTGTATATACAGACGCGTCAGAGGCTGGCTAGAACGCGACTGATTGAGTTCC 119
G. lamblia GS     AGTTACGTCTGTATGTACAGGCGCGTCAGAGGTTGGCTAGAGCGTGACTGGTTGAGTTCC 115
                  ************ *  **** ***********  ******* ** ***** *********

G. lamblia WB     CAGAGCGATCTGGGTGATTAGCAGTCATACAGTCCTTTACTTAAGC----CTACT----- 166
G. lamblia P15    CAGAGCAATCTGGGTGATTGGCAGTCATACAATCCTTTACTTAACT----CTACT----- 170
G. lamblia GS     TGGGGAAGCCTGGGTGATTAACAGTCATATAGTCCTTTACTCAAGTTGGGCTTCCAGTCT 175
                    * *    **********  ******** * ********* **      ** *
```

21. GlsR21

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR04	1151475	1151640	-	166
G. lamblia GS	ACGJ01002311	35663	35829	-	167
G. lamblia P15	contig696	15261	15427	+	167

```
G. lamblia WB     ACATTTTAATTGTCG-CTTCAAGCAAAAGTGTACTGTATAAAAACCAATATTACTACCAT 59
G. lamblia P15    ACATTTTAATTGCCG-TTTCAAGCAAAAGTGTACTGTATAAAAACCAATATTACTACCAT 59
G. lamblia GS     ACATTTTAATTGTTGGTTTCTAGCAAA-GTATACTGCATAAAAACCAATATTACTACCAT 59
                  ************  *  *** ****** ** ***** **********************

G. lamblia WB     CGGTCCTCACCACTAGATCGGTGTTATGCTTTGTTGGGATAGCAGGCCGTGCCAGTTGGA 119
G. lamblia P15    CGGTCCTCACCACTAGATCGGTGCTATGCTTTGTTGGGACAGCAGGCTGTGCCAGTTGGA 119
G. lamblia GS     CGGTCCTCACTACTAGATCGGTGTTATGCTTTGTTGGGATAACATGCCGTGCCAATCAGG 119
                  ********** ************ *************** * ** ** ****** *  *

G. lamblia WB     CAGCCAAGGTCCACCT-CTGGTTCGGCACACATTTATTCAAGACATCT 166
G. lamblia P15    CAACTAAGGTCCATCTACTGGTCCAGCACACATTTATTCAAGATATTT 167
G. lamblia GS     CAGCTAGTGTCTCTCTACTGGTGCGGCATACATTTATTCAAGACATCT 167
                  ** * *  ***   ** ***** * *** ************** ** *
```

22. GlsR22

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR03	1225005	1225170	-	166
G. lamblia GS	ACGJ01002895	16172	16337	-	166
G. lamblia P15	contig18	1797	1962	-	166

```
G. lamblia WB     GTCAATTCTATCATATTTTTTTGACAGCCTGCGACGCAAGCCCTCTAGCAAGATGCAGGC 60
G. lamblia P15    GTCAATTCTATCGTATTTTTTTGATAGCCTGAGACGCAAGCCCTCTAGCAAGATGCAGGC 60
G. lamblia GS     GTAAATTCTATAGCATTTCTTTGACAGCCTGTGACGCAAGCCCTCCAGCAAGGTGCAGAC 60
                  ** ********   **** ***** ****** ************* ****** ***** *

G. lamblia WB     CGGAGCCTGTGTCTCGTTCCCTGGGGCGATAGCTCTTGCTGGCAGGTCTTGCAGTGTCCA 120
G. lamblia P15    CGGAGTCTGTGCCTCGTTCCCTGGGGCGATAGCTCTTGCTGGCAGGTCTTGCAGTGTCCA 120
G. lamblia GS     CGGAGTCTGTGCCTCGTTCCCTGGGGCGATAGCTACTGCTGGAGGGTTTTGCATTATCCG 120
                  ***** ***** **********************  ******  *** ***** * ***

G. lamblia WB     TACCCGGGCAACACGTTTTCCAGCTACACCTTTACTCAACGTGCAC 166
G. lamblia P15    TGCTTGGGCAACACGTTCTCCAGCTATACCTTTACTCAACGTGTAT 166
G. lamblia GS     TATCTGGATAACACGTCCCCCAGCTATACCTTTACTCAAAGTGCGC 166
                  *    **  *******   ******* ************ ***
```

23. GlsR23

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR04	1890711	1890876	-	166
G. lamblia GS	ACGJ01001044	23638	23804	-	167
G. lamblia P15	contig54	22832	22998	+	167

```
AssemblageA       GACTTAGATACTGACATACTGCAGCTG-CGATAAAAAA-TCGTCTGGCCGCCTCGTGCCT 58
G. lamblia P15    GTCCTGGATACTGACA-ACCGCGGCTGGTAACAAAAAAATCGCCTTGTCGTCTCATGCCT 59
G. lamblia GS     GATCTGAGTGTTGGTATGCTGCAAAGGACGATAAAAAA-TCGCCATATGGTCTCATGCCT 59
                  *   *   *  **  *  * **    *   * ****** *** *     * *** *****

G. lamblia WB     ACGATGGGCTAGGGAAATGCCGTGACGAGACACGCACTGGGTGGCCATTGCGTCTGCGGT 118
G. lamblia P15    ACGATGGGCTAAAGAAGCACTGTGGCGAAACATGCACTGGGTGGCCATTGCGTCTGCGGT 119
G. lamblia GS     ACGATGGGCCAGAGGAGCATTTTGGCAAGACACACACTGGGCGGCCATTGCGTCTGCGGT 119
                  ********* *  * *      ** * * ***  ******* ******************

G. lamblia WB     AGATCCGCCGATTCCACAGCCCAGAAACACCCTTTACTCAAGCTGGCT 166
G. lamblia P15    AGACCCGCCGATTCCACAGCCCAGAAACATCCTTTACTTAAGCTGGCT 167
G. lamblia GS     AGACCCGCCGATTCCACAACCCAGGAACAGCCTTTACTCAAACCGATT 167
                  *** ************** ***** **** ******** ** * *  *
```

24. GlsR24

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR05	1301238	1301403	+	166
G. lamblia GS	ACGJ01002360	10351	10498	-	120
G. lamblia P15	contig463	79	259	-	181

```
G. lamblia WB     ----------------ACCGCCTGTAAACGCCTTCCCCGAATAATTTCGGG-CAGTCCT 42
G. lamblia P15    ACCAGCGCCTCTGAGCCCCCGCCTGTAAACGCCTTCCCCAAAA---TTCGGG-CAGTCCT 56
G. lamblia GS     ----------------------------------CTCCAATTTATTTCGGGGCAGTTCT 25
                                                   * ** *     ****** **** **

G. lamblia WB     TGCCCGCGGAGGGCACTTAAGGCTCCGGGGCCCGGGGCAGAGTCGGCCCTCCCAGAGCCC 102
G. lamblia P15    TGCCCGCGGAGGATCCTTAATGCCCCGGGGCCCGGGGCAGAGTCGGCCCTCCCAGAGCCC 116
G. lamblia GS     TGCCCGGGGAGGATACTTAATGCTCGGGTCCGCGCAGCAGAGTCAGTCCTCCCAGAGCCC 85
                  ****** *****   ***** ** * **  * **  ******** * **************

G. lamblia WB     GCCGACGCCCCCGAGCGCCAGCCCGGCGCAGGGGCCCGGCCACACTCATTTATTAAAC-A 161
G. lamblia P15    GCCGGCGCCCCCGAGCGTCAGCCCGGCGCAGGGGGCCGGCCACACTCATTTATTAAACCA 176
G. lamblia GS     GCCGGGACCCCCATGCGTCTGGCAGGCGCGGGGG-CCGGCCACATTCATTTATTAAAC-G 143
                  ****   *****  *** * * * ***** **** ********* **************
```

25. GlsR25

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR05	1459333	1459498	+	166
G. lamblia GS	ACGJ01001903	4369	4535	+	167
G. lamblia P15	contig778	18935	19100	+	166

```
G. lamblia WB     CTTTGCAACGGCACTTGTAACGAAAAAGTAAATCGAGGCTGCTAAAACACAGGGCTGCAC 60
G. lamblia P15    CTTTGCAACGGCACTTGCAACGAAAAAGTAAATTGAGGCTGCTAAAACACAGGGCTGCGC 60
G. lamblia GS     CTTTGCAACGGCATTTATAACGAAAAAGTAAATTGAGGCTGCTAAAACACAGGGCTGCAC 60
                  ************* **  *************** ********************** *

G. lamblia WB     AGCATCCTTGCACCTGCGTAGCCGATAGGTACGGGTGACCGTTTATCCCGGGCTCGTGTG 120
G. lamblia P15    AGCATCCTTGCACCTGCGCAGCCGACAGGTACGGGTGACCGTTTATCCCGGGCCCACGTG 120
G. lamblia GS     AGCATCCTTGCACCTGCGTAGCCGATAGACGCATTTGGCTGTTTATCTAGGATCTGCGTG 120
                  ****************** ****** **   *   ** * *******  **      ***

G. lamblia WB     GGCCCGGGTAGGCACGGTCAAAGAGTTTCCTTCATTCAA-TTTTTAG 166
G. lamblia P15    GATCCGAGCAGGCACGGTCAAAGACTTTCCTTCATTCAA-TTTTTAG 166
G. lamblia GS     GGTCTGAACAGGCACGGTCAAACAGTTTCCTTTACTCAAGTTTTTAG 167
                  *  * *   ************* * ******* * **** *******
```

26. Candidate-16 [as per reference (Chen *et al.* 2007)]

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR05	1006432	1006597	-	166
G. lamblia GS	ACGJ01001794	6880	7044	-	165
G. lamblia P15	contig399	33225	33390	+	166

```
G. lamblia WB     TGGATTGCTTCTTAAAAGATGGCCG-GAAGAGAAAAAGATCAAAAGCAAGGCTAGAGCCA 59
G. lamblia P15    TGGATTGCTTTTTAAAAAATGGCTG-GAGGAGAAAAAGATCAAAAGCAAGGCTAGAGCCA 59
G. lamblia GS     TGGATAGCTTTTTAAA-GATGGCTACAAAGAGAAAAAGATCAAAAGCAAGGCTAAAGCCA 59
                  ***** **** *****  *****    * ************************ *****

G. lamblia WB     TGGAGCGCGGATCTGCGCTCTGCCAGATACGCCGACAGAAAGCACCAAGGAAGGATGTGG 119
G. lamblia P15    TGGAGCGCAGACCTGCGCTCTGCCAGATACGCCGATAGGAAGCGCCAAGGAAGGACGCGG 119
G. lamblia GS     TGGAGCGCGGATTCGCGTTCTGCCAGATACGCCGATAGAAAGCAACAAGGAAGGACGTGG 119
                  ******** **   *** ***************** ** ****  ********** * **

G. lamblia WB     ATCTCCATGTCTGCCGTGTGCGCGCATATCCTTTACTCAATCTGTGT 166
G. lamblia P15    GCCTCCGTGTCTGCCGTGTGCGCGCATATCCTTTACTCAATCTGTGT 166
G. lamblia GS     -TTTCCATGTCTGCCATGTGCGCGCATATCCTTTAATCAATTTACGT 165
                     *** ******** ****************** ***** *  **
```

27. GlsR26

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR05	1459471	1459636	+	166
G. lamblia GS	ACGJ01001903	4507	4674	+	168
G. lamblia P15	contig778	19073	19238	+	166

```
G. lamblia WB     CAAAGAGTTTCCTTCATTCAA-TTTTTAGATCCATCTGGCGCACGTTACGAGGCCTGGCT 59
G. lamblia P15    CAAAGACTTTCCTTCATTCAA-TTTTTAGATCCATCTGGCGCACGTTACGAGGCCTGGCT 59
G. lamblia GS     CAAACAGTTTCCTTTACTCAAGTTTTTAGATCCATCTGGCGCACGTTACGAGGCTATGCT 60
                  **** * ******* * **** ******************************   ***

G. lamblia WB     CCCGGTCCGGCCCTACCTTGCGTGCGCATATCTCCGGGATCTGCGCCGCGTCTGCTCGCG 119
G. lamblia P15    TCCGGTCCGGCCCTACCTTGCGTGCGCATAGCTCCGGGATATGCGCCGTGTCTGCTCACG 119
G. lamblia GS     TCCGGCCTGGCCCTACCTTGCGTGCGCATAGCTCCGGGACATACACTACGTCTGGTCGTG 120
                   **** * ********************** ********  * * *   ***** **  *

G. lamblia WB     GCGATTTCGGTTATGCCGGCCCGAACA---CTCCTTCATTCAACAGGCCC 166
G. lamblia P15    GCGATTTCGGCTATGCCGGCCCGAACA---CTCCTTCGTTCAACAGGCTC 166
G. lamblia GS     ACGACTGCGACTGTACTGACCCGAACATTGCTCCTTCATTCAACAAAC-- 168
                   *** * **  * * * * ********   ******* *******  *
```

28. GlsR27

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR01	1371240	1371405	+	166
G. lamblia GS	ACGJ01002332	21871	22119	-	249
G. lamblia P15	contig9	61259	61425	-	167

```
G. lamblia WB     GGGCGAGATAT-------------------------CTT-------------------- 14
G. lamblia P15    GGGCGAGATAT-------------------------CTC-------------------- 14
G. lamblia GS     GAGCGAGAGATTGTACGGTATAGTTTGAGTAAAATCCCTTAACAGGAAATTTGCTTAGCC 60
                  * ****** **                          **

G. lamblia WB     -----TCA------------------------------GCATTTATAAC-CA-AAAAAT 36
G. lamblia P15    -----TCA------------------------------GTATCTATGGCTCA-AAAAAT 37
G. lamblia GS     AGTAATCAATGGGTCATTGCATTAAGTCAAGTAGGTGCTGTATTTACAAGCCACAAAAAT 120
                       ***                                 * ** **     ** ******

G. lamblia WB     TAAGCTCACCCAAAGTCAACGGAGCGCCAGCTACGTGTTATGGGCAGCGAAAGTACCAGA 96
G. lamblia P15    TAAGCTCACCCAAAGTCAACGGAGCGTCAGCTACGTGTTATGGGCAGCGAAAGTACCAGA 97
G. lamblia GS     TAAGCTCACCCAAAGTCAACGAAACGCTAGCTTCGTGTTATGGGCAGCGAAAGTGCCAGA 180
                  ********************* * **  **** ********************* *****

G. lamblia WB     GCCAAAGAGTTCCTCTGATCGCCTGGCCGGAGCACATTTGTGATCTCCTATACCTTCATT 156
G. lamblia P15    GCCAAAGAGTTCCTCTGATCGCCTGGTCGGAGCACATCTGTGATCTCCTACACCTTCACT 157
G. lamblia GS     GCCAAAGAGTTCCTCTGATCGCCTGGCTGGAGCACACATGTGACCTCCTACACCTTTATT 240
                  *************************  ********  ***** ****** ***** * *

G. lamblia WB     TAATTAGCGT 166
G. lamblia P15    TAATTAGCCT 167
G. lamblia GS     TAAC-GGCCT 249
                  ***   ** *
```

Spliceosomal snRNA Candidates and RNase MRP RNA

29. Gl U1 snRNA

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR03	661194	661359	-	166
G. lamblia GS	ACGJ01002471	10962	11125	+	164
G. lamblia P15	contig621	4079	4244	+	166

```
G. lamblia WB     TGAGCAGGTCAAAAATTGAAGGTAATTTTA ACTTACCTCAAGGGTGGCGACGAGCCAGTG 60
G. lamblia P15    TGAGCAGGCTAAAAATTGGAGGTAATTTTAACTTACCTCAAGGGTGGCAACGAGCCAGTG 60
G. lamblia GS     TGAGCAGGTCAAAA-TTGAAGACAATTTTAACTTACCTTAAGGGTGGCGATGAGCCATTG 59
                  ********  **** *** **  ********************* ********* * ****** **

G. lamblia WB     TTCGGGCCAGGCTGGTGCTGCGCATACCGCGCTGGCACTGGTCACGGGGCAGTGCTCTCA 120
G. lamblia P15    TTCGGGCCAGGCTGGTGTTGCGCATACCGCGCTAGCACCGGTCACGGGGCAGTGCTCTCA 120
G. lamblia GS     TTCAGGCC-GGCTAATGCTGCGCATACCGCGCTGGTGTTTGTCACAGAGCAGTGCTCTCA 118
                  *** **** ****  ** *************** *     ***** * ************

G. lamblia WB     GACCTGCTACCGTACCCTTTTAATTTTTCCTTCACTTAAGGCCCAT 166
G. lamblia P15    GACCTGTTACCGTACCCTTTTAATTTTTCCTTCACTTAAGGCCCAT 166
G. lamblia GS     GACCTGCTACTGTACCCTTTTAATTTTTCCTTCGCTCAAGGCCTAT 164
                  ****** *** ********************** ** ****** **
```

30. Gl U2 snRNA

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR04	581396	581561	-	166
G. lamblia GS	ACGJ01002726	93831	93991	+	161
G. lamblia P15	contig698	112750	112915	+	166

```
G. lamblia WB     AGGCAAAAATTAAAATCAGAGTCGGCTTCGACTTTAGTGTAGTTACTGTTTCGTCGGCTT 60
G. lamblia P15    AGGCAAAAATTAAAATCAGAGTCGGCTTCGACTTTAGTGTAGTTACTGTTTTATCGGCTT 60
G. lamblia GS     -----GTAATTAAAACCAGAGTCGGCTTCGACTTTAGTGTAGTTACTGTTTTGTCGGCTT 55
                        ******** ********************************  *******

G. lamblia WB     AACCGCCGATCCACTACATGCAAGGGGCAGCCGGGCTGTGAGGCAGCTGCCAGGATGGTC 120
G. lamblia P15    AGCCGCCGATCCACTGCATGCAAGGGGCAGTCGGGCTGTGAGGCAGCTGCCAGGATGGTC 120
G. lamblia GS     AACCGCCGATCCATTACATTCAAGGGGCAGTCGGGCTGTGAGGCAGCTGCCAGGATGGTC 115
                  * *********** * *** ********** ****************************

G. lamblia WB     CTGCCCTTGTCCCGGCTGGCGCCGTCCACCTTTATTCAAGTTTTCT 166
G. lamblia P15    CTGCCCTTGTCCCGGCTGGCGCCGTCCACCTTTATTCAAGCTTTCT 166
G. lamblia GS     CTACCCTTGTTCCGGCTGGCGCCGTCCACCTTTACTCAAGCTTTCT 161
                  ** ******* ********************** ***** *****
```

31. Gl U4 snRNA

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR02	1194980	1195145	+	166
G. lamblia GS	ACGJ01002266	10833	10998	-	166
G. lamblia P15	contig371	58049	58214	-	166

```
G. lamblia WB     TGTAAAATAAACTATTTTAAATTCAATTCTGTAAAATAAATTTTTTATTTTTTGACT CTA 60
G. lamblia P15    TGTAAAATTAACCATTTTAAATTCAATTTTGTAAAAATCAGTTTTTGTTTTTTGACTCTA 60
G. lamblia GS     TGTAGAATAAACTATTTTAAATTCAATTCTATTAAACTAATTTCTTTTTTCATGACTCTA 60
                  **** *** *** ************** * * ***   * ** ** ***  ********

G. lamblia WB     GGCTGAAGCTGCCAAGGTGCGTGATCCCTCGGTGATGCCTTGAGTGTTGCTTCACCAAAG 120
G. lamblia P15    GGCTGAAGCTGCCAAGGTGCGTGATCCCTCGGTGATGCCTTGAGTGTTGCTTCACCAAAG 120
G. lamblia GS     GACCGAAGCTGCCAAGGTGCGTGATCCCTCGGTGATGCCTTGAGTGTTGCTTCGCCAAAA 120
                  * * ********************************************** *****

G. lamblia WB     AACAACCACACGGCACAGCCGAATCTCTCATTTTTTAAACTTTCTC 166
G. lamblia P15    AACAACCACACGGCACAGCCGAATCTCTCATTTTTTAAACTCTTCT 166
G. lamblia GS     AACAACCACACGGCCTAGCCGAATTTCTCATTTTTTAAATCTCTCC 166
                  **************  ******** **************
```

32. Gl U6 snRNA

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR04	1813414	1813579	-	166
G. lamblia GS	ACGJ01001509	9049	9214	+	166
G. lamblia P15	contig24	136825	136989	-	165

```
G. lamblia WB     TAAACCATTTTAAATTGAAATAGGCGGTTGGAAATAAAAGCGCGGCGTGGTTAACAAAAA 60
G. lamblia P15    TAAACTATTTTAAATTGAAATAGATGATTAGAAATAAAAGCGTAGCGTGGTTAACAAAAA 60
G. lamblia GS     TAAACAATTTTAAATTAAAATCTACAATTCAGAATAAAAATGGAGCGTGGTTAACAAAAA 60
                  ***** ********** ****      **   *******  *  ****************

G. lamblia WB     CAGAGACAGTTAGCACCAGCTTCAGTCTAGAGTCGCTGGGGGACCTCTGGTTTCGCGGGA 120
G. lamblia P15    CAGAGACAGTTAGCACCAGCTTCAGTCTAGAGTCGCTGGGAGACCTCTGGTTTCGCGGGA 120
G. lamblia GS     CAGAGACAGTTAGCACCAGCTTCGGTCTAGAGTCGCTGGGGAACCTCTGGTTTCGCGGGA 120
                  ********************** ****************  ******************

G. lamblia WB     GCCCGTTGGCGCGTGCTTGCACCCCGCTCCTTTTCTCAATCTTCGC 166
G. lamblia P15    GCCCGTTGGCGCGTGCTTGCACCCCGCTCCTTTTCTCAATCCT-GC 165
G. lamblia GS     GCCTGTTGGCGCGTGCTTGCACCCCGCTCCTTTTCTCAATTGCTGC 166
                  *** ********************************    **
```

33. RNase MRP RNA

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR01	479835	480000	+	166
G. lamblia GS	ACGJ01002287	25852	26017	+	166
G. lamblia P15	contig818	88174	88339	-	166

```
G. lamblia WB     TCCCTGGGCGTCGGGCAGAAAGTGCCGGTCCTCTGGATTCCGGGGAGTGTCTGGTGCCGA 60
G. lamblia P15    TCCCTGGGCGTTGGGCAGAAAGTGCCGGTCCTCTGGACTCCGGGGAGTGTCTGGTGCCGA 60
G. lamblia GS     TCCCTGGGCGTTGGGCAGAAAGTGCCAGTCCTCTGGACTCCGGGGGGTGTCTGGTGCTAA 60
                  *********** ************** ********** ******* ***********  *

G. lamblia WB     TCGGACACTCCCTAGCCGCCACACTGACAGTTATGGTTGCAGGACAAGCTTAGCGAGTCC 120
G. lamblia P15    TCGGACACTCCCCAGCCGCCACACTGACAGTTATGGTTGCAGGACGAGCTTGGCGAGTCC 120
G. lamblia GS     TCGGACACCCCCTAGCCGCCACACTGACAGTTATGGTTGCAAGACGAGCTTAGCGAGTCT 120
                  ******** *** *************************** *** ***** *******

G. lamblia WB     GAACTCGACAGGGATACTCTACAGCGTTCCTTTATTCAATCATTGA 166
G. lamblia P15    GAACTCGATAGGGATACTCTACAGCGTTCCTTTATTCAATCGTTGG 166
G. lamblia GS     GAACTTGACAGGGATACTCTGCGGCGTTCCTTTATTAAATTGCGGG 166
                  ***** ** *********** * ************* ***    *
```

34. Telomerase RNA Candidate (GlsR28)

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR01	978028	978193	-	166
G. lamblia GS	ACGJ01002347	4321	4497	-	177
G. lamblia P15	contig59	153500	153665	+	166

```
G. lamblia WB     AAACTGCACCCTTGTGTTACTC-TGGTGTGTTC TTTATTACCCTACTCTGTCTCGTGAAC 59
G. lamblia P15    AAACTGCACCCTTGCGTTACTC-CGGTGTGTTCTTTATTACCCTACTCTGTCTCGTGAAT 59
G. lamblia GS     AAACTGCACCCTTGCGTCACCCCTGGTGTGTTCTTTATTACCCTACTCTGTCTAGTGATC 60
                  ************** ** ** *  *************************** ****

G. lamblia WB     CCTCACCACCAGTTCTTCTCGCAGTGGTCTCTGTCAAAGACTGCCGCACGGTACACCAGA 119
G. lamblia P15    TCTCACCACCAGTTCTTCTCGCAGTGGTCTCTGTCAAAGACTGCCGCATGGTACACCGGA 119
G. lamblia GS     TCTCACCACCAGTTCTTCTCGCAGTGGTCTCTGTCAAAGACTGCCGCATGGTACACCAGA 120
                   ********************************************* ******** **

G. lamblia WB     AGCAAGGGGAAGGATCCCATCCACGCAGTCCTTTACTTAAACA--------TGGG-- 166
G. lamblia P15    AGCAAGGGGAAGGATCCCATCCACGCAGTCCTTTACTTAAACG--------CGGG-- 166
G. lamblia GS     AGCAAGGGAAAGGAACCCATCCACGCAGTCCTTTACTTAAATAAGACCCTACGGGGG 177
                  ******** ***** *************************           ***
```

NcRNAs with no Assigned Function

35. Candidate-3 [as per reference (Chen *et al.* 2007)]

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR01	702197	702362	-	166
G. lamblia GS	ACGJ01002928	73869	74040	-	172
G. lamblia P15	contig780	40672	40838	+	167

```
G. lamblia WB     CCATTTCAGATAAAGTGGCGCATCTAGAACGGTCAAAAA-GGACCGATCGAAGACCAAGC 59
G. lamblia P15    CCATTTCAAATAAAATTGCATGTCTGGAGTGGCCAAAAA-GGACCGATCGAAGACCAAGC 59
G. lamblia GS     CCATTTCAGATGAAATTACACATCTGGAGCAACCAAAAAAGGATCGATCGAAGACCAGGC 60
                  ******** ** ** *  *   *** **     ****** *** ************ **

G. lamblia WB     GGTGCTAGGTTCAAGCCAGGTCCAAGACCCGGGCAGTCTGTGCTGTGGGGCGCCGCTGTA 119
G. lamblia P15    AGTGCTAGGTTCAAGCCGGGTCCAAGACCTGGGCAGTTTGTGCTGCGGGGCGTTGCTGTA 119
G. lamblia GS     AGTGCTAGGTTCAAACCGTATCCNAGAGTTTAGCAGCCTGTACCGTGGAGTGTTGATGTA 120
                   ************* **   *** ***     ****  *** * * ** * *  * ****

G. lamblia WB     GACGTCTTCCGAACACACCTGCGATAA-A-CCCTTTATTTAAA---AGATTA 166
G. lamblia P15    GACGTCTTTCGAACACACCTGCAACAA-AACCCTTTATTTAAA---AGGTTA 167
G. lamblia GS     GACGTCTTCTGAACACGGCTGCAACAATAATCCTTTATTTAAATCTAAGCTA 172
                  ********  ******  **** * ** *  ************   *   **
```

36. Candidate-5

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR03	299369	299534	+	166
G. lamblia GS	ACGJ01002279	13909	14070	-	162
G. lamblia P15	contig236	1707	1900	-	194

```
G. lamblia WB     GTGAAGGGAAGCAGACTCCATGGCATAAATAAATGCAAAATT---------------CTT 45
G. lamblia P15    GTGAAGGGAAGCGGACTCCATGACATAAATAAATGAAAAATTTCTTAATGTATAATTCTT 60
G. lamblia GS     GTGAAGTGAAGCAGACTCCATGGCATAAAAAA--GCAAATTT---------------CT 42
                  ****** ***** ********* ****** **  * *** **               *

G. lamblia WB     TAACCTGAAAACA------------AAATGGCTAGCAACACGAGGAAACGAGTGTTTCG 92
G. lamblia P15    TAATCTAAAAACATAACCCGGAAATAAAATGGCCAGCAACACGAGGGAACGAGTGTTCTG 120
G. lamblia GS     CAGTCC-AAAATA------------AAATGGCTAAGAGCATGAGGAAACGACTGCTCTG 88
                   *  *  **** *             ******* *  * ** **** ***** ** *  *

G. lamblia WB     CCGGGCATAACTGGGCATGCATTTTCCTTGCCCAGTCTGCCTCCATACTAATTTCTCCTT 152
G. lamblia P15    CCGGGCATAACTGGGCATGCATTTTCCTTGCCCAGTCTGCCCTTGTATTAATTTCTCCTT 180
G. lamblia GS     CCGGGCATAACTGGGCATGCATTTTCCTTGCTCAGTCTGCCTCCTTACAAATTTCTCCTT 148
                  ****************************** *********    **  ***********

G. lamblia WB     TACTCAATCAGGAT 166
G. lamblia P15    TATTAAATCGAAGT 194
G. lamblia GS     TAATAAATCGAAGT 162
                  ** * ****    *
```

37. Candidate-12

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR03	474659	474824	-	166
G. lamblia GS	ACGJ01002930	21466	21613	-	148
G. lamblia P15	contig34	18717	18871	+	155

```
G. lamblia WB     TTCTTAGTCTCTTCTTAGCATCCAGAATAAATCACATTAAATGTATTTTAATTTGAATTT 60
G. lamblia P15    -----------TTCTTAGTATCCAGAATAAATTAAATTAAATGTATTTTAATTTGACTTT 49
G. lamblia GS     -----------TTCTCACTCACCAGAGTAAATTACAC-AAATGTATTTTGATTTGACTTC 48
                             **** *    ***** ***** * *  ********** ****** **

G. lamblia WB     TGATCCCCCGAGAAAAAGAACCCCAACCCGATGACGAATAGCTGTCCTGGCGGAGGCGGT 120
G. lamblia P15    TGATCCTTCAAGAAAAAGGGTCCCAACCCGATGACGAGTAGCTGTCCTGGCGGAGGCGGT 109
G. lamblia GS     CGGTATCCCGAGAAAAGAAAGCCCAACCCGATGACGAATAGCTGTCCTGGCGGAGGCGGT 108
                   * *    * ******     ********** ********** **********

G. lamblia WB     CATGACGACGAAGCCAT-CACGTAGGATCCCTTCACTCAACCTCTGC 166
G. lamblia P15    CATGACGACGATGCCAT-TACGTAGGATTCCTTCACTCAACCTTTGC 155
G. lamblia GS     CATGACGACAATGCCATATGCGTAGGTTCCCTTTACTCAA------- 148
                  ********* * *****   ****** * **** ******
```

38. Candidate-15

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR02	350512	350677	+	166
G. lamblia GS	ACGJ01002258	38250	38415	+	166
G. lamblia P15	contig571	3380	3545	-	166

```
G. lamblia GS     CGGCCTCTGGCTTGGACCCCGTGGCGCTGTCGGCCCCCGCGGAGGCAGGGGTTGGCTCGT 60
G. lamblia P15    TGGCCTCTGGCTTGGACCCCGTGGCGTTGTCGGCCCCCGCGGAGGCAGGGGCCGGCCCGC 60
G. lamblia WB     CGGCCTCTGGCTTGGACCCCGTGGCGTCGCCGGCCTCCGCGGAGGCAGGGGCCGGCCCGC 60
                   ************************  * ***** **************  *** **

G. lamblia GS     CTTCAACTCAGCTGGACAGCCGGAGGCCGGAGACGGAGCACGGTCAGGCGGGCGGGGTGC 120
G. lamblia P15    CTTCAACTCAGCTGGACAGCCGGAGGCCGGAGACGGAGCACGGTCAGGCGGGCGGGGTGC 120
G. lamblia WB     CTTCAACTCAGCCGGACAGCCGGAGGCCGGAGACGGAGCACGGTCAGGCGGGCGGGGTGC 120
                  ************ ********************************************

G. lamblia GS     AGTGCCAGCCTCAGTCGCAGAGCGGCTTCCTTTACTCAAGATCGGG 166
G. lamblia P15    AGTGCCAGCCTCAGTCGCAGAGCGGCTTCCTTTACTCAAGATCGGG 166
G. lamblia WB     AGTGCCAGCCCCAGCCGCAGAGCGGCTTCCTTTACTCAAGATCGGG 166
                  ********** *** ******************************
```

39. Candidate-17

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR03	1601283	1601448	-	166
G. lamblia GS	ACGJ01002096	4585	4750	-	166
G. lamblia P15	contig753	29559	29724	-	166

```
G. lamblia WB     GAGTTAATACCACCAAACCCCTGTGCGTACATGTCGCCCCCTAACCTTCTGATGCGGATA 60
G. lamblia GS     GAGTTAATACCACCAAACCCCTGTGCGTACATGTCGCCCCCTAACCTTCTGATGCGGATA 60
G. lamblia P15    GAGTTAACACCACCAAACCCCTGTGCGTACATGTCGCCCCCTAACCTTCTGATGCGGATA 60
                  ******* **************************************************

G. lamblia WB     CCTTGCCGCAGGGCCGTTAAGCGAGGCTTGGCCCGTGCGACGATGAGGCTCCCTGCGGGG 120
G. lamblia GS     CCTTGCCGCAGGGCCGTTAAGCGAGGCTTGGCCCGTGCGACGATGAGGCTCCCCACGGGG 120
G. lamblia P15    TCTTGCCGCAGGGCCGTTAAGCGAGGCTTGGCCTGTGCGACGATGAGGCTCCCTGCGGGG 120
                   ******************************** ******************  *****

G. lamblia WB     AAGCCCTGCGGCGCGTCTTAAGGAGGCTCCTTCACTCAACGGCGTC 166
G. lamblia GS     AAGCCCTGCGGCGCGTCTTAAGGAGGCTCCTTCACTCAACGGCGTC 166
G. lamblia P15    AAGCCCTGCGGCTCGTCTTAAGGAGGCTCCTTCACTCAATGGCGTC 166
                  ************ ************************* ******
```

40. Candidate-21

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR05	1896857	1897022	-	166
G. lamblia GS	ACGJ01002272	33335	33498	+	164
G. lamblia P15	contig632	33226	33399	+	174

```
G. lamblia WB     CGCCAGCTATTCTACGTCTGTGGCCGTTCTGGCTGCGCTGGACGATGAACTGGAGATGCT 60
G. lamblia P15    CACCAGCTATTCTACGTCTGTGGCCGTTCTGGCTGCGCTAGGCGATGAACTGGAGATGTT 60
G. lamblia GS     CACCAGCTATTCTACGCCTGTGGCCGTTCTGGTTGCGCTGGGTGTTGAACAGGTGATG-T 59
                  * ************** ************** ****** *  * ***** ** **** *

G. lamblia WB     GGACACGGCTTTGCTCTCCCACCGGAGCACATATGCTGCAGGATGACCGGCGCCTGTCTC 120
G. lamblia P15    GGACACGGCTTTGCTCTCCCACCGGAGAACATATGCTGCAGAACGGACGGCACCTGTCCC 120
G. lamblia GS     GTACACGACTCTGCTCTCCCACTGGAGTACACCTGCTGTAGAACGGACGGTACCTGTCCC 119
                  * ***** ** ********** **** ***  ***** ** * *  ***  ****** *

G. lamblia WB     CCACCACGTGCCAGCTAAACTGCAGCC--------ACATTTATTCAACCTTTTC 166
G. lamblia P15    CCACCATGTGCCAGATAAACTGCAGCAACATTACAACATTTATTCAACCTCTTC 174
G. lamblia GS     CCACTGTGTGCCAGCTAAATTACAGCA--------ACATTTACTCAACCCTTC- 164
                  ****   ******* **** * **** *****   ******* ******  *
```

Trans-spliced Intron 5′ Halves

41. Hsp90 Exon 1-Intron 5′ half

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR05	2515245	2515410	+	166
G. lamblia GS	ACGJ01002286	15902	16067	-	166
G. lamblia P15	contig421	16140	16305	-	166

```
G. lamblia WB     TCGACGGCGCCGCCCAGTTCCGCGGGATCCTCTTCATCCCCAAGCGCGCGCCCTTCGACA 60
G. lamblia P15    TAGACGGCGCCGCTCAGTTCCGCGGTATCCTCTTCATCCCCAAGCGCGCGCCCTTCGACA 60
G. lamblia GS     TCGATGGCGCCGCACAGTTCCGCGGCATCCTCTTCATCCCTAAGCGCGCGCCCTTCGACA 60
                  * ** ******** *********** ************** ******************

G. lamblia WB     TGTGGGACGCTCAGAAGAAGAAGACGGGCATCAAGCTCATGGTCAAGAAAGTGTATGTT- 119
G. lamblia P15    TGTGGGACGCTCAAAAGAAGAAGACGGGCATCAAGCTCATGGTCAAGAAAGTGTATGTT- 119
G. lamblia GS     TGTGGGACGCTCAGAAGAAGAAGACGGGCATCAAGCTCATGGTCAAGAAAGTGTATGTTT 120
                  ************* ******************************************

G. lamblia WB     ATGTTTGTATGCTGTATGTGTGCGAGACTCCTTTACTCAAATTTGCG 166
G. lamblia P15    ATGTTTGTATGCTGTATGTGTGCGAGACTCCTTTACTCAAATTCGCG 166
G. lamblia GS     ATGTATGTGTGTTGTATGTGTGCGAGACTCCTTTACTCAA-TTCGTG 166
                  **** *** ** *************************** ** * *
```

42. DHC β Exon 2-Intron 5′ half

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR03	577578	577743	+	166
G. lamblia GS	ACGJ01002314	1045	1211	-	167
G. lamblia P15	contig30	114285	114449	-	165

```
G. lamblia WB     GCCCGTAAGAAATACAAAGCTGTCATGGACAACGTTGCAAAGCTGAACAAAAAGCTCCAA 60
G. lamblia P15    GCCCGTAAGAAGTACAAAGCTGTCATGGACAACGTTGCAAAGCTGAACAAGAAGCTTCAA 60
G. lamblia GS     GCCCGTAAGAAGTACAAAGCCGTCATGGACAACGTTTCGAAGCTCAACAAGAAGCTCCAA 60
                  *********** ******** *************** * ***** ***** ***** ***

G. lamblia WB     ACTCTCAAGGATCAATTTGACAAGGGTATGTTACTGGTGAAACGCTACTTATGTATGTA 120
G. lamblia P15    ACTCTCAAGGATCAATTTGACAAGGGTATGTTACTAGGTGAAACGCTACTTATGTGTGTA 120
G. lamblia GS     ACTCTCAAAGACCAATTCGACAAGGGTATGTTACC-GGTGAAACGCTACTTATGTGTGTA 119
                  ******** ** ***** ***************  ****************** ****

G. lamblia WB     TGCTTATATGT-CTTCGCGCTCAGGCGCTCCTTTACTCAAT-TATCAG 166
G. lamblia P15    TGTCTATATGT-CT-CGCGCTCGGGCGCTCCTTTACTCAAT-TATCAA 165
G. lamblia GS     TGTCTATATGTTCTTCGCGCTTAGGCGCTCCTTTACTCAAAACACCAA 167
                  **  ******* ** ******  *****************   * **
```

43. DHC β Exon 3-Intron 5′ half

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR05	4266308	4266473	+	166
G. lamblia GS	ACGJ01002906	4407	4562	+	156
G. lamblia P15	contig39	71963	72134	+	172

```
G. lamblia WB     GTTTGCTCTGATCTTCTTCCATGCAATCGTCATCGAGAGACGGAAGTTCGGTCCTATAGG 60
G. lamblia P15    GTTTGCTCTGATTTTCTTCCATGCAATTGTCATCGAGAGACGAAAGTTCGGTCCTATAGG 60
G. lamblia GS     GTTTGCTTTAATATTCTTCCATGCAATCGTCATCGAGAGACGAAAATTCGGTCCCATAGG 60
                  ******* * ** ************** ************** ** ******** *****

G. lamblia WB     GTATGTTTGTAATCTGTGTAGTCGCAGTATGCCATTATTTTATAACGTGTATGTCATTAT 120
G. lamblia P15    GTATGTTCGTAATCTGTGTAGTTGCAGTATGCCATTGTTTTATAACGTGTATGTCACTAT 120
G. lamblia GS     GTATGTTCACG--CTGTGTAGT-GCAGTATGCCATT--TTTAAA-----TATGTTAGTAT 110
                  *******      ********* *************  **** *     ***** * ***

G. lamblia WB     GTCAGTATGCCAGTGCGCTGGTG------AGTTTCCTTTATTCAAATGTTGT 166
G. lamblia P15    GTCAGTATGTCAGTACGCCAGTACAAGTAATTTTCCTTTACTCAAATATTTT 172
G. lamblia GS     GTTAATATGTCAGTACGCTAGTAC------ATTTCCTTTACTCAAGTGCTAT 156
                  ** * **** **** ***  **         ********* **** *  * *
```

44. DHC γ Exon 1-Intron 5′ half

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR03	967465	967630	+	166
G. lamblia GS	ACGJ01002918	40974	41139	+	166
G. lamblia P15	contig38	36690	36855	-	166

```
G. lamblia GS     TCTTGCGGTATCTTCATCACGATGAACCCCGGTTACGCCGGGCGTCAAGAACTTCCAGAG 60
G. lamblia P15    TCTTGCGGTATCTTCATCACGATGAACCCCGGTTACGCCGGGCGTCAGGAACTTCCAGAG 60
G. lamblia WB     TCTTGCGGTATTTTCATCACGATGAACCCCGGTTACGCCGGGCGTCAAGAACTTCCAGAG 60
                  *********** ********************************** ************

G. lamblia GS     AATCTCAAAGCTTTATTCCGTAGTGTTGCAATGATATGTTTACAGGTGGTTCGGTGTGTA 120
G. lamblia P15    AATCTCAAAGCCTTATTCCGTAGCGTTGCAATGATATGTTTACAGGTGGTTCGGTGTGTA 120
G. lamblia WB     AATCTCAAAGCCTTATTCCGTAGCGTTGCAATGATATGTTCACAGGTGGTTTGGTGTGTA 120
                  *********** *********** ***************** ********** ********

G. lamblia GS     TGCTTGGCGTGTATGTGTGTATGTCCTTCCTTTACTCAATGCCCAG 166
G. lamblia P15    TGCTTGGCGTGTATGTGTGTATGTCCTTCCTTTACTCAATGCTTGG 166
G. lamblia WB     TGCTTGGCGTGTATGTGTGTATGTTCCTCCTTTACTCAATACTTGG 166
                  ********************** * ************* *   *
```

Table A.2.1. Motif sequence variants in *Giardia* WB, P15 and GS isolates

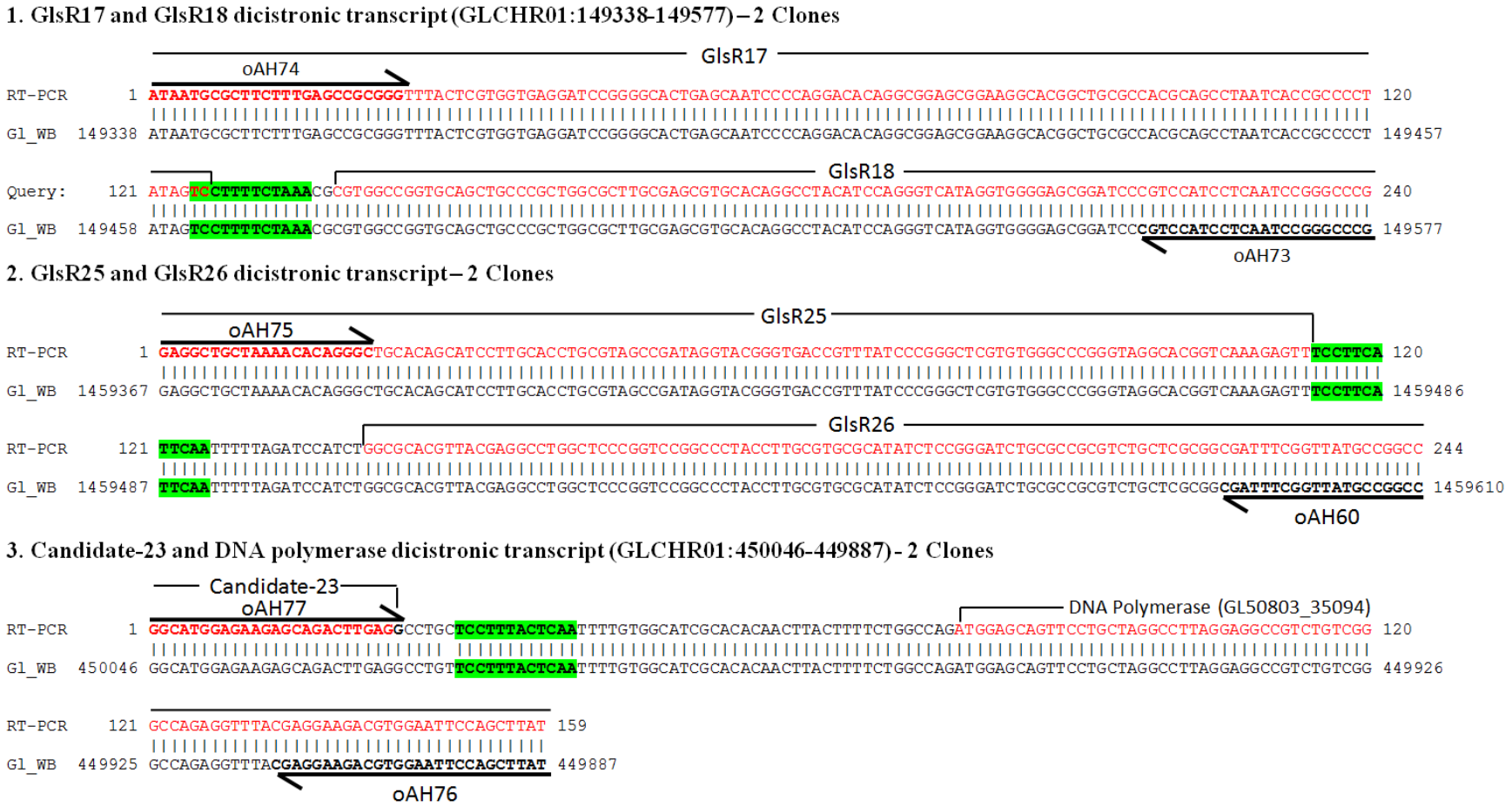
Occurrences of variant motif sequences in ncRNA and *trans*-spliced intron containing genes in *G. lamblia* WB, P15 and GS isolates (132 total motif instances) identified in this study are annotated in descending order of observed frequency. The consensus 'TCCTTTACTCAA' motif sequence was observed 34 times and nucleotides differing in motif variants are highlighted in bold red text.

No.	Motif Variant	WB	P15	GS	Frequency
1	TCCTTTACTCAA	GlsR7 GlsR15 Candidate-1 Candidate-23 GlsR18 Candidate-16 Candidate-15 Hsp90 Exon 1 DHC β Exon 2 DHC γ Exon 1 Candidate-5	GlsR2 GlsR7 GlsR15 Candidate-1 Candidate-23 Candidate-16 Candidate-15 Hsp90 Exon 1 DHC β Exon 2 DHC β Exon 3 DHC γ Exon 1	GlsR2 GlsR5 GlsR7 GlsR15 GlsR18 GlsR20 GlsR25 Candidate-15 Hsp90 Exon 1 DHC β Exon 2 DHC β Exon 3 DHC γ Exon 1	34
2	TCCTTTACT**T**AA	GlsR2 Candidate-13 GlsR20 GlsR28	Candidate-13 GlsR20 GlsR23 GlsR28	GlsR28	9
3	TCCTT**C**ACTCAA	GlsR5 GlsR6 Candidate-17	GlsR6 Candidate-12 Candidate-17	GlsR6 GlsR8 Candidate-17	9
4	**A**CCTTTACTCAA	GlsR22 GlsR24	GlsR22 GlsR24	GlsR22 GlsR24 Gl U2 snRNA	7
5	TCCTTT**T**CTCAA	GlsR10 Gl U6 snRNA	GlsR10 Gl U6 snRNA	Gl U6 snRNA	5
6	TCCTTTA**T**TCAA	RNase MRP DHC β Exon 3	GlsR18 RNase MRP	Candidate-23	5
7	**A**C**A**TTTA**T**TCAA	GlsR21 Candidate-21	GlsR21 Candidate-21	GlsR21	5
8	TCCTTTA**A**T**A**AA	GlsR9	GlsR9	GlsR9 Candidate-5	4
9	**C**CCTTTACTCAA	GlsR23		GlsR13 Candidate-1 Candidate-12	4
10	TCCTT**C**A**T**TCAA	GlsR25 GlsR26	GlsR25	GlsR26	4
11	TCCTT**C**A**T**T**T**AA	GlsR1	GlsR1	GlsR1	3
12	**C**CCTT**C**ACTCAA	GlsR8 Candidate-12	GlsR8		3
13	TCCTT**C**AA**T**T**A**A	GlsR19	GlsR19	GlsR19	3
14	TC**A**TTT**T**T**A**AA	Gl U4 snRNA	Gl U4 snRNA	Gl U4 snRNA	3
15	TC**A**TT**C**ACTCAA	GlsR4	GlsR4		2
16	**C**CCTTTACT**T**AA	GlsR13	GlsR13		2
17	**C**CCTTTA**T**TCAA			GlsR14 Candidate-13	2

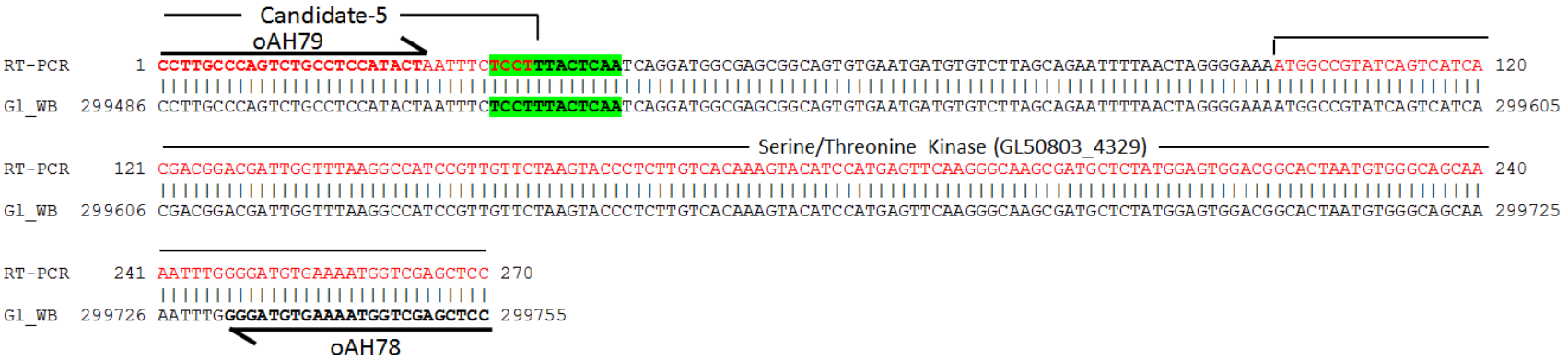
Table A.2.1 (continued)

No.	Motif Variant	WB	P15	GS	Frequency
18	TCCTTTGCTCAA	Candidate-2	Candidate-2		2
19	TCCTTTTCTAAA	GlsR17	GlsR17		2
20	TCCTTTATTTAAA		Candidate-5	RNase MRP	2
21	CCCTTTATTTTAA	Candidate-3	Candidate-3		2
22	TCCTTCACTTAA	Gl U1 snRNA	Gl U1 snRNA		2
23	ACCTTTATTCAA	Gl U2 snRNA	Gl U2 snRNA		2
24	TCATTTACTCAA			GlsR4	1
25	ACATTTACTCAA			Candidate-21	1
26	ACCTTCACTTAA		GlsR27		1
27	ACCTTCATTTAA	GlsR27			1
28	ACCTTTATTTAA			GlsR27	1
29	CCCTTTGCTCAA		GlsR14		1
30	GCCTTTACTCAA			GlsR23	1
31	TCCTTCGCTCAA			Gl U1 snRNA	1
32	TCCTTCGCTTAA			GlsR17	1
33	TCCTTCGTTCAA		GlsR26		1
34	TCCTTCTCTCAA			GlsR10	1
35	TCCTTTAATCAA			Candidate-16	1
36	TCCTTTACTAAA		GlsR5		1
37	TCCTTTATTTAA			Candidate-3	1
38	TCCTTTCCTCAA	GlsR14			1
39	TCCTTTTCTTAA			Candidate-2	1
Total Motif Occurrences					**132**

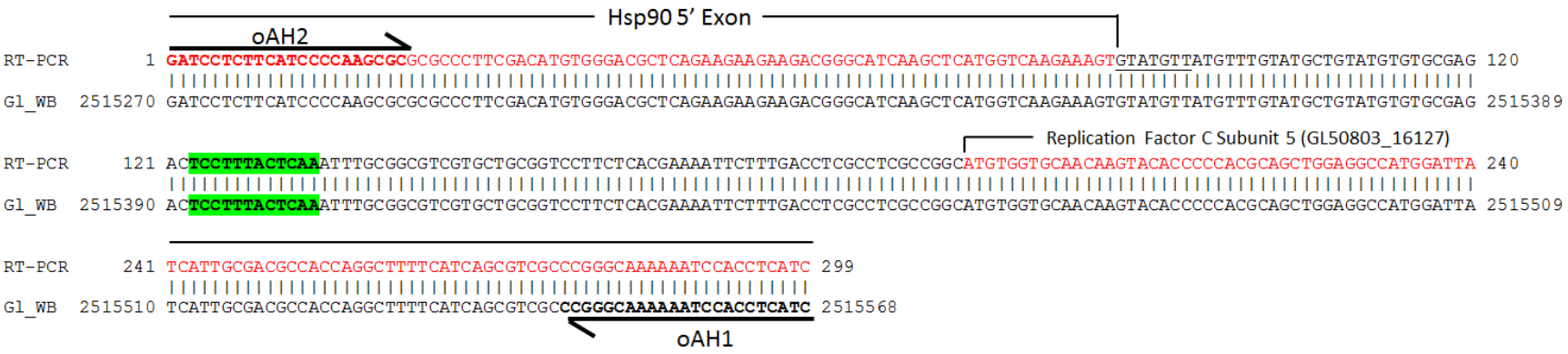
Figure A.2.2. RT-PCR detection of motif-containing ncRNA and *trans*-spliced intron precursor transcripts.
RT-PCR product sequences (top) are compared with *Giardia* WB isolate genomic DNA sequences (bottom, Gl_WB). Genomic locations of amplified regions are given as well as the number of unique clones that were isolated and sequenced. Regions specifying ncRNA or protein coding sequences are indicated in red text with motif sequences highlighted in green. *Trans*-spliced intron 5′ splice sites are underlined. Primer annealing sites are indicated with arrows that denote the direction of amplification during PCR.



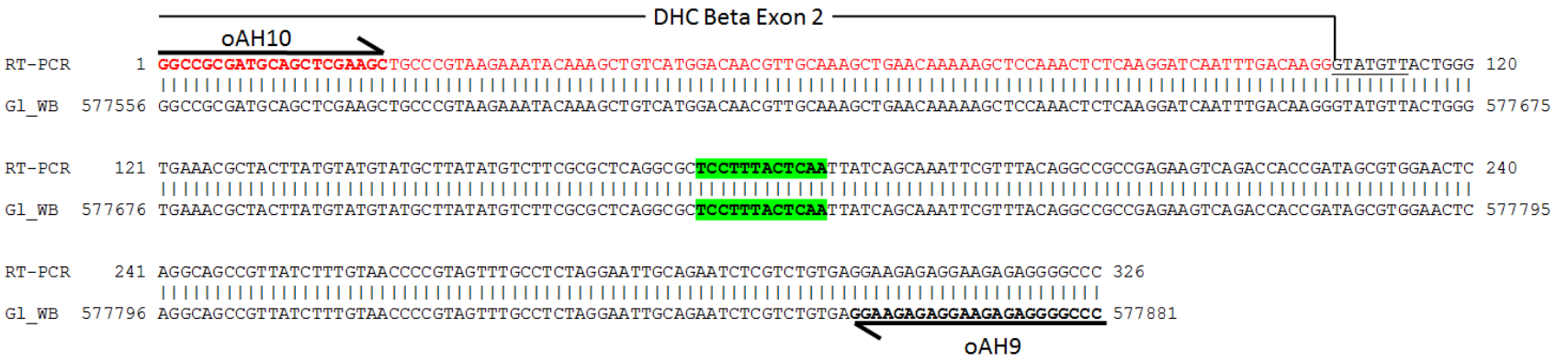
4. Candidate-5 and Serine/Threonine Kinase (GLCHR03: 299486-299755) – 2 Clones



5. Hsp90 Exon-Intron 5' half and Replication Factor C Subunit 5 dicistronic transcript (GLCHR05:2515270-2515568) – 2 Clones



6. DHC Beta Exon 2-Intron 1 5' half (GLCHR03:577556-577881) – 2 Clones



7. DHC Beta Exon 3-Intron 2 5' half (GLCHR05:4266265-4266624) - 2 clones

Figure A.2.3. 5′ and 3′ RACE analysis of *Giardia* ncRNAs and *trans*-spliced introns.
Giardia WB isolate genomic sequences are shown with sequencing results from 5′ and 3′ RACE experiments overlaid in red text and motif sequences highlighted in green. *Trans*-spliced intron 5′ exons are in grey and the 5′ splice sites are highlighted in blue. Names of primers used for RACE experiments are indicated to the left of sequences and indicate whether they were used for 3′ (3′R) or 5′ (5′R) RACE analysis. Primer annealing positions for RACE experiments are underlined. Sequences reported by Chen *et al*. 2007 are also shown.

Box H/ACA RNAs

GlsR26 — GLCHR05:1459484-1459636

```
>ODE3 3′R #1 TCATTCAATTTTTAGATCCATCTGGCGCACGTTACGAGGCCTGGCTCCCGGTCCGGCCCTACCTTGCGTGCGCATATCTCCGGGATCTGCGCCGCGTCTGCTCGCGGCGATTTCGGTTATGCCGGCCCGAACACTCCTTCATTCAACAGGCCC
>ODE3 3′R #2 TCATTCAATTTTTAGATCCATCTGGCGCACGTTACGAGGCCTGGCTCCCGGTCCGGCCCTACCTTGCGTGCGCATATCTCCGGGATCTGCGCCGCGTCTGCTCGCGGCGATTTCGGTTATGCCGGCCCGAACACTCCTTCATTCAACAGGCCC
>oAH60 5′R #1TCATTCAATTTTTAGATCCATCTGGCGCACGTTACGAGGCCTGGCTCCCGGTCCGGCCCTACCTTGCGTGCGCATATCTCCGGGATCTGCGCCGCGTCTGCTCGCGGCGATTTCGGTTATGCCGGCCCGAACACTCCTTCATTCAACAGGCCC
>oAH60 5′R #2TCATTCAATTTTTAGATCCATCTGGCGCACGTTACGAGGCCTGGCTCCCGGTCCGGCCCTACCTTGCGTGCGCATATCTCCGGGATCTGCGCCGCGTCTGCTCGCGGCGATTTCGGTTATGCCGGCCCGAACACTCCTTCATTCAACAGGCCC
```

GlsR27 — GLCHR01:1371253-1371405

```
>ODE4 3′R #1 TTCAGCATTTATAACCAAAAAATTAAGCTCACCCAAAGTCAACGGAGCGCCAGCTACGTGTTATGGGCAGCGAAAGTACCAGAGCCAAAGAGTTCCTCTGATCGCCTGGCCGGAGCACATTTGTGATCTCCTATACCTTCATTTAATTAGCGT
>ODE4 3′R #2 TTCAGCATTTATAACCAAAAAATTAAGCTCACCCAAAGTCAACGGAGCGCCAGCTACGTGTTATGGGCAGCGAAAGTACCAGAGCCAAAGAGTTCCTCTGATCGCCTGGCCGGAGCACATTTGTGATCTCCTATACCTTCATTTAATTAGCGT
>OAH62 5′R #1TTCAGCATTTATAACCAAAAAATTAAGCTCACCCAAAGTCAACGGAGCGCCAGCTACGTGTTATGGGCAGCGAAAGTACCAGAGCCAAAGAGTTCCTCTGATCGCCTGGCCGGAGCACATTTGTGATCTCCTATACCTTCATTTAATTAGCGT
>OAH62 5′R #2GTCAGCATTTATAACCAAAAAATTAAGCTCACCCAAAGTCAACGGAGCGCCAGCTACGTGTTATGGGCAGCGAAAGTACCAGAGCCAAAGAGTTCCTCTGATCGCCTGGCCGGAGCACATTTGTGATCTCCTATACCTTCATTTAATTAGCGT
>OAH62 5′R #3TTCAGCATTTATAACCAAAAAATTAAGCTCACCCAAAGTCAACGGAGCGCCAGCTACGTGTTATGGGCAGCGAAAGTACCAGAGCCAAAGAGTTCCTCTGATCGCCTGGCCGGAGCACATTTGTGATCTCCTATACCTTCATTTAATTAGCGT
>OAH62 5′R #4TTCAGCATTTATAACCAAAAAATTAAGCTCACCCAAAGTCAACGGAGCGCCAGCTACGTGTTATGGGCAGCGAAAGTACCAGAGCCAAAGAGTTCCTCTGATCGCCTGGCCGGAGCACATTTGTGATCTCCTATACCTTCATTTAATTAGCGT
```

NcRNAs without Assigned Function

Candidate-5 — GLCHR03:299356-299534

```
>CHEN ET AL. GACTCCATGGCATAAATAAATGCAAAATTCTTTAACCTGAAAACAAAATGGCTAGCAACACGAGGAAACGAGTGTTTCGCCGGGCATAACTGGGCATGCATTTTCCTTGCCCAGTCTGCCTCCATACTAATTTCTCCTTTACTCAATCAGGAT
>ODE8 3′R #1 GACTCCATGGCATAAATAAATGCAAAATTCTTTAACCTGAAAACAAAATGGCTAGCAACACGAGGAAACGAGTGTTTCGCCGGGCATAACTGGGCATGCATTTTCCTTGCCCAGTCTGCCTCCATACTAATTTCTCCTTTACTCAATCAGGAT
>ODE8 3′R #2 GACTCCATGGCATAAATAAATGCAAAATTCTTTAACCTGAAAACAAAATGGCTAGCAACACGAGGAAACGAGTGTTTCGCCGGGCATAACTGGGCATGCATTTTCCTTGCCCAGTCTGCCTCCATACTAATTTCTCCTTTACTCAATCAGGAT
```

Candidate-12 — GLCHR03:474672-474824

```
>CHEN ET AL. CTTAGCATCCAGAATAAATCACATTAAATGTATTTTAATTTGAATTTTGATCCCCCGAGAAAAAGAACCCCAACCCGATGACGAATAGCTGTCCTGGCGGAGGCGGTCATGACGACGAAGCCATCACGTAGGATCCCTTCACTCAACCTCTGC
>ODE2 3′R #2 CTTAGCATCCAGAATAAATCACATTAAATGTATTTTAATTTGAATTTTGATCCCCCGAGAAAAAGAACCCCAACCCGATGACGAATAGCTGTCCTGGCGGAGGCGGTCATGACGACGAAGCCATCACGTAGGATCCCTTCACTCAACCTCTGC
```

Candidate-14 — GLCHR04:581409-581561

```
>CHEN ET AL. AATCAGAGTCGGCTTCGACTTTAGTGTAGTTACTGTTTCGTCGGCTTAACCGCCGATCCACTACATGCAAGGGGCAGCCGGGCTGTGAGGCAGCTGCCAGGATGGTCCTGCCCTTGTCCCGGCTGGCGCCGTCCACCTTTATTCAAGTTTTCT
>ODE9 3′R #1 AATCAGAGTCGGCTTCGACTTTAGTGTAGTTACTGTTTCGTCGGCTTAACCGCCGATCCACTACATGCAAGGGGCAGCCGGGCTGTGAGGCAGCTGCCAGGATGGTCCTGCCCTTGTCCCGGCTGGCGCCGTCCACCTTTATTCAAGTTTTCT
>ODE9 3′R #2 AATCAGAGTCGGCTTCGACTTTAGTGTAGTTACTGTTTCGTCGGCTTAACCGCCGATCCACTACATGCAAGGGGCAGCCGGGCTGTGAGGCAGCTGCCAGGATGGTCCTGCCCTTGTCCCGGCTGGCGCCGTCCACCTTTATTCAAGTTTTCT
```

Candidate-15 — GLCHR02:350525-350677

```
>CHEN ET AL. GGACCCCGTGGCGTCGCCGGCCTCCGCGGAGGCAGGGGCCGGCCCGCCTTCAACTCAGCCGGACAGCCGGAGGCCGGAGACGGAGCACGGTCAGGCGGGCGGGGTGCAGTGCCAGCCCCAGCCGCAGAGCGGCTTCCTTTACTCAAGATCGGG
>ODE5 3′R #1 GGACCCCGTGGCGTCGCCGGCCTCCGCGGAGGCAGGGGCCGGCCCGCCTTCAACTCAGCCGGACAGCCGGAGGCCGGAGACGGAGCACGGTCAGGCGGGCGGGGTGCAGTGCCAGCCCCAGCCGCAGAGCGGCTTCCTTTACTCAAGATCGGG
>ODE5 3′R #1 GGACCCCGTGGCGTCGCCGGCCTCCGCGGAGGCAGGGGCCGGCCCGCCTTCAACTCAGCCGGACAGCCGGAGGCCGGAGACGGAGCACGGTCAGGCGGGCGGGGTGCAGTGCCAGCCCCAGCCGCAGAGCGGCTTCCTTTACTCAAGATCGGG
```

Candidate-17 — GLCHR03:1601296-1601448

```
>CHEN ET AL. CAAACCCCTGTGCGTACATGTCGCCCCCTAACCTTCTGATGCGGATACCTTGCCGCAGGGCCGTTAAGCGAGGCTTGGCCCGTGCGACGATGAGGCTCCCTGCGGGGAAGCCCTGCGGCGCGTCTTAAGGAGGCTCCTTCACTCAACGGCGTC
>ODE1 3′R #1 CAAACCCCTGTGCGTACATGTCGCCCCCTAACCTTCTGATGCGGATACCTTGCCGCAGGGCCGTTAAGCGAGGCTTGGCCCGTGCGACGATGAGGCTCCCTGCGGGGAAGCCCTGCGGCGCGTCTTAAGGAGGCTCCTTCACTCAACGGCGTC
>ODE1 3′R #1 CAAACCCCTGTGCGTACATGTCGCCCCCTAACCTTCTGATGCGGATACCTTGCCGCAGGGCCGTTAAGCGAGGCTTGGCCCGTGCGACGATGAGGCTCCCTGCGGGGAAGCCCTGCGGCGCGTCTTAAGGAGGCTCCTTCACTCAACGGCGTC
```

Candidate-18 GLCHR01:479848-480000

```
>CHEN ET AL. GGCAGAAAGTGCCGGTCCTCTGGATTCCGGGGAGTGTCTGGTGCCGATCGGACACTCCCTAGCCGCCACACTGACAGTTATGGTTGCAGGACAAGCTTAGCGAGTCCGAACTCGACAGGGATACTCTACAGCGTTCCTTTATTCAATCATTGA
>ODE10 3'R #1GGCAGAAAGTGCCGGTCCTCTGGATTCCGGGGAGTGTCTGGTGCCGATCGGACACTCCCTAGCCGCCACACTGACAGTTATGGTTGCAGGACAAGCTTAGCGAGTCCGAACTCGACAGGGATACTCTACAGCGTTCCTTTATTCAATCATTGA
>ODE10 3'R #1GGCAGAAAGTGCCGGTCCTCTGGATTCCGGGGAGTGTCTGGTGCCGATCGGACACTCCCTAGCCGCCACACTGACAGTTATGGTTGCAGGACAAGCTTAGCGAGTCCGAACTCGACAGGGATACTCTACAGCGTTCCTTTATTCAATCATTGA
```

Candidate-23 GLCHR01:445011-450163

```
>CHEN ET AL. TTATACCAAAATTAATTGCAGAGTTCTCCTTTTTCAAAAAGCCTCTCTGTAGGTAGGGCCGATGAGCTATTTTGTACCACTCTGACCGTGAGGCGTATGCCTAGGGCATGGAGAAGAGCAGACTTGAGGCCTGTTCCTTTACTCAATTTTGTG
>oDE7 3'R #1 TTATACCAAAATTAATTGCAGAGTTCTCCTTTTTCAAAAAGCCTCTCTGTAGGTAGGGCCGATGAGCTATTTTGTACCACTCTGACCGTGAGGCGTATGCCTAGGGCATGGAGAAGAGCAGACTTGAGGCCTGTTCCTTTACTCAATTTTGTG
>oDE7 3'R #2 TTATACCAAAATTAATTGCAGAGTTCTCCTTTTTCAAAAAGCCTCTCTGTAGGTAGGGCCGATGAGCTATTTTGTACCACTCTGACCGTGAGGCGTATGCCTAGGGCATGGAGAAGAGCAGACTTGAGGCCTGTTCCTTTACTCAATTTTGTG
```

Spliceosomal snRNA Candidates

U1 Candidate GLCHR03:258103-258039

```
>oAH133 3'R#1AATTGAAGGTAATTTTAACTTACCTCAAGGGTGGCGACGAGCCAGTGTTCGGGCCAGGCTGGTGCTGCGCATACCGCGCTGGCACTGGTCACGGGGCAGTGCTCTCAGACCTGCTACCGTACCCTTTTAATTTTTCCTTCACTTAAGGCCCAT
>oAH133 3'R#2AATTGAAGGTAATTTTAACTTACCTCAAGGGTGGCGACGAGCCAGTGTTCGGGCCAGGCTGGTGCTGCGCATACCGCGCTGGCACTGGTCACGGGGCAGTGCTCTCAGACCTGCTACCGTACCCTTTTAATTTTTCCTTCACTTAAGGCCCAT
>oAH95 5'R#1 AATTGAAGGTAATTTTAACTTACCTCAAGGGTGGCGACGAGCCAGTGTTCGGGCCAGGCTGGTGCTGCGCATACCGCGCTGGCACTGGTCACGGGGCAGTGCTCTCAGACCTGCTACCGTACCCTTTTAATTTTTCCTTCACTTAAGGCCCAT
>oAH95 5'R#2 AATTGAAGGTAATTTTAACTTACCTCAAGGGTGGCGACGAGCCAGTGTTCGGGCCAGGCTGGTGCTGCGCATACCGCGCTGGCACTGGTCACGGGGCAGTGCTCTCAGACCTGCTACCGTACCCTTTTAATTTTTCCTTCACTTAAGGCCCAT
>oAH95 5'R#3 AATTGAAGGTAATTTTAACTTACCTCAAGGGTGGCGACGAGCCAGTGTTCGGGCCAGGCTGGTGCTGCGCATACCGCGCTGGCACTGGTCACGGGGCAGTGCTCTCAGACCTGCTACCGTACCCTTTTAATTTTTCCTTCACTTAAGGCCCAT
>oAH95 5'R#4 AATTGAAGGTAATTTTAACTTACCTCAAGGGTGGCGACGAGCCAGTGTTCGGGCCAGGCTGGTGCTGCGCATACCGCGCTGGCACTGGTCACGGGGCAGTGCTCTCAGACCTGCTACCGTACCCTTTTAATTTTTCCTTCACTTAAGGCCCAT
```

Candidate-14 (U2 Candidate) GLCHR04:581409-581561

```
>CHEN ET AL. AATCAGAGTCGGCTTCGACTTTAGTGTAGTTACTGTTTCGTCGGCTTAACCGCCGATCCACTACATGCAAGGGGCAGCCGGGCTGTGAGGCAGCTGCCAGGATGGTCCTGCCCTTGTCCCGGCTGGCGCCGTCCACCTTTATTCAAGTTTTCT
>ODE9 3'R #1 AATCAGAGTCGGCTTCGACTTTAGTGTAGTTACTGTTTCGTCGGCTTAACCGCCGATCCACTACATGCAAGGGGCAGCCGGGCTGTGAGGCAGCTGCCAGGATGGTCCTGCCCTTGTCCCGGCTGGCGCCGTCCACCTTTATTCAAGTTTTCT
>ODE9 3'R #2 AATCAGAGTCGGCTTCGACTTTAGTGTAGTTACTGTTTCGTCGGCTTAACCGCCGATCCACTACATGCAAGGGGCAGCCGGGCTGTGAGGCAGCTGCCAGGATGGTCCTGCCCTTGTCCCGGCTGGCGCCGTCCACCTTTATTCAAGTTTTCT
>OAH120 3'R#1AATCAGAGTCGGCTTCGACTTTAGTGTAGTTACTGTTTCGTCGGCTTAACCGCCGATCCACTACATGCAAGGGGCAGCCGGGCTGTGAGGCAGCTGCCAGGATGGTCCTGCCCTTGTCCCGGCTGGCGCCGTCCACCTTTATTCAAGTTTTCT
>OAH120 3'R#2AATCAGAGTCGGCTTCGACTTTAGTGTAGTTACTGTTTCGTCGGCTTAACCGCCGATCCACTACATGCAAGGGGCAGCCGGGCTGTGAGGCAGCTGCCAGGATGGTCCTGCCCTTGTCCCGGCTGGCGCCGTCCACCTTTATTCAAGTTTTCT
>OAH140 5'R#1AATCAGAGTCGGCTTCGACTTTAGTGTAGTTACTGTTTCGTCGGCTTAACCGCCGATCCACTACATGCAAGGGGCAGCCGGGCTGTGAGGCAGCTGCCAGGATGGTCCTGCCCTTGTCCCGGCTGGCGCCGTCCACCTTTATTCAAGTTTTCT
```

Candidate-11 (U4 Candidate) GLCHR02:1194993-1195145

```
>Chen et al. ATTTTAAATTCAATTCTGTAAAATAAATTTTTTATTTTTTGACTCTAGGCTGAAGCTGCCAAGGTGCGTGATCCCTCGGTGATGCCTTGAGTGTTGCTTCACCAAAGAACAACCACACGGCACAGCCGAATCTCTCATTTTTTAAACTTTCTC
>oAH118 3'R#1ATTTTAAATTCAATTCTGTAAAATAAATTTTTTATTTTTTGACTCTAGGCTGAAGCTGCCAAGGTGCGTGATCCCTCGGTGATGCCTTGAGTGTTGCTTCACCAAAGAACAACCACACGGCACAGCCGAATCTCTCATTTTTTAAACTTTCTC
>oAH118 3'R#2ATTTTAAATTCAATTCTGTAAAATAAATTTTTTATTTTTTGACTCTAGGCTGAAGCTGCCAAGGTGCGTGATCCCTCGGTGATGCCTTGAGTGTTGCTTCACCAAAGAACAACCACACGGCACAGCCGAATCTCTCATTTTTTAAACTTTCTC
>oAH139 5'R#1ATTTTAAATTCAATTCTGTAAAATAAATTTTTTATTTTTTGACTCTAGGCTGAAGCTGCCAAGGTGCGTGATCCCTCGGTGATGCCTTGAGTGTTGCTTCACCAAAGAACAACCACACGGCACAGCCGAATCTCTCATTTTTTAAACTTTCTC
```

U6 Candidate GLCHR04:1813427-1813579

```
>oDE14 3'R#1 ATTGAAATAGGCGGTTGGAAATAAAAGCGCGGCGTGGTTAACAAAAACAGAGACAGTTAGCACCAGCTTCAGTCTAGAGTCGCTGGGGGACCTCTGGTTTCGCGGGAGCCCGTTGGCGCGTGCTTGCACCCCGCTCCTTTTCTCAATCTTCGC
>oDE14 3'R#2 ATTGAAATAGGCGGTTGGAAATAAAAGCGCGGCGTGGTTAACAAAAACAGAGACAGTTAGCACCAGCTTCAGTCTAGAGTCGCTGGGGGACCTCTGGTTTCGCGGGAGCCCGTTGGCGCGTGCTTGCACCCCGCTCCTTTTCTCAATCTTCGC
>oAH72 5'R#1 ATTGAAATAGGCGGTTGGAAATAAAAGCGCGGCGTGGTTAACAAAAACAGAGACAGTTAGCACCAGCTTCAGTCTAGAGTCGCTGGGGGACCTCTGGTTTCGCGGGAGCCCGTTGGCGCGTGCTTGCACCCCGCTCCTTTTCTCAATCTTCGC
>oAH72 5'R#2 ATTGAAATAGGCGGTTGGAAATAAAAGCGCGGCGTGGTTAACAAAAACAGAGACAGTTAGCACCAGCTTCAGTCTAGAGTCGCTGGGGGACCTCTGGTTTCGCGGGAGCCCGTTGGCGCGTGCTTGCACCCCGCTCCTTTTCTCAATCTTCGC
```

Telomerase RNA component (TERC) Candidate (GlsR28)

Trans-Spliced Introns

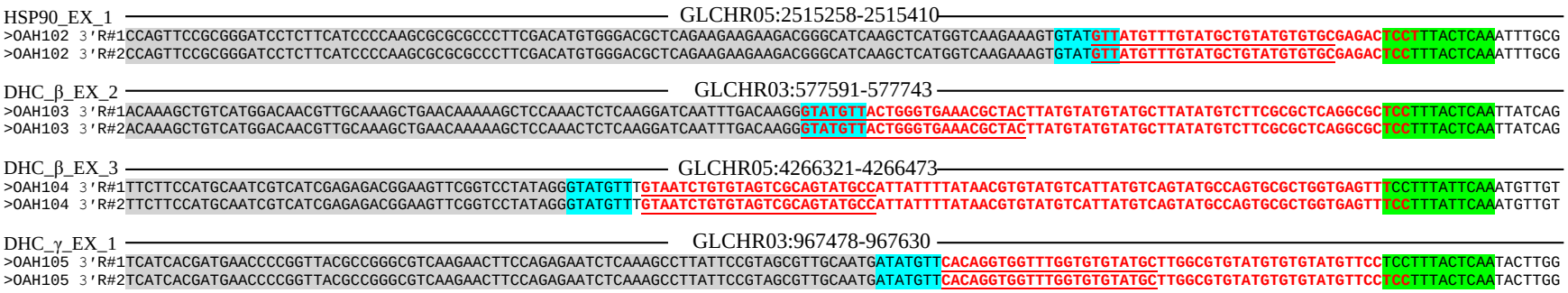


Figure A.2.4. 5′ RACE mapping of regions downstream of RNA motif sequences.
Genomic sequences (WB isolates) encoding 3′ end regions of *G. lamblia* ncRNAs (**A/B**) or *Hsp90 trans*-intron 5′ half (**C**) are shown with red text indicating regions specifying ncRNA/*trans*-introns and nearby downstream protein coding sequences. RNA processing motif sequences are highlighted in green. Reverse primer binding sites used for 5′ RACE are underlined and nucleotide sequences from clones are in bold text. *Giardia* total RNA samples were either (**i**) solely treated with DNase I and used 5′ RACE primers which annealed immediately downstream of motif sequences or (**ii**) treated with DNase I, CIP and TAP and used 5′ RACE primers annealing downstream to detect capped 5′ RACE products for mRNAs encoded downstream of motif sequences (primer binding sites not shown). Potential A-T rich transcription start sites (TSS?) for downstream ORFs are indicated under alignments.

A) Candidate-23 ncRNA and DNA polymerase catalytic subunit delta (GLCHR01:450046- 449927)

i) DNase I treated only

```
                                                                                (oAH219 reverse primer)
>Clone_1   GGCATGGAGAAGAGCAGACTTGAGGCCTGCTCCTTTACTCAATTTTGTGGCATCGCACACAACTTACTTTTCTGGCCAGATGGAGCAGTTCCTGCTAGGCCTTAGGAGGCCGTCTGTCGG
>Clone_2   GGCATGGAGAAGAGCAGACTTGAGGCCTGCTCCTTTACTCAATTTTGTGGCATCGCACACAACTTACTTTTCTGGCCAGATGGAGCAGTTCCTGCTAGGCCTTAGGAGGCCGTCTGTCGG
>Clone_3   GGCATGGAGAAGAGCAGACTTGAGGCCTGCTCCTTTACTCAATTTTGTGGCATCGCACACAACTTACTTTTCTGGCCAGATGGAGCAGTTCCTGCTAGGCCTTAGGAGGCCGTCTGTCGG
>Clone_4   GGCATGGAGAAGAGCAGACTTGAGGCCTGCTCCTTTACTCAATTTTGTGGCATCGCACACAACTTACTTTTCTGGCCAGATGGAGCAGTTCCTGCTAGGCCTTAGGAGGCCGTCTGTCGG
>Clone_5   GGCATGGAGAAGAGCAGACTTGAGGCCTGCTCCTTTACTCAATTTTGTGGCATCGCACACAACTTACTTTTCTGGCCAGATGGAGCAGTTCCTGCTAGGCCTTAGGAGGCCGTCTGTCGG
>Clone_6   GGCATGGAGAAGAGCAGACTTGAGGCCTGCTCCTTTACTCAATTTTGTGGCATCGCACACAACTTACTTTTCTGGCCAGATGGAGCAGTTCCTGCTAGGCCTTAGGAGGCCGTCTGTCGG
>Clone_7   GGCATGGAGAAGAGCAGACTTGAGGCCTGCTCCTTTACTCAATTTTGTGGCATCGCACACAACTTACTTTTCTGGCCAGATGGAGCAGTTCCTGCTAGGCCTTAGGAGGCCGTCTGTCGG
>Clone_8   GGCATGGAGAAGAGCAGACTTGAGGCCTGCTCCTTTACTCAATTTTGTGGCATCGCACACAACTTACTTTTCTGGCCAGATGGAGCAGTTCCTGCTAGGCCTTAGGAGGCCGTCTGTCGG
>Clone_9   GGCATGGAGAAGAGCAGACTTGAGGCCTGCTCCTTTACTCAATTTTGTGGCATCGCACACAACTTACTTTTCTGGCCAGATGGAGCAGTTCCTGCTAGGCCTTAGGAGGCCGTCTGTCGG
>Clone_10  GGCATGGAGAAGAGCAGACTTGAGGCCTGCTCCTTTACTCAATTTTGTGGCATCGCACACAACTTACTTTTCTGGCCAGATGGAGCAGTTCCTGCTAGGCCTTAGGAGGCCGTCTGTCGG
```

ii) DNase I/CIP/TAP treated (oAH182 reverse primer – anneals 166 nt downstream of 'ATG' codon of DNA Pol. Delta ORF)

```
           --------Candidate-23-----------|                                                    |---------DNA Pol. Delta Subunit---------
>Clone_1   GGCATGGAGAAGAGCAGACTTGAGGCCTGCTCCTTTACTCAATTTTGTGGCATCGCACACAACTTACTTTTCTGGCCAGATGGAGCAGTTCCTGCTAGGCCTTAGGAGGCCGTCTGTCGG
>Clone_2   GGCATGGAGAAGAGCAGACTTGAGGCCTGCTCCTTTACTCAATTTTGTGGCATCGCACACAACTTACTTTTCTGGCCAGATGGAGCAGTTCCTGCTAGGCCTTAGGAGGCCGTCTGTCGG
>Clone_3   GGCATGGAGAAGAGCAGACTTGAGGCCTGCTCCTTTACTCAATTTTGTGGCATCGCACACAACTTACTTTTCTGGCCAGATGGAGCAGTTCCTGCTAGGCCTTAGGAGGCCGTCTGTCGG
>Clone_4   GGCATGGAGAAGAGCAGACTTGAGGCCTGCTCCTTTACTCAATTTTGTGGCATCGCACACAACTTACTTTTCTGGCCAGATGGAGCAGTTCCTGCTAGGCCTTAGGAGGCCGTCTGTCGG
>Clone_5   GGCATGGAGAAGAGCAGACTTGAGGCCTGCTCCTTTACTCAATTTTGTGGCATCGCACACAACTTACTTTTCTGGCCAGATGGAGCAGTTCCTGCTAGGCCTTAGGAGGCCGTCTGTCGG
                                                                  TSS?
```

B) Candidate-5 ncRNA and Serine/Threonine Kinase (GLCHR03:299486-299605)

i) DNase I treated

```
           ----------Candidate-5------------|                          (oAH220 reverse primer)                  |--Ser/Thr Kinase--
>Clone_1   CCTTGCCCAGTCTGCCTCCATACTAATTTCTCCTTTACTCAATCAGGATGGCGAGCGGCAGTGTGAATGATGTGTCTTAGCAGAATTTTAACTAGGGGAAAATGGCCGTATCAGTCATCA
>Clone_2   CCTTGCCCAGTCTGCCTCCATACTAATTTCTCCTTTACTCAATCAGGATGGCGAGCGGCAGTGTGAATGATGTGTCTTAGCAGAATTTTAACTAGGGGAAAATGGCCGTATCAGTCATCA
>Clone_3   CCTTGCCCAGTCTGCCTCCATACTAATTTCTCCTTTACTCAATCAGGATGGCGAGCGGCAGTGTGAATGATGTGTCTTAGCAGAATTTTAACTAGGGGAAAATGGCCGTATCAGTCATCA
>Clone_4   CCTTGCCCAGTCTGCCTCCATACTAATTTCTCCTTTACTCAATCAGGATGGCGAGCGGCAGTGTGAATGATGTGTCTTAGCAGAATTTTAACTAGGGGAAAATGGCCGTATCAGTCATCA
>Clone_5   CCTTGCCCAGTCTGCCTCCATACTAATTTCTCCTTTACTCAATCAGGATGGCGAGCGGCAGTGTGAATGATGTGTCTTAGCAGAATTTTAACTAGGGGAAAATGGCCGTATCAGTCATCA
>Clone_6   CCTTGCCCAGTCTGCCTCCATACTAATTTCTCCTTTACTCAATCAGGATGGCGAGCGGCAGTGTGAATGATGTGTCTTAGCAGAATTTTAACTAGGGGAAAATGGCCGTATCAGTCATCA
>Clone_7   CCTTGCCCAGTCTGCCTCCATACTAATTTCTCCTTTACTCAATCAGGATGGCGAGCGGCAGTGTGAATGATGTGTCTTAGCAGAATTTTAACTAGGGGAAAATGGCCGTATCAGTCATCA
```

ii) DNase I/CIP/TAP treated (oAH183 reverse primer – anneals 278 nt downstream of 'ATG' codon of Ser/Thr Kinase ORF)

```
          ----------Candidate-5------------|                                                                  |--Ser/Thr Kinase--
>Clone_1  CCTTGCCCAGTCTGCCTCCATACTAATTTCTCCTTTACTCAATCAGGATGGCGAGCGGCAGTGTGAATGATGTGTCTTAGCAGAATTTTAACTAGGGGAAAATGGCCGTATCAGTCATCA
>Clone_2  CCTTGCCCAGTCTGCCTCCATACTAATTTCTCCTTTACTCAATCAGGATGGCGAGCGGCAGTGTGAATGATGTGTCTTAGCAGAATTTTAACTAGGGGAAAATGGCCGTATCAGTCATCA
>Clone_3  CCTTGCCCAGTCTGCCTCCATACTAATTTCTCCTTTACTCAATCAGGATGGCGAGCGGCAGTGTGAATGATGTGTCTTAGCAGAATTTTAACTAGGGGAAAATGGCCGTATCAGTCATCA
>Clone_4  CCTTGCCCAGTCTGCCTCCATACTAATTTCTCCTTTACTCAATCAGGATGGCGAGCGGCAGTGTGAATGATGTGTCTTAGCAGAATTTTAACTAGGGGAAAATGGCCGTATCAGTCATCA
                                                                                                  TSS?
```

C) Hsp90 5′ Exon and Replication Factor C Subunit 5 (GLCHR05:2515390-2515509)

i) DNase I treated

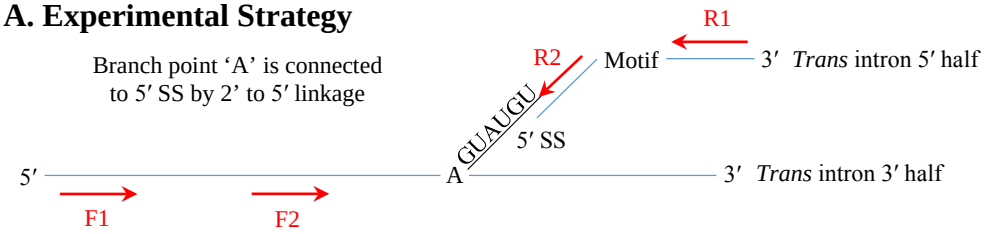
```
                                                (oAH221 reverse primer)                  |---------Replication Factor C subunit
5-------
>Clone_1   ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
>Clone_2   ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
>Clone_3   ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
>Clone_4   ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
>Clone_5   ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
>Clone_6   ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
>Clone_7   ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
>Clone_8   ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
>Clone_9   ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
>Clone_10  ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
>Clone_11  ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
>Clone_12  ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
```

ii) DNase I/CIP/TAP treated (oAH184 reverse primer – anneals 162 nt downstream of 'ATG' codon of Replication Factor C ORF)

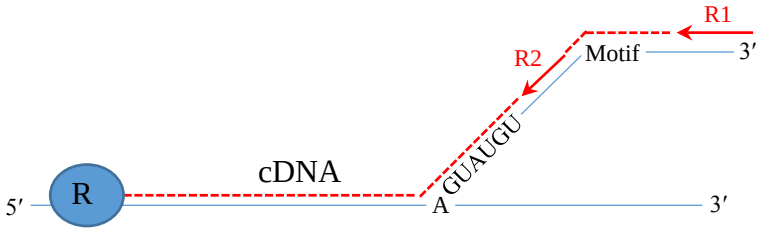
```
                                                                                                  |---------Replication Factor C
subunit 5-------
>Clone_1   ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
>Clone_2   ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
>Clone_3   ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
>Clone_4   ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
>Clone_5   ACTCCTTTACTCAAATTTGCGGCGTCGTGCTGCGGTCCTTCTCACGAAAATTCTTTGACCTCGCCTCGCCGGCATGTGGTGCAACAAGTACACCCCCACGCAGCTGGAGGCCATGGATTA
                                                  TSS?
```

Figure A.2.5. RT-PCR mediated detection of *trans*-spliced introns after the first step of splicing.
(A) Experimental strategy to detect intron splicing intermediates with and without RNA motif cleavage. **1. Reverse Transcriptase Step -** Reverse primers annealing upstream (R2) or downstream (R1) of the RNA motif sequence were used to generate first-strand cDNAs with reverse transcriptase (RT). **2. PCR Step #1 -** cDNAs were used as template in PCR reactions, using an upstream forward primer (F1) and reverse primer used for cDNA synthesis. **3. PCR Step #2 -** Products from PCR #1 were used as template in a second round of PCR using a downstream forward primer (F2) and reverse primer used for cDNA synthesis.
(B) Sequencing results from RT-PCR experiments. Individual RT-PCR sequence clones (traces for each clone shown above) were aligned and annotated using Geneious 4.8 software. Coloured bars under alignments indicate primer binding sites for PCR, regions of the PCR product corresponding to each *trans* intron half with splicing elements and RNA motif sequences indicated. A vertical red line emphasizes the boundary between the branch point sequence of the 3′ intron half and the 5′ splice site of the 5′ intron half.

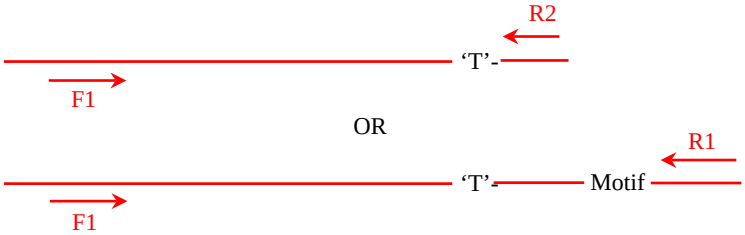
A. Experimental Strategy



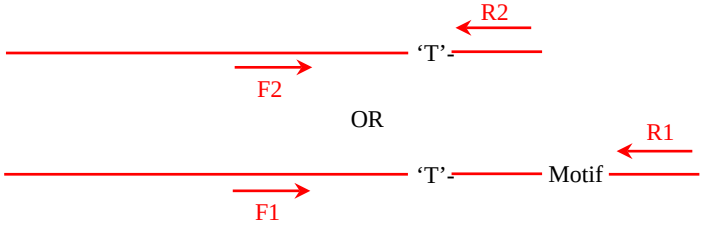
1. Reverse Transcriptase Step



2. PCR Step #1

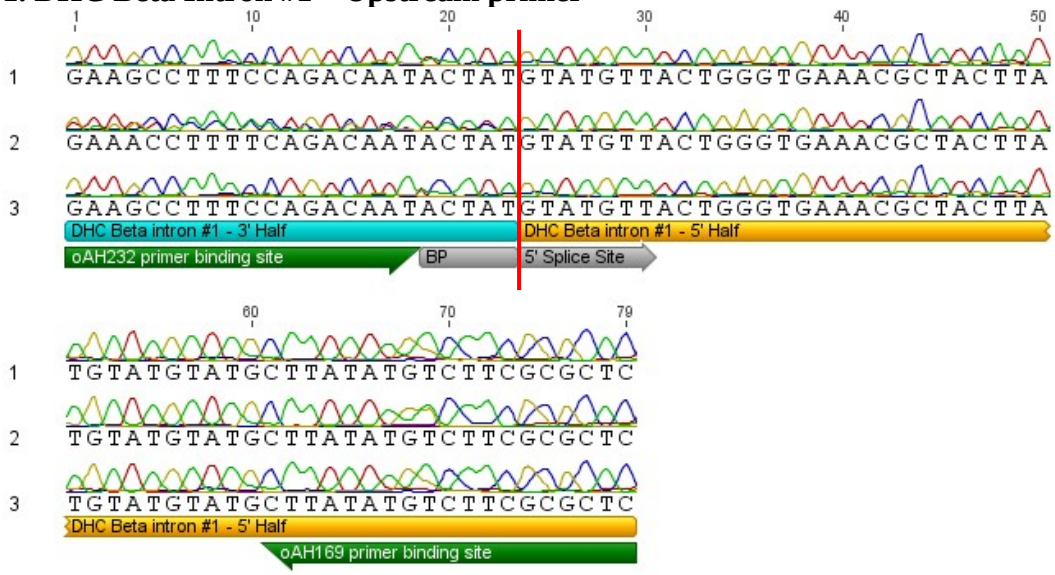


3. PCR Step #2 (nested PCR)



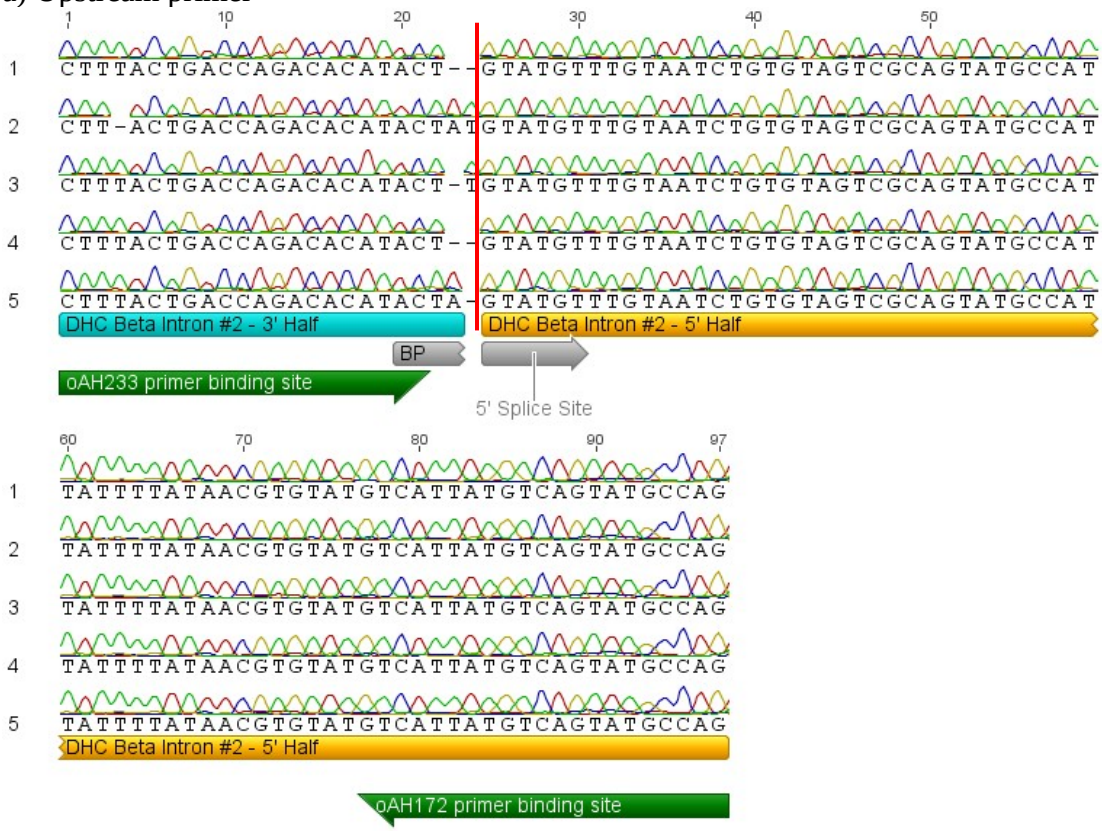
B. Sequencing results from RT-PCR experiments.

1. DHC Beta Intron #1 – Upstream primer

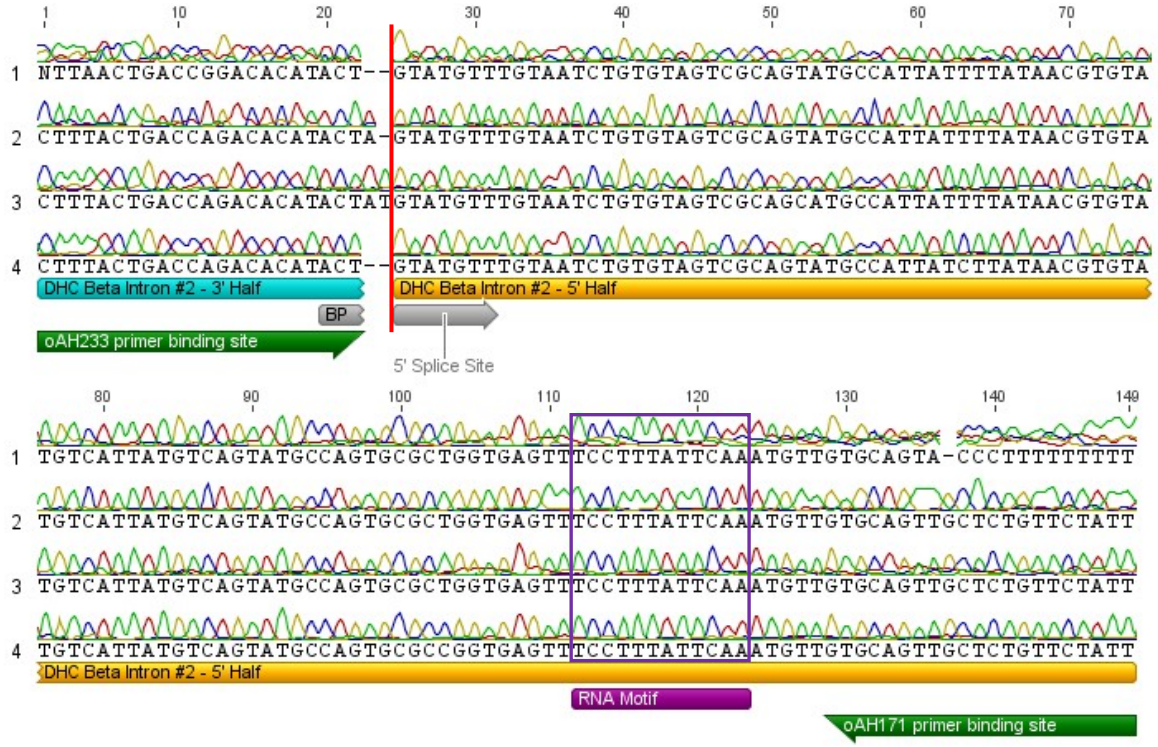


2. DHC Beta Intron #2 γ

a) Upstream primer

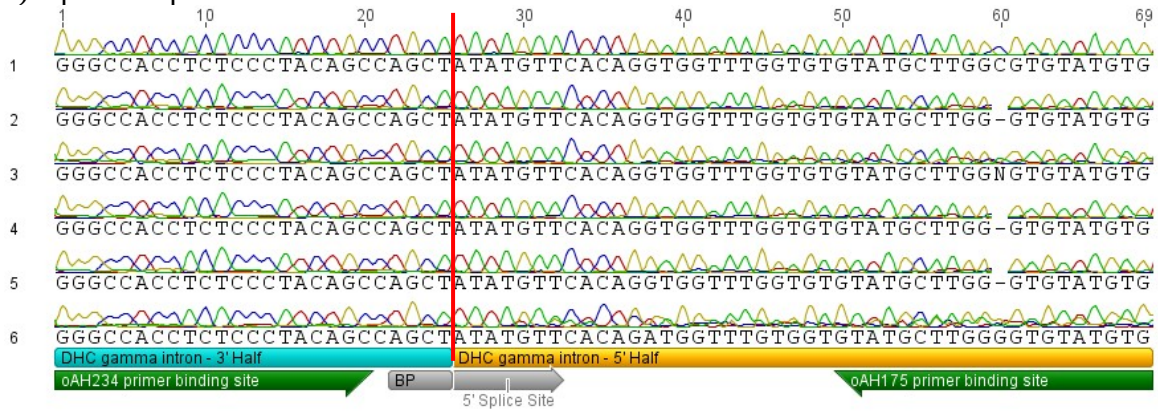


b) Downstream Primer



2. DHC Gamma Intron

a) Upstream primer



b) Downstream primer

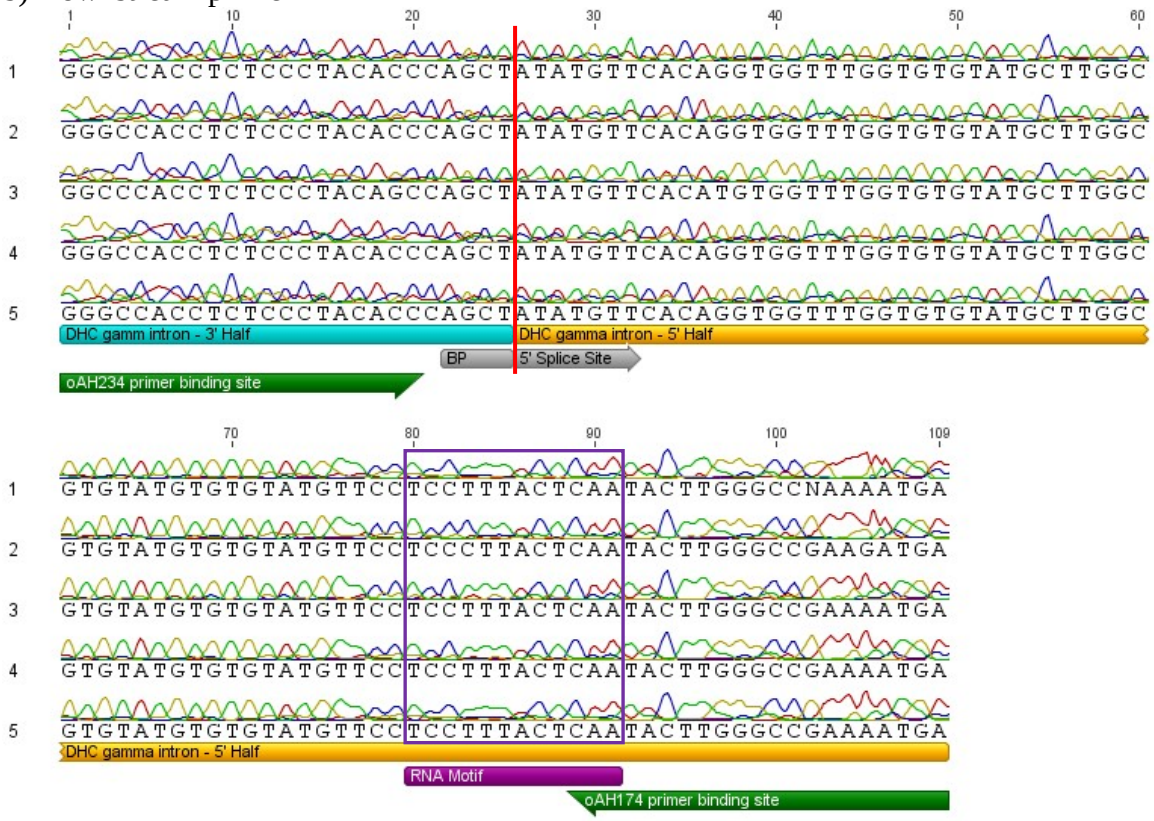
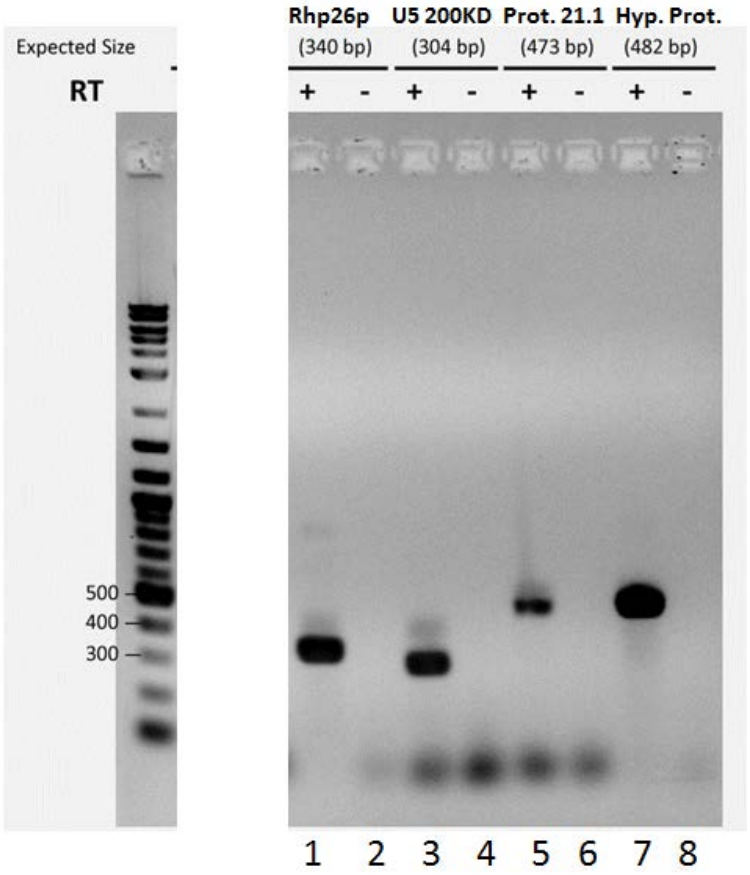


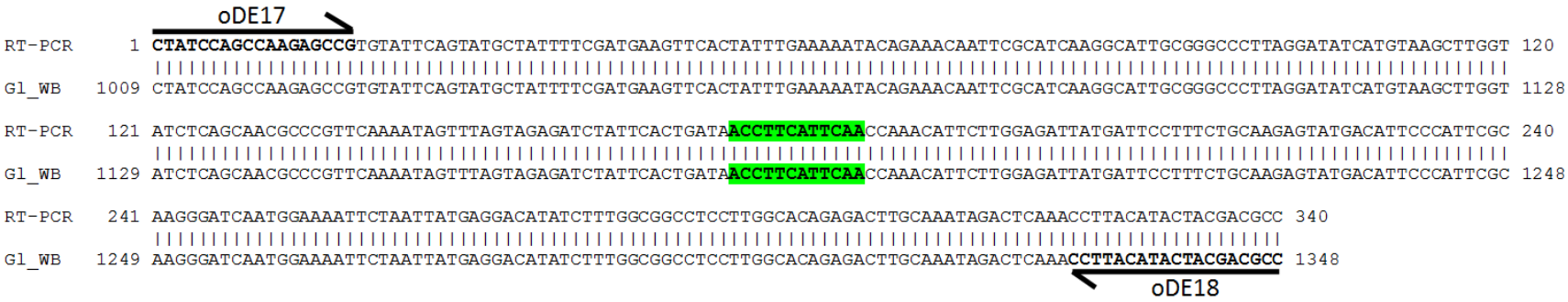
Figure A.2.6. Motif inclusion (no cleavage) in mature mRNAs.
(**A**) RT-PCR analysis of expression of *G. lamblia* protein coding sequences containing internal motif sequences. Experiments were carried out in either the presence (RT +) or absence (RT -) of SuperScript II™ reverse transcriptase and PCR products were resolved by 2% agarose gel electrophoresis. Expected sizes for products of Rhp26p mRNA (lanes 1 and 2), U5 snRNP 200 kDa helicase mRNA (3 and 4), protein 21.1 mRNA (5 and 6) and hypothetical protein (7 and 8) are indicated in parentheses and size marker (M) bands are in base-pairs. (**B**) Sequencing results for RT-PCR experiments (top sequence) are compared to *G. lamblia* WB isolate genomic sequences (bottom sequence). Open reading frame (ORF) identifiers are provided with the specific regions amplified indicated. Motif sequences within ORFs are highlighted in green and primer binding sites are shown in bold text with arrows indicating direction of amplification during PCR.

A



B

1. DNA repair and recombination protein Rhp26p (GL50803_87205:1009-1348)



2. U5 small nuclear ribonucleoprotein 200 kDa helicase, putative (GL50803_9352: 2570-2873)

3. Protein 21.1 (GL50803_25296:542-1014)

4. Hypothetical Protein (GL50803_7350:3160-3641)

Figure A.2.7. Primary and secondary structural features of previously predicted *Giardia lamblia* U1, U2, U4 and U6 snRNA candidates.
(A to D) *G. lamblia* (*G. intestinalis)* WB isolate U1, U2, U4 and U6 snRNA candidate primary nucleotide sequences identified by (Chen *et al*. 2008) are aligned with syntenic genomic regions of the P15 and GS isolates using ClustalW 2.0 (Larkin *et al.* 2007). SnRNA candidate nucleotides predicted to constitute functionally critical regions are highlighted in grey with nucleotide changes observed in P15 and GS isolates indicated with red text. Where appropriate, consensus sequences for evolutionarily-conserved snRNA elements are shown above the alignments to highlight unexpected nucleotide substitutions. "5′SS" = Intron 5′ splice site interacting sequence. **(E to G)** MFOLD based secondary structural predictions for *G. lamblia* WB isolate snRNA candidate sequences (Chen *et al.* 2008) are shown with nucleotide changes observed in the P15 (black arrows) and GS (grey arrows) isolates indicated.

A) U1 snRNA Candidate [identified in reference (Chen *et al.* 2008)]

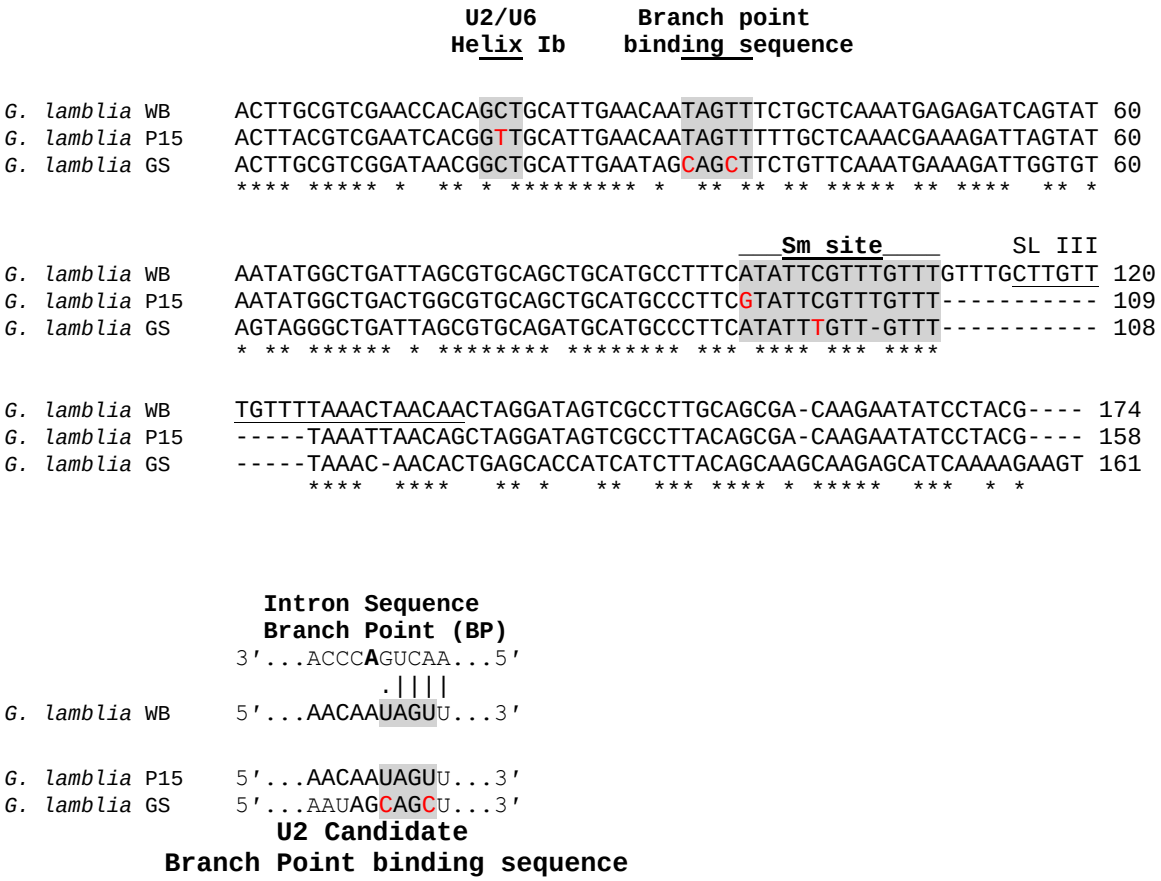
Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR02	846330	846451	+	122
G. lamblia GS	ACGJ01002714	394	515	+	122
G. lamblia P15	contig48	66075	66196	+	122

```
                                U1-70K Binding Site
                          5′SS_          (GATCACGAAGG)
G. lamblia WB     AAACATCAGCGGCATCGTCATCACGAAGATGAGCAAAAGCATAAAGTTCGAGATCCTCAT 60
G. lamblia P15    AAACATTAGCGGCATCGTCATCATGAAGATGAGCAAAAGCATGAAGTTCGAGATCCTCAT 60
G. lamblia GS     GAACATCAAGGGCAGTGTCATCATAAACACGAATAGAAGCATGAAATTTGAAATTCTCAG 60
                   ***** *  ****  *******  ** * **  * ****** ** ** ** ** ****

                                                                 Sm site
G. lamblia WB     CGTGTCTGCGAAGAGGAGGTTGACCAGGTTGCCGGCGGCAGAATTTTGGCGGGTGATGTC 120
G. lamblia P15    TGTGTCTGCGAAGAGGAGGTTGACCAGGTTACCGGCAGCAGAATTTTGGCGGGTGATGTC 120
G. lamblia GS     TGTGTCTGCAAAGAGGAGATTGACTAGGTTGCCGGCAGCAGAGTTCTGACGAGTGATATC 120
                   ******** ******** ***** ***** ***** ***** ** ** ** ***** **

G. lamblia WB     CG 122
G. lamblia P15    CG 122
G. lamblia GS     CG 122
                  **
```

Maps to antisense strand of gene encoding multidrug resistance-associated protein 1 (Gene ID: GL50803_115052)

B) U2 snRNA Candidate (Chen *et al.* 2008)

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR05	2371858	2372031	+	174
G. lamblia GS	ACGJ01000614	18412	18563	-	152
G. lamblia P15	contig59	52096	52253	-	158



Overlaps Rrmp3 helicase protein coding sequence on opposite strand (Gene ID: GL50803_16747)

C) U4 snRNA Candidate (Chen *et al.* 2008)

Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR05	1169325	1169457	-	133
G. lamblia GS	ACGJ01000362	2873	3005	-	133
G. lamblia P15	contig380	17649	17781	+	133

```
                    U4/U6
                 _helix II_                  ____5' stem-loop region____
G. lamblia WB    AATATTGCGAGAAAACCCTCTTAGAATTGATAGAAGACAGTCCTGGCGGGATTCCAATAG 60
G. lamblia P15   AATATTGTGAGAAAACTCTCTTAGAATTGATAGAAGACAGTCCTGGCGGAATTCCAATAG 60
G. lamblia GS    AATATTGTGAGAAAACTCTCTTAGAATTGATAGAAGACAGCCCTGGGGGAATTCCAATTG 60
                 ******* ******** ********************** ***** ** ******** *

                              U4/U6
                         ____helix I____                    _Sm site_
G. lamblia WB    AAACTGTTAAGCTTCTAACCTTTCAGATGCTTCGTGGTGTCGAATTTTTGTGGGAGTTCA 120
G. lamblia P15   AAACCGTTAAGCTTCTAACCTTTCAAATGCTTCGCGGTGTCGAATTTTTGTGGGAGTTCA 120
G. lamblia GS    AAACTGTTAAGCTTTTAACCTTTCAAATGCTTCGTGGTGTCGAATTTTTGTGGGAATTCA 120
                 **** ********* ********** ******** ******************** ****

G. lamblia WB    TGGAGATATGTCA 133
G. lamblia P15   TGGAGATATGTCA 133
G. lamblia GS    TGGAGATATGCCA 133
                 ********** **
```

Genomic organization: Maps to intergenic region

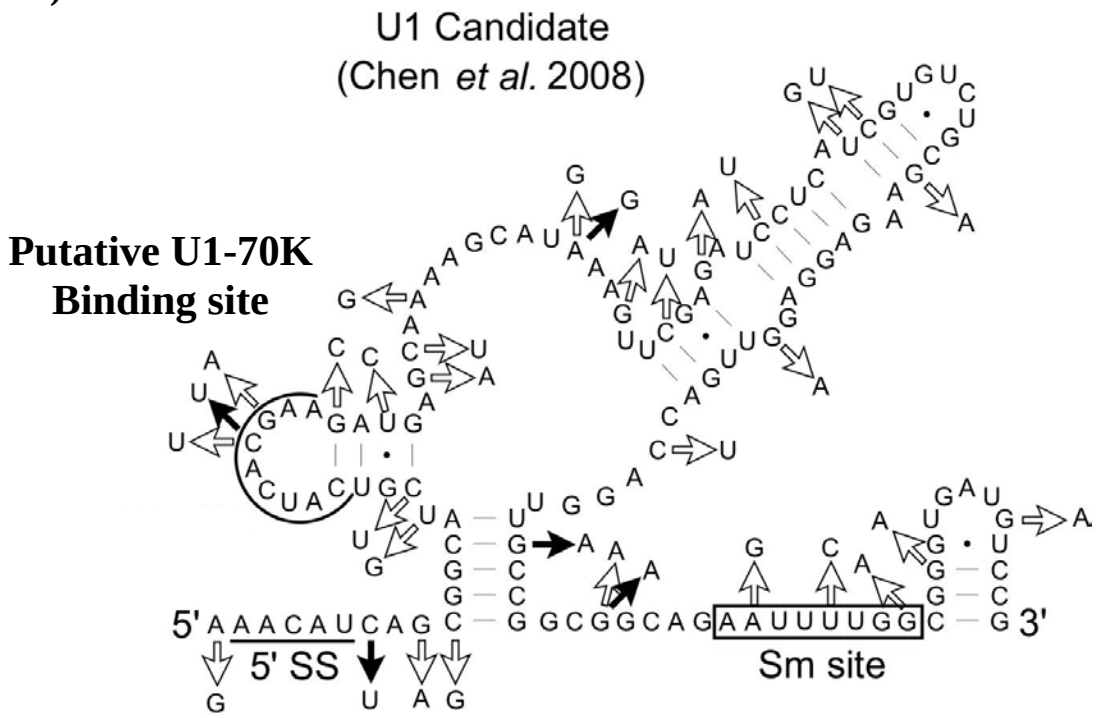
D) U6 snRNA Candidate (Chen *et al.* 2008)

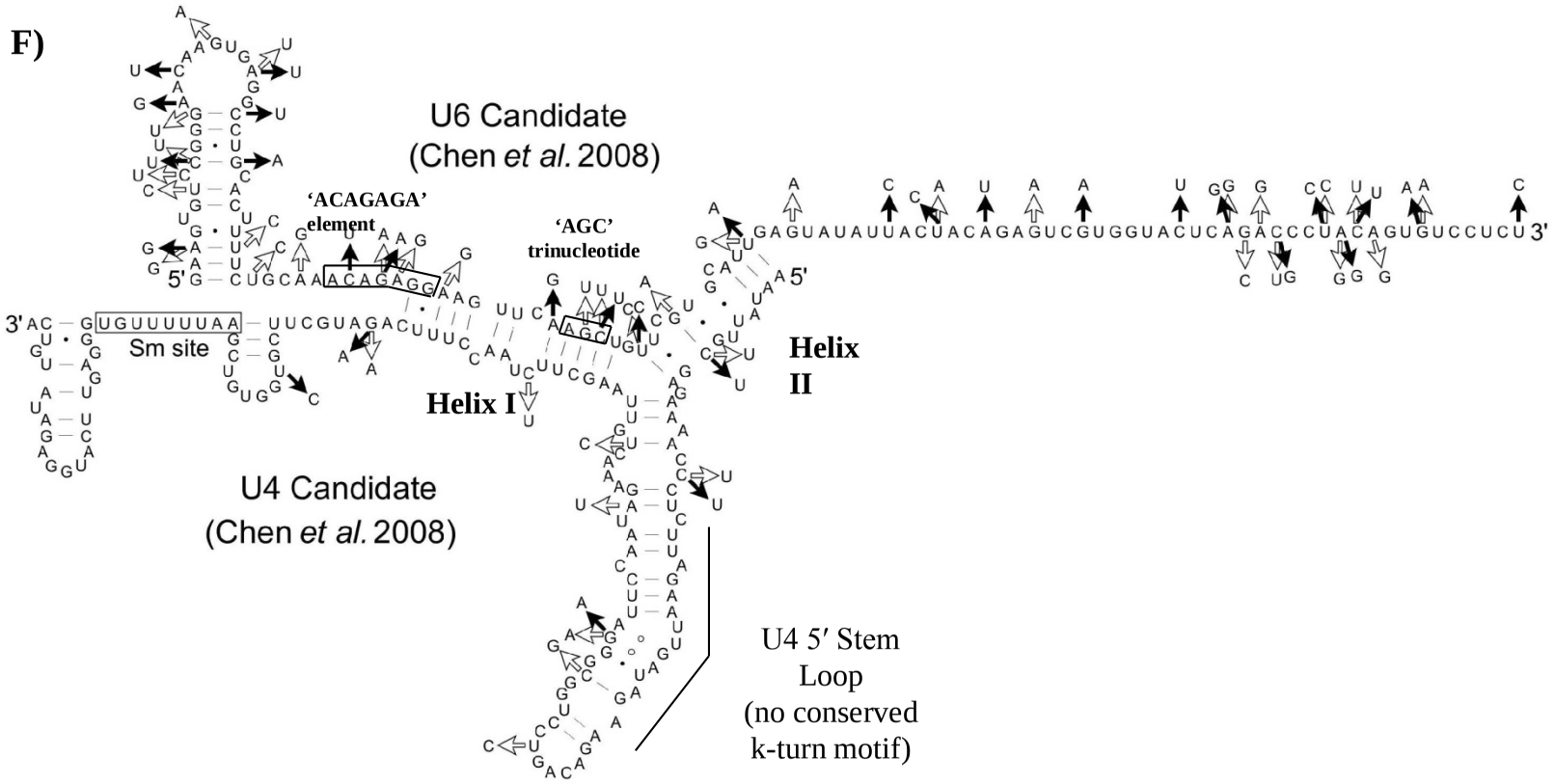
Genome	Sequence	Start	End	Strand	#Nucleotides
G. lamblia WB	GLCHR05	3863206	3863322	-	117
G. lamblia GS	ACGJ01002923	23027	23143	+	117
G. lamblia P15	contig393	54509	54625	-	117

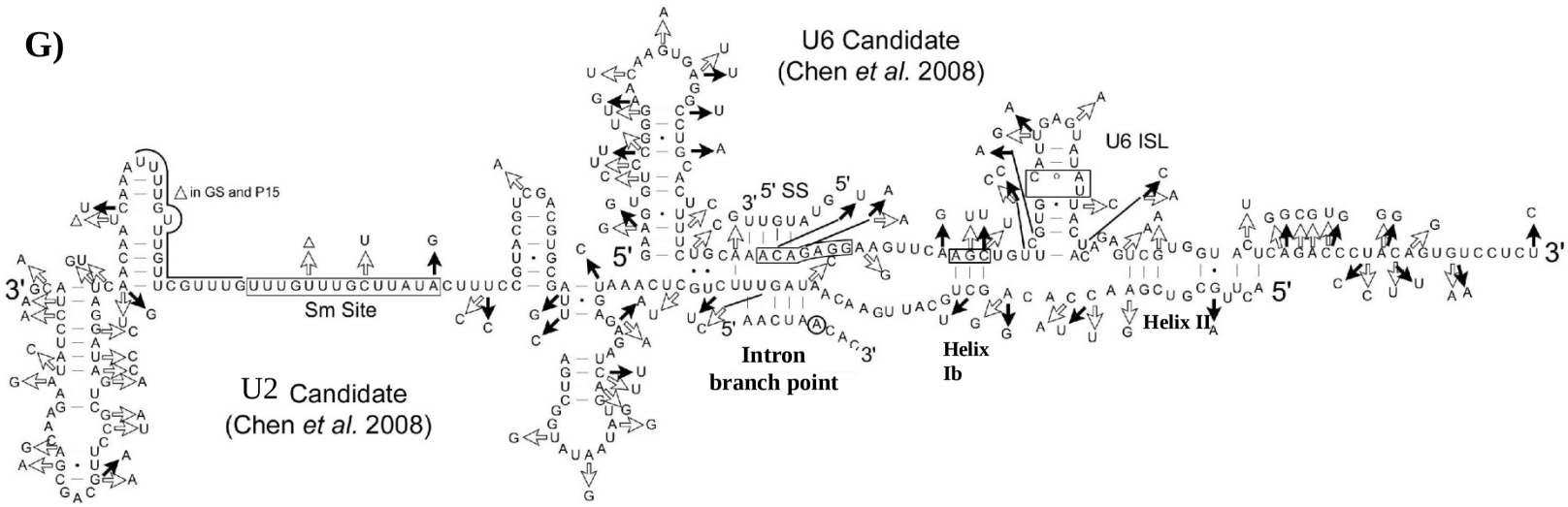
```
                                                        'ACAGAGA'     'AGC'
                                                         Element   trinucleotide
G. lamblia WB    GAAGTGTCCGGGAACAAGTGAGGCCTGCACTTTTCTGCAAACAGAGGAAGTTCAAGCTGT 60
G. lamblia P15   GAGGTGTCTGGGGATAAGTGTGGTCTACACTTTTCTGCAAATAAAGGAAGTTCGAGTTGC 60
G. lamblia GS    GAGGTGCTTGGTAACAAATGTGGCCTGCACTTCTCCGCGAACAAGGGGAGTTCAATTTGC 60
                 ** ***   **  * ** ** ** ** ***** ** ** ** *  ** ***** *  **
                 ________ISL__________
G. lamblia WB    TCGTGCATTGAGTATATTACTACAGAGTCGTGGTACTCAGACCCTACAGTGTCCTCT 117
G. lamblia P15   TCATGCATAGAGTATATCACCACTGAGTCATGGTATTCGGAGCCCGTAGTATCCTCC 117
G. lamblia GS    TCGTGCATGGAATATATTACAACAGAATCGTGGTACTCGCGTCCCGTGGTATCCTCT 117
                 ** ***** ** ***** ** ** ** ** ***** **    **    ** *****
```

Maps to antisense region of Hypothetical Protein coding sequence (GL50803_21048)

E)


U1 Candidate
(Chen *et al.* 2008)
Putative U1-70K
Binding site
5' AAACAUCAGC
5' SS
GCGGCAG
AAUUUUGGC 3'
Sm site


F)
U6 Candidate
(Chen et al. 2008)
'ACAGAGA' element
'AGC' trinucleotide
Sm site
Helix I
Helix II
U4 Candidate
(Chen et al. 2008)
U4 5′ Stem Loop (no conserved k-turn motif)


G)
U6 Candidate
(Chen et al. 2008)
U6 ISL
in GS and P15
Sm Site
U2 Candidate
(Chen et al. 2008)
5' SS
Intron
branch point
Helix
Ib
Helix II
3'
5'

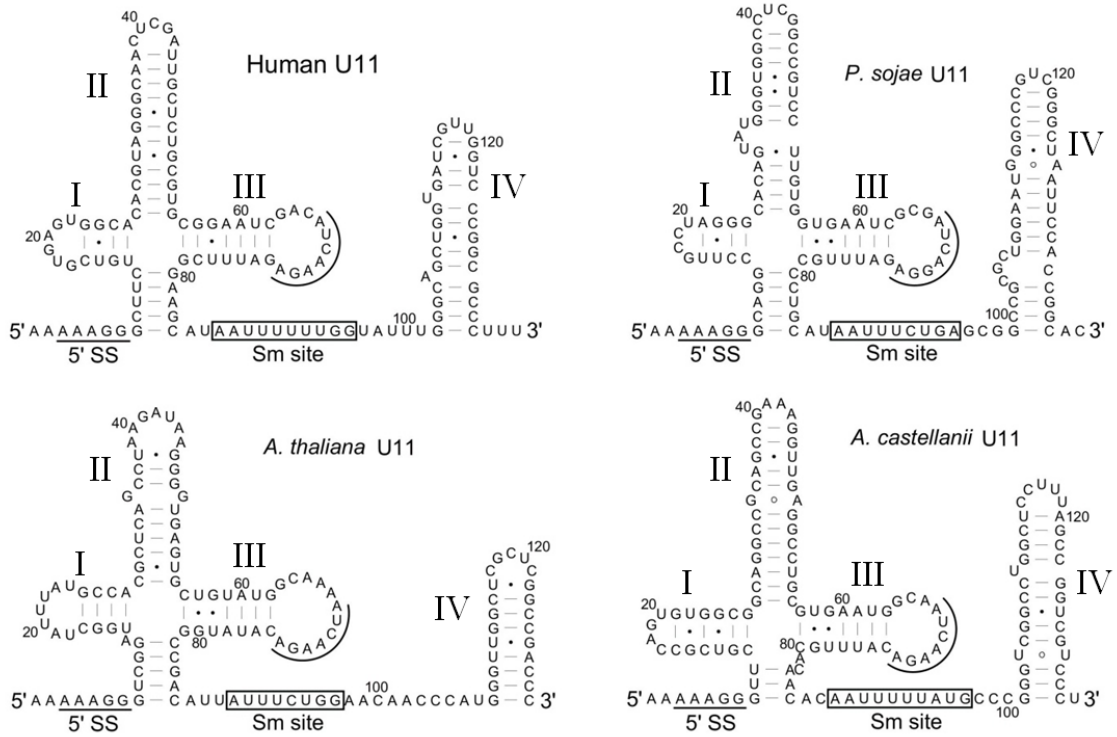


Figure A.2.8. A novel conserved sequence motif in U11 snRNA stem-loop III.
Secondary structural predictions of representative eukaryotic U11 snRNAs from humans (Patel and Steitz 2003), *Phythophthora sojae* (Russell *et al.* 2006), *Arabidopsis thaliana* and *Acanthamoeba castellanii* (Lopez *et al.* 2008) are shown with their 5′ splice site interacting sequences (5′ SS) underlined and Sm protein binding sites (Sm site) in boxes. We note a novel conserved sequence element of unknown function (underlined loop sequence with consensus 'AUCARGA') within stem-loop III (SL III) that is also present at the same relative position within SL III from the novel *Giardia* U1 snRNA.

Figure A.2.9. Primary sequence comparison of *G. lamblia* U2, U4 and U6 snRNA candidates with representative U2- and U12-dependent spliceosomal snRNAs.
Full-length *G. lamblia* WB isolate U2, U4 and U6 snRNA primary sequences (this study) were aligned with U2- and U12-dependent spliceosomal snRNAs from diverse eukaryotes using ClustalW 2.0 (Larkin *et al.* 2007). Regions of snRNAs predicted to form snRNA-snRNA, snRNA-intron or intramolecular base pairings for the *G. lamblia* snRNAs are indicated above alignments. Letters below alignments indicate *G. lamblia* snRNA nucleotides which are more major/U2-dependent ('M') or minor/U12-dependent ('m') spliceosomal-like (i.e. nucleotides which are exclusively conserved in the majority of representative eukaryotic major or minor snRNAs). (**A and B**) Grey boxed nucleotides are conserved in at least four snRNAs. (**C**) Nucleotides which are universally conserved between all U6 and U6atac snRNAs are in red text. "BP" = intron branch point interacting sequence, "SL" = stem-loop. Alignments were constructed using *Giardia lamblia* (Gl) U2 [JX416862], U4 [JX416863] and U6 [JX416864]; *Acanthamoeba castellanii* (Acan_cas) U2 [GenBank CW933695: nucleotide positions 787-579], U12 [CW917526:369-205], U6 [CW934080:725-629] and U6atac [AEYA01001292:233679-233780]; *Phytophthora* spp. (Phy_species) U2 [*Phythophthora ramorum* genome release V1.0 scaffold_1672:234-416], U12 [AAQY02000248:644532-644696], U4 [AATU01001408:348748-348878], U6 [AATU01006594:28185-28081] and U6atac [AATU01001737:24057-24182]; *Arabidopsis thaliana* (Arabi) U2 [X06478:200-359], U12 [CP002684:22603122-22603295], U4 [X67145:194-344], U4atac [CP002687:9096362-9096515], U6 [X52527:306-408] and U6atac [CP002688:16166306-16166185]; and *Homo sapiens* (Human) U2 [NR_002716:1-187], U12 [L43846:331-480], U4 [NR_003137:1-141], U4atac [AC073911.38:96778-96890], U6 [M14486:329-435] and U6atac [NR_023344:1-125] snRNA sequences.

A-1) U2 Alignment

```
                                        U2/U6 hI                U2/U6
                    U2/U6 hII_    _Ib___Ia_    ___BP___    hIII
Acan_cas_U2         -ACATCTT-CTCGGCCCAAGTGGCTAAGATCA-TGTGAAGTATCTGTTCTTATCAGCTTA 57
Phy_ramorum_U2      ---ACCTT-CTCGGCCTTT-TGGCTAAGATCA-AGTGTAGTATCTGTTCTAATCAGTGTG 54
Arabi_U2            -ATACCTTTCTCGGCCTTT-TGGCTAAGATCA-AGTGTAGTATCTGTTCTTATCAGTTTA 57
Human_U2            -ATCGCTT-CTCGGCCTTT-TGGCTAAGATCA-AGTGTAGTATCTGTTCTTATCAGTTTA 56
Gl_U2               --------TAAAATCAGAGTCGGCTTCGACTTTAGTGTAGTTACTGTT-TCGTCGGCTTA 51
                              M      MMMM  MM    MMMMMMMM  MMMMM M  MM M MMM
                                                                ___Sm site___
Acan_cas_U2         ATCTCTGGTAGTGAGGCCTCCTGTGCCTCACCTCAAGGTTAGACTTATTTTTCTTGTTGG 117
Phy_ramorum_U2      AAAACTGGT---TCCGACGTTTTTCGTT--GGTCTTTTTCACATTCATTTTTGG------ 104
Arabi_U2            ATATCTGATATGTGGGCCATCGGC-CCACACGAT---ATTAACTCTATTTTTTAAGGGAG 113
Human_U2            ATATCTGATA--CGTCCTCTATCCGAGGAC--AATATATTAAATGGATTTTTGGA----- 107
Gl_U2               ACCGCCGAT------------CCACTACATGCAAGGGGCAGCCGG------GCTGTGAG 92
                    M   M   M                                  M

Acan_cas_U2         GC-TCCTGGCACCATGCCCTTCCAGCTATGCTGTGGGCAGTCCAGAGAGCAGTGATC--- 173
Phy_ramorum_U2      GCATCCCGATGTCGCGC------AGCT-TGCTGTGCGAGGTC---GGGGCGGTTTCCGGG 154
Arabi_U2            AAAGCCCGTTAAGAT--------AGCT-TGCTATCTG---------------------- 141
Human_U2            GCAGGGAGATGGAATAGG-----AGCT-TGCTCCGTCCACTC-------CACGCATC--- 151
Gl_U2               GCAGC-TGCCAGGATG--------GTCCTGCCCTTGTC-----------------CC--- 123
                    MMM M  M     MM          M   MMM  M                      M

Acan_cas_U2         AGCTTTGTACTGCACCACCCTGCAAAGTTCTTCAAAT- 210
Phy_ramorum_U2      GGCTTT-CACCTCTCC--CCCGCGAGGC----CAAC-- 183
Arabi_U2            GGCTTT--------------CGCGAGTCGCC-CA---- 160
Human_U2            GACCTGGTATTGCAGTACCTCCAGGAACGGTGCACC-- 187
Gl_U2               GGCT------------------GGCGCCGTCCACCTT 142
                     MMMM                  M   M    MM
```

A-2) U12 Alignment

```
                          U2/U6 hI          U2/U6
                         _Ib___Ia__      ___BP___ _hIII_
Human_U12      -----------ATGCCTTAAACTTA-TGAGTAAGGAAAATAACGATTCGGGGTGA------------ 43
Phy_sojae_U12  -------------GGCTTAAACTCAATGAGTAAGGAAACTAACGC-CCTACCTGACAAGTAGGGCTGC 54
Acan_cas_U12   -------------GCCTTAAACT-AATGAGTAAGGAAAATACCGC-ACCGGTTGAC-ACCGGTGTTAT 52
Arabi_U12      -------------GCCTTAAACT-AATGAGTAAGGAAAACAAAGC-GTCCGGTGAGAACCG--GTCGC 51
Gl_U2          TAAAATCAGAGTCGGCTTCGACTTTA---GTGTAGTTACTGTTTCGT-------------------- 44
                           m mmm  mmm  m   mm   m  m m    m
                                                 _Sm_site
Human_U12          CGCCCGAATCCTC-ACTGCTA--ATGTGAGACGAATTTTTGA-GCGGGTAAA--GGTC-- 95
Phy_sojae_U12      CTGTTGGATCG---AGTGATC-CATGGGTTTCTAATATTTGACGGAGGTGGAATGGCC-- 108
Acan_cas_U12       GGCGTGAACCG-----CGTTCACGTCGGCTTCCAATTTCTGGCG-GGCTACA--GGCCTC 104
Arabi_U12          GGCGCTAATCGTAAAACACAAAATTGGCGCAATAATTTATGGAGGGGTTATA--GGCT-- 107
Gl_U2              CGGCTTAACCGCCGATCCACTACATG------CAAGGGGCAGCCGGGCTGTGA-GGCA-- 95
                    m    mm mm   m          mm       mm          mm m     mmm

Human_U12          -GCCCTCAA--GGTGACC--CGCCTACTTTG---CGGGATGCCT-----GGGAGTTGCGA 142
Phy_sojae_U12      -GTCATTTT--AGGAACC--CG-CTACTTT----TAGGGTGACTAAAAAGGGGGCCGGTC 158
Acan_cas_U12       GGCCCTCTTGTAGTGACCTGCGCCTACTTTCGCGCGGGATGCTCGA---GTTGGCTGGCC 161
Arabi_U12          GGCCGATGT--GTTGACGC-TGCTTACTTTT--GCAGAACTCACCTCGTGCGGGCCTCCC 162
Gl_U2              -GCTGCCAG--GATGGTC-----CTGCCCT---------TGTCCCG---GCTGGC-GCCG 134
                    mm          mm  m     mm m  m         mm        m  mmm m

Human_U12          TCTGCCCG---- 150
Phy_sojae_U12      CCCGCCC----- 165
Acan_cas_U12       TCCG-------- 165
Arabi_U12          TACACCCATCCC 174
Gl_U2              TCCACCTT---- 142
                   m m mm
```

B) U4/U4atac Alignment

```
                    U4/U6 or U4a/U6a hII_      __U4_5' Stem-loop_____
Human_U4atac        --ACCATCCTTCTCGTGGG-GTTGTGTTACTGTCCAGTGAGCGCATGGT-GAGGGCA--A 54
Arabi_U4atac        AACCCGTTTCTGTCAGAGGTGAAGGATGATCCGTCAATGATCGTTTAGA-GACGGCGG-A 58
Human_U4            -----AGCTTTGCGCAGTG-GCAGTATCGTAGCC-AATGAGGTT-TATCCGAGGCGCGAT 52
Arabi_U4            -----ATCTTTGCGCTTGGGGCAATGACGCAGCT-AATGAGGTTCTAACCGAGGCGCGTC 54
Phy_inf_U4          -----ATCTTTGTGCTTGGGGCAATACGATAGTGT-GTGAAGCTCTGCT-GATGCATCGT 53
Gl_U4               -----GACTCTAGGCT-GAAGC--TGCCAAGGTGC-GTGATCCCTCGGT-GATGCCTTGA 50
                              MM      M         M                       M
                                                   m      m     m       m   m
                    U4/U6 or
                    U4a/U6a hI
Human_U4atac        -TACTGCTAACGC----CTACA-C-AACACA---CCCACATCAA--------------CT 90
Arabi_U4atac        -TCGTGCCGACACAGAATTTGA-CGAACATAATTTTCAAGGCGAGTGGGCCTTGCCTTACT 117
Human_U4            -TATTGCTAATTGAAAACTTTTCCCAATAC----CCCGCCATGA--CGACTTGAAA----T 102
Arabi_U4            -TATTGCTGGTTGAAAACTATTTCCAA-AC----CCCCTCCT----AGGCCTAAG------ 99
Phy_inf_U4          -GATTGCTAGTTGAAAACTACTCC-AACAC-----CCGTGAGAA---GGCCAC-------- 96
Gl_U4               GTGTTGCTTCACCAAAGAAC----AACCACA----------------------------- 77
                       M         M
                              m
                                        _Sm_site_
Human_U4atac         ATGGTG--GTGC---------AATTTTTTGAAAA-------------------- 113
Arabi_U4atac         TTGGTT--GGGCCTGCCCGTCAATTTTTGGAAGC------CTCGA--------- 154
Human_U4             ATAGTC---GGCAT--TGG-CAATTTTTGACAGT---CTCTACGGAGA------ 141
Arabi_U4             CTTGTCTTAGGCCT--TCGAGAATTTCTGGAAGGGCTCCCTTTGGGGTAAAGCC 151
Phy_inf_U4           -TGGC---CAGCTC------CAATTTCTGTTTTATCTCCCACTAT--------- 131
Gl_U4                -CGGCA--CAGC-------CGAATCTCTCATT---------------------- 99
                                          M
```

C) U6/U6atac Alignment

```
                                                  U6/U2 or U6a/U2a   U6/U2 or U6a/U2a
                                                        hIII              __ hI___
Human_U6            GTGCTCGCTTCGGCAGCA-CATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGG 59
Arabi_U6            ---GTCCCTTCGG--GGA-CATCCGATAAAATTGGAACGATACAGAGAAGATTAGCATGG 54
Acan_cas_U6         -----------GGAGGCTCCATCTGTTAAAATTGGAACGATACAGAGAAGATTAGCATGG 49
Phy_inf_U6          --GACCACTTCGGTGGT--CATCCGTAAAAATTGGAACGATACAGAGAAGATTAGCATGG 56
Phy_inf_U6atac      ----------------------------GTGTTCGTTGAGCCGAGAGAAGGTTAGCATC- 31
Acan_cas_U6atac     ----------------------------GTGCTGGTTGAGCCGAGAGAAGGTTAGCATC- 31
Human_U6atac        ----------------------------GTGTTGTATGAAAGGAGAGAAGGTTAGCACT- 31
Arabi_U6atac        ----------------------------GTGTTCGTAGAAAGGAGAGATGGTTGGCATC- 31
Gl_U6               ----------------------------GTGGTTAACAAAAACAGAGACAGTTAGCACCA 32
                                                                  MM
                                                  mmm       m           m       m

                    U6/U4 or U6a/U4a
                    ____hII_____                                 _____ U6atac 3′ SL _____
Human_U6            CCCCTGCGCAAGGATGACA---------CGCA------AA--------TTCGTG--AAGC 94
Arabi_U6            CCCCTGCGCAAGGATGACA---------CGCAT-----AA--------ATCGAG--AAAT 90
Acan_cas_U6         CCCCTGCGCAAGGATGACA---------CGCA------AA--------ATCGAG--AAGA 84
Phy_inf_U6          CCCCTGCGCAAGGATGACA---------CGCAT-----AA--------ATCGAG--AAG- 91
Phy_inf_U6atac      TCCCTCGACAAGGACGGGATTGCGCGTTTGCGTATC-CAAC-CACTGGATGGT-TTAAGC 88
Acan_cas_U6atac     TCCCTGCATAAGGACGGGAAAAGAC-TCCG-GTCTT-CAACTCAC---ATCGTGTTAAGG 85
Human_U6atac        CCCCTTGACAAGGATGGAAGAG-GCCCTCGGGCCTGACAACACGC---ATACGGTTAAGG 87
Arabi_U6atac        TCCTCTGACAGAGACGGGATTTGACCTTCGGGTCTTTGAAC--AC---ATCCGGTTAAGG 86
Gl_U6               GCTTCAGTCTAGAGTCGCTGGGGGACCTCTGGTTTCGCGGG-------AGCCCGTTGGCG 85
                                  M
                          m           m           m    m   m  m                  mm    m

                    __U6atac 3′ SL__                 _Lsm Site_
Human_U6            GTT-----------------------CCATATTTTT-- 107
Arabi_U6            GGT-----------------------CCAAATTTTT-- 103
Acan_cas_U6         TAC-----------------------CCAACTTTTT-- 97
Phy_inf_U6          TAT-----------------------CGCACTTTTG- 105
Phy_inf_U6atac      T-CTGTCATCCTTCTGGAAGACATCTACCAGTTTTTTTT 126
Acan_cas_U6atac     C-TAGTAAC-----------------ACTAACTTTT--- 103
Human_U6atac        CATTGCCACCTACTTCGTGGCATCTAACCATCGTTTTT- 125
Arabi_U6atac        C-TCTCCACATTCGTGTGGATCTAAACCCAATTTTTT-- 122
Gl_U6               CGTGCTTGCACCCCGCTCCT------------------- 105

                    m       m
```

Appendix 3 - Supplementary Material for Chapter 4:

Conservation of Spliceosomal Intron Structures and snRNA Divergence in Diplomonad and Parabasalid Lineages

Table A.3.1. Spliceosomal introns in conserved protein coding genes from *Spironucleus vortens*.
Genomic sequences encoding predicted and EST-confirmed spliceosomal intron-containing genes from *S. vortens* are shown with intron sequences in lower case red text and start/stop codons are bolded. GenBank expressed sequence tags (ESTs) confirming intron splicing and the phase of intron insertions are indicated.

Host Gene	Gene Coding Sequence 5′-3′ (introns in red text)	Genomic trace accession	Intron Phase	Confirming ESTs
Rpl7a	**ATG**CACCGCTGCCAAGCCGAACACGTCAACAACC TGACCCGCATGGTCAAGTGGCCCGCCTACATCCG CATCCAGCGCCAGAAGGCCCTCCTCCAGCACCGC CTGAAGGTCCCCGGCGTCGTCAACATGTTCCGCA ACCCGCTGAACGCCAACGCCACCAAGGAGATCCT GAAGTTCGCCGCCAAGTACCAGCCGGAGACCAAG GAGGCCAGACAGCAGCGCCTTGTCCAGGCTGCCG ACAAGAAGACCACCATCAACGCCCCAGTGTCCTT CAACTACAACATCCACAAGGTTGTTGAGGCCGTC GAGAAGAAGGAGGCCAAGCTGGTCCTCATCGCCC ACGACGTCGACCCAATCGAG**gtaagtctaaacta** **actgtgatccgcggatcgcttgagttacgaaact** **ttgctaacaaactag**CTCGTCCTGTACCTGCCAA CCCTCTGCCACAAGAACAACATCCCATATGCCAT CGTTCGCTCCCGCACCGAGCTCGGCAAGCTGGTT CACTGCACCAAGTGCACCTCCATCGCCTTCACCA CCATCAAGCCGGAGGACACCGCCGCCTTCAAGTC CATCCTGGACACTGTTGCCCACGAGGTCGACTAC GTCCACGCCATCAAGACCCACGGTGGTGTTTCCC GCTCCAACAAGTCCCTTGCTAAGGAGGCCAAGAA GAACAAGATCGGCAAGAAA**TGA**	Ti:2141515448: 16-733	0	GH187119.1, GH187120.1, GH184167.1, GH184166.1, GH195615.1
Rpl30	**ATG**GATCGCGTATCT**gtgagttgaaacaaactga** **gaccagaaacctggatccagtacaaacgaaactt** **tactaacaaactag**AAGAAGTCTTCTGAATCGGC CGCCTTGCAGCTTGCTCTTGTCGTCAAGTCCGGC AAGTACACCCTTGGTGTCAACCAGGCTCTTAAGT CCATCCGCAACCTGAAGGCCAAGCTCGTCATCAT CACCTCCAACCTTCCACCCCTGGTCGCCTCCCAG ATCGAGTACCTCTGCATGCTCAGCGGCATCCCAG TCCACGCCTTCCCGTCCAACTCCCGCGAGTTCGG CGTTACCCTCGGTAAGCAGTTCAACGTCGGCGTT ATGGCCGTGACTGAAGCCGGCGACGCTGACCTCG CTGCCTTCAAG**TGA**	Ti:2141591039: 516-129	0	GH194962.1, GH190822.1, GH188072.1, GH188071.1
Rps4	**ATG**gtaagtctaaaatgtgtgcgcacggcgcatc atctatttgcttgaactaacaaactagGCTCGTG GTCCAAAACTTCATATGAAACGTCTTAACGCTCC ATCCCACTGGCAGCAGGACAAGCTTGGCGGCATC TACTCCACCAAGTGCAACCTCTCCACCCACAGGA TCAATGAGTGCGTCCCAATGTCCCTCGTTCTCCG CAACCGCCTGAACCTCGCCAAGACCCTCCGCGAG TGCAAGCTCATCCTCGACACCAAGAACATCCTGG TTGACGGCAAGGTCCGCACCGACTCCCAGTTCTC CGTCGGCTTCATGGACGTCCTCGAGGTCAAGAAG CTCAACAAGCTGTACCGCATCCTCTTCGACATCA AGGGCCGCCTGACCCTCCAGCCGATCGACAAGAA GGAGGCCGAGTTCAAGCTCCTCCGCATCAACAAG GTCTTCCTCGGTGAGAAGGGCGTCCGCTACGGTG TCTCTCACGACGGCCGCACCATCCGCTTCCTCCC GGACGACGTCAAGGTCAACGACACCGTCAAGTTC GACCTGAAGACCGGCAACATCGTTGAGAAGGCCC AGTTCAACATCGGCCAGATGGCCTGTGTCACCAT CGGTCAGAACGTCGGCTCCATCGGCAAGATCACC CAGGTCGACCCGCACAACGGCTCCTACACCATGG TCCACCTCGTCGACGCCGCCGGCCAGAAGTTCAT CACCCGCCGCGAGAACGTCTTCATCCTCGGCAAC CAGGGTAACTCCTTCATCTCTATCCCCAAGGAGA AGGGCGTCAAGCCGACCATCTTCCAGGAGCGTGA	Ti:2141550682: 123-914 and Ti:2141654036: 986-256	0	GH185743.1

	CCTCCGCCTGGCCTCCATCGCCAAGCACCAGAGA AACGAG**TGA**			
Rps12	**ATG**TCAACgtaagtctagagctgagcagtcaact ttactaacaaaatagTGACCAACTGAAGACTTTC TGTAAGAAGATCCGCGTCCACGGTGCTATGGTCT CCGGCGTCCGCCAGGTCGTGCGCGCCGTCGAGAA CCACGCCACCTCCAACGTGAAAGTCATTCTCCTG GCCAACGACTGCAAGGAAGCCGGCATCAAGAACC TCGTCAAGGCCCTCGCCAAGCAGCACTCCATCGG CGTCTGCGAGAAGTTCGGCGCCGCCCACCTCGGC GAGCTCGCCCACCAGTACGTGATCAAGGGCCACG TCACCGAGGGCAAGATCGGCAAGGTCAGAAACGC CTCCTGCATGGCCATCCAGAACTTCGGCACCCTC ACCGCTGAGGATCAGGCCGCTTTCAACGCTCTCC TCCAG**TGA**	Ti:2141614707: 43-458	2	GH185220.1
Rps24	**ATG**CAGATCAAGTATCGCGAAATTGTCAACAACC CGATCCTCGATCGTACTCAAATGgtaagtctaaa tctcatgtataactaatactaacaagttagAAGC TCAAGATCGTCCACCCAGGTAAGTCCGTGGGTAC CATCGAGGCTCTCCGCGAGCTCGTCCAGAAGGAT CGTAAGATCAAGGACATCAAGCAGGTTGTCGTCT TTGACTGCCACACCAAGCACGGTGGTAACCTCAG CACTGCTTCTTGCCACATCTACGGCAACGTTGAG ACCCTGAAGAAGGTTGAGCCGAAGTACACCATCA TCCGCCTTGGTTACATCGAGAAGCCGAAGCCAGT CTCCCGCAAGATGATCAAGAACCACAAGAACAAG CTCATCCGCAAGTTCGGTACTGCCAAGAGCAAGA TCGTCATGTCTGGTAAGAAGAAC**TGA**	Ti:2141541737: 116-549	0	GH184038.1, GH193644.1, GH184039.1, GH192698.1, GH192697.1
FolC-like	**ATG**TAGTACCCTCAAGTGCTTgtaagtcaacttt tgccatcaaacttttgctaacaaattagGATTCC CTCTCAAAAGTATCATAAACTGTCAAGTCGCATT GGTCATAGCTACCAGATCTACTAAATAAGTAATT TCGACCGGAATTGCTGTTTCACGTCACTGGTTCA AAAGGTAAAACGTCTATTTCGCACTATTTGAGCA AAAATATAACAAATAGCGGACTGTTTACATCTCC CCACCTTATTAGCTTCCGCGATAGAATCAAGGTG AACAACACGCCTCTTACACCACTAGAGTTTACAA CCCTATATAATTATCACTTATCCAATATATAGAA TTTACCACCGTTTCAAAAATCGTTCGTGTTGGCT CAAAACCACTTCTCAAATCTGTCGCTCCCCGTTC AGATTTTTGAGGTCGGTATAGGCGGCTTGCACGA CTCCACATAAGCA**TGA**	Ti:2141479887: 480-23	0	GH189627.1
Rps15 (Predicted 5′ UTR intron)	TTCATTTTGATATTAAGTCTTTTTAGTTATGTTT TCCACACCTTTTCTTTTGGGTAACTAATTgtaag tctaaatctattgcgaaactttgctaacaaggtc ctggaacgggccctag**AATG**GGTCGTACTAATGT TCTCAATGACGTTCTCAAGCAGATCACCAACGCT CAGCGCCTTGGCAGACGCCAGTGCATCCTGCACC CAGTCAACTCCGTCACCCTGAAGGTCCTGGAGAT CATGCAGAAGGAGGGCTACATCGGCGACTTCACC TTCGTTGATGACCGCCGCGGCAACAAGGTCGTCG CGAACCTGACCGGCCGCCTCAACAAGGCTGCCGT CATCTCCCCGCGTTTCGACGTCTCCCACAACGAC CTCTCCAAGTGGGTTGTCAACCTCCTCCCGTCCC GCCTCTTCGGCCACATCCTGCTCTCCACCACCGT CGGCATCATCGACCACAACGAGGCCCAGCACCGC AACATTGGAGGCAAGATAATCGGTTTCTTCTAC**T** **GA**	Ti:2141512834: 214-725	N/A	None
Hypothetical ORF 1 (Predicted – 3′ end unknown)	**ATG**TCCGAAACCTCGTCCTCCAGCGACGCTGGAG ACGCTTTCGAGTAATATCgtaggtctaattgata tggataactttactaacaaactagTGCAATAGCG CCAGAAAAAGAACCAAGAAATCGAAGAATTAAAG AATCAAACGCAGTAACTGCTCTTGTAAACCCAGG AAAAGGAATAAATCGACCCGTACGCAGATATTAT GAACTCCCTAAAAGACACGAAAAATCTCGTCGGC AAAGACGTCGACTAATGGGAGCAAACGATTAAAA CTGACTTCGCTGAGCCAGATGTTCAAGATCAGAA TGACATCTTATAAGAGACACAACAATCACTCGCT CAACAATCACAACTCATTGATGTCGCGTCGCCAG TTCATGAATAGCAGTCTCTGGCTGCCACCATGTC GCCGCAGTAGCCGCAACCTGTCGACCTGACAACA	Ti:2141664662: 761-1	1	None

	TAAAATGACACATTGCGAATGGAAAAAGAGGACT TCTATTTTAAAAAATCGCTGAGCCAACAGTAGAA ACTGCAGGCTGTATTGAACTAGCAAAACAAATAG TAAAAGCTGATAAAGTAGAAAAACAAAGACGAAT TGGAGTAAATAAGCTAGCTCTTCAAAACTATGGT AAATAAACCAAATCCATAGGCGGCCAACTAGCAA GTAAAGCATGCCCTCGTCGCGCCCCAGCAGCAAA TCTAAAACTACTTCGCCGACTACCAGAATTTCAT GCTCAGCGTCCAGTCGATATCGCCCATAATCATC CAGAGGGCCGCTA			
Hypothetical ORF 2 (Predicted)	**ATG**GAAGCAAGTCACTATGAATAGCTTAGCTgta agtctaaattcgatcgtataaactctgctaacaa aatagAACTCGAGGCCCTGACGAAGATTTTGAAG CTCAATTTGACTCGTGAGCAGCTATCTGCATTGA TGGAGTTGACCGAGACGGGCGTGAATCCGGAGGC AATTGCGGCCACAATCGCGGAGATCATG**TGA**	Ti:2141558755: 563-763	+1	None
Hypothetical ORF 3 (Predicted)	**ATG**TCTGATAAATTACCATTTGAGGTTCTTTCAG ACCAGGAAAAGTAGTAAGATGAAATgtaagtcta ataaatacaggaaactttcgctaacaagatagAA TTAAGAAGCTTACGTATATTGCTTCATAGCAGCA ACATATGATCTAACTGTCTGAATTGACATTTAAA TAGTTGGCTAGTTCATTTGATGCAAGCTCAAAAA TGCTTCAATAATATGTTACCGATCTTCATGAGGC AGTCAACATAACTCAGAAATTCAAAAAA**TGA**	Ti:2141517615: 182-450	+2	None

Figure A.3.1. ClustalW2 alignment of ribosomal protein sequences.
Translated nucleotide sequences for *S. vortens* intron-containing RP genes are aligned with RP sequences from various eukaryotes using ClustalW2. RP proteins are represented in single letter amino acid code with an asterisk ('*') indicating an in-frame stop codon. Amino acids which are conserved in four or more organisms are highlighted in black and translated intron sequences are in grey highlighting. Sequences used for alignments are: **(A) RP L7a** – *Homo sapiens* (Hs – NCBI Accession: EAW81017), *Arabidopsis thaliana* (At - NP_191846), *Trypanosoma brucei* (Tb - XP_846969), *Giardia lamblia* (Gl - XP_001706321) and *Spironucleus vortens* (Sv); **(B) RP L30** – *Homo sapiens* (Hs - NP_000980), *Drosophila melanogaster* (Dm - NP_524687), *Trypanosoma cruzi* (Tc - XP_810701), *Saccharomyces cerevisiae* (Sc - NP_011485), *Trichomonas vaginalis* (Tv - XP_001584528) and *Spironucleus vortens* (Sv); **(C) RP S4** - *Homo sapiens* (Hs - EAW71816), *Drosophila melanogaster* (Dm - NP_729871), *Arabidopsis thaliana* (At - NP_001189539), *Saccharomyces cerevisiae* (Sc - NP_012073), *Dictyostelium discoideum* (Dd - XP_644913) and *Spironucleus vortens* (Sv); **(D) RP S12** – *Homo sapiens* (Hs - NP_001007), *Saccharomyces cerevisiae* (Sc - NP_015014), *Dictyostelium discoideum* (Dd - XP_638666), *Arabidopsis thaliana* (At - AAD15398), *Trypanosoma brucei* (Tb - XP_828505), *Encephalitozoon cuniculi* (Ec – CAD24962); *Spironucleus vortens* (Sv) **(E) RP S24** - *Homo sapiens* (Hs - EAW54618), *Drosophila melanogaster* (Dm - NP_611693), *Saccharomyces cerevisiae* (Sc - NP_012195), *Arabidopsis thaliana* (At - NP_187143), *Perkinsus marinus* (Pm - XP_002766228), *Tetrahymena thermophila* (Tt - XP_001025062), *Thalassiosira pseudonana* (Tp - XP_002291077) *Giardia lamblia* (Gl - XP_001709806), *Spironucleus vortens* (Sv).

(A) Ribosomal protein L7a (*Rpl7a*)

```
Hs                --------MPKGKKAKGKKVAPAPAVVKKQEAKKVVNPLFEKRPKNFGIGQDIQPKRDLT 52
At                -------MAPK----KGVKVAS------KKKPEKVTNPLFERRPKQFGIGGALPPKKDLS 43
Tb                MAGKEVKKAVKPTKKAGVPYKKPEVTQKKAKASAAAPSPFVARPKDFGIGRDVPYARDLS 60
Gl                ------------------------------MSKVSGSDIKRALAVPENKSRSKCDFDLT 29
Sv                ---------------------------------------MHRCQAEHVN--------NLT 13
Sv (with intron)----------------------------------------MHRCQAEHVN--------NLT 13

Hs                RFVKWPRYIRLQRQRAILYKRLKVPPAINQFTQALDRQTATQLLKLAHKYRPETKQEKKQ 112
At                RYIKWPKSIRLQRQKRILKQRLKVPPALNQFTKTLDKNLATSLFKILLKYRPEDKAAKKE 103
Tb                RFMRWPTFVTMQRKKRVLQRRLKVPPALNQFTKVLDRSSRNELLKLVKKYAPETRKARRD 120
Gl                PFVRWPRQVRIQRQKAVLQRRLKVPPTVNQFMNPISRNLTNEIFNLARKYSPESKEEHKA 89
Sv                RMVKWPAYIRIQRQKALLQHRLKVPGVVNMFRNPLNANATKEILKFAAKYQPETKEARQQ 73
Sv (with intron)RMVKWPAYIRIQRQKALLQHRLKVPGVVNMFRNPLNANATKEILKFAAKYQPETKEARQQ 73

Hs                RLLARA-EKKAAGKGDVPTKRPPVLRAGVNTVTTLVENKKAQLVVIAHDVDPIELVVFLP 171
At                RLLNKA-QAEAEGK-PAESKKPIVVKYGLNHVTYLIEQNKAQLVVIAHDVDPIELVVWLP 161
Tb                RLTKVAEEKKKNPKGTVSTKAPLCVVSGLQEVTRTIEKKTARLVLIANNVDPIELVLWMP 180
Gl                RLLQIA-DAKANGKPLPEKSNKLVIASGIRRITSLVESKRAKLVLIANDVDPLELVLWLP 148
Sv                RLVQAA-DKKTTIN-APVSFN-----YNIHKVVEAVEKKEAKLVLIAHDVDPIELVLYLP 126
Sv (with intron)RLVQAA-DKKTTIN-APVSFN-----YNIHKVVEAVEKKEAKLVLIAHDVDPIEVSLNQL 126

Hs                ALCRKMGVPYCIIKGKARLGRLVHRKTCTTVAFTQVNSEDKGALAKLVEAIRTNYNDRYD 231
At                ALCRKMEVPYCIVKGKSRLGAVVHQKTAAALCLTTVKNEDKLEFSKILEAIKANFNDKYE 221
Tb                TLCRANKIPYAIVKDKARLGDAIGRKTATCVAFTDVNAEDQAALKNLTRSVNARFLARSD 240
Gl                TLCHKMGVPYAIVRTKGDLGKLVHLKKTTSVCFTDVNPEDKPTFDKILAAVAH--EVDYA 206
Sv                TLCHKNNIPYAIVRSRTELGKLVHCTKCTSIAFTTIKPEDTAAFKSILDTVAH--EVDYV 184
Sv (with intron)*

Hs                EIRRHWGGNVLGPKSVARIAKL-EKAKAKELATKLG 266
At                EYRKKWGGGIMGSKSQAKTKAK-ERVIAKEAAQRMN 256
Tb                VIRRQWGGLQLSLRSRAELRKKRARTAGNDAAAKAA 276
Gl                KAMKTYGGGVRRED----EAQ--------------- 223
Sv                HAIKTHGGVSRSNKSLAKEAKKNKIGKK-------- 212
```

(B) Ribosomal protein L30 (*Rpl30*)

```
Hs                  MVAAKKTKKSLESINSRLQLVMKSGKYVLGYKQTLKMIRQGKAKLVILANNCPALRKSEI 60
Dm                  MVAVKKQKKALESTNARLALVMKSGKYCLGYKQTLKTLRQGKAKLVLIASNTPALRKSEI 60
Tc                  --MAKKNKTKVDTINTKIQLVMKSGKYVLGTKQTLKTLRQGRSKLVVISSNCPPIRKAEI 58
Sc                  MAPVKS----QESINQKLALVIKSGKYTLGYKSTVKSLRQGKSKLIIIAANTPVLRKSEL 56
Tv                  -MGRKKLSRARESIINSLSLVTKSGKYSLGISQTLKSLRNGEAKLVIFASNVAPADRSLI 59
Sv                  MDRVSK--KSSESAALQLALVVKSGKYTLGVNQALKSIRNLKAKLVIITSNLPPLVASQI 58
Sv (with intron)MDRVSVS*

Hs                  EYYAMLAKTGVHHYSGNNIELGTACGKYYRVCTLAIIDPGDSDIIRSMPEQTGEK 115
Dm                  EYYAMLAKTEVQHYSGTNIELGTACGKYFRVCTLSITDPGDSDIIRSLETA---- 111
Tc                  EYYCTLSKTPMHHYAGNNLDLGTACGRHFRSCVLSITDVGDSDITSA-------- 105
Sc                  EYYAMLSKTKVYYFQGGNNELGTAVGKLFRVGVVSILEAGDSDILTTLA------ 105
Tv                  EYYAMLSGCDILPFDGDNVDLGTACGKYFRSSVISIIDAGESEILKMIKQKDE-- 112
Sv                  EYLCMLSGIPVHAFPSNSREFGVTLGKQFNVGVMAVTEAGDADLAAF-------- 105
```

(C) Ribosomal protein S4 (*Rps4*)

```
Dm                  MARGPKKHLKRLAAPKAWMLDKLGGVFAPRPSTGPHKLRESLPLLIFLRNRLKYALNGAE 60
Sc                  MARGPKKHLKRLAAPHHWLLDKLSGCYAPRPSAGPHKLRESLPLIVFLRNRLKYALNGRE 60
Hs                  MARGPKKHLKRVAAPKHWMLDKLTGVFAPRPSTGPHKLRECLPLIIFLRNRLKYALTGDE 60
Dd                  MARGPKKHLKRLAAPNHWMLDKLSGKWAPRPSSGPHKLRECLPLILVLRNRLKYALTKKE 60
At                  MARGLKKHLKRLNAPKHWMLDKLGGAFAPKPSSGPHKSRECLPLVLIIRNRLKYALTYRE 60
Sv                  MARGPRKHLKRLNAPKHWLLDKMGGIWAPRPTNGPHGLRECIPLILILRNRLHYANTYAE 60
Sv (with intron)MVSLKCVRTAHHLFA*

Dm                  VTKIVMQRLVKVDGKVRTDPTYPAGYMDVITLEKTGEFFRLVYDVKGRFVIHRISAEEAK 120
Sc                  VKAILMQRHVKVDGKVRTDTTYPAGFMDVITLDATNENFRLVYDVKGRFAVHRITDEEAS 120
Hs                  VKKICMQRFIKIDGKVRTDITYPAGFMDVISIDKTGENFRLIYDTKGRFAVHRITPEEAK 120
Dd                  VTLILMQRLVKVDGKVRTDPNYPAGFMDVISIEKTKENFRLLFDPKGRFTLQRITPEEAK 120
At                  VISILMQRHIQVDGKVRTDKTYPAGFMDVVSIPKTNENFRLLYDTKGRFRLHSIKDEEAK 120
Sv                  TNMILKDKNVLIDNKPRIDPTFPIGFMDVFEIPKVHKVFRVLYDVKGRFTLIPIQSNEAG 120

Dm                  YKLCKVKKTQLGAKGVPFLVTHDGRTIRYPDPLIHANDSVQVDIASGKITDYIKFDSGNL 180
Sc                  YKLGKVKKVQLGKKGVPYVVTHDGRTIRYPDPNIKVNDTVKIDLASGKITDFIKFDAGKL 180
Hs                  YKLCKVRKIFVGTKGIPHLVTHDARTIRYPDPLIKVNDTIQIDLETGKITDFIKFDTGNL 180
Dd                  FKLARVTRVETGNQGIPYVHTDDGRTIRYPDPAISIHDTIKIDIESGKITAFIPFEVNNL 180
At                  FKLCKVRSIQFGQKGIPYLNTYDGRTIRYPDPLIKPNDTIKLDLEENKIVEFIKFDVGNV 180
Sv                  FKLCRVQKIFLGDKGMPYLVTHDARTIRFPHPDIKTNDTIKINLKTGKIDEWYKFDIGKV 180

Dm                  CMITGGRNLGRVGTVVNRERHPGSFDIVHIKDSQGHVFATRLTNVFIIGKGNKPYISLPK 240
Sc                  VYVTGGRNLGRIGTIVHKERHDGGFDLVHIKDSLDNTFVTRLNNVFVIGEQGKPYISLPK 240
Hs                  CMVTGGANLGRIGVITNRERHPGSFDVVHVKDANGNSFATRLSNIFVIGKGNKPWISLPR 240
Dd                  CMIVGGHNLGRVGAVTHREKHPGSFDIVHVTDTAGHQFATRLSNVFIIGKASQTFVSLPA 240
At                  VMVTGGRNRGRVGVIKNREKHKGSFETIHIQDSTGHEFATRLGNVYTIGKGTKPWVSLPK 240
Sv                  VMVTGGRNCGRIGTIQAIDKHMGSYTMIRMKDSEGAEFITRLCNVFIIGNDS-PAVTVPS 239

Dm                  GKGVKLSIAEERDKRLAAKTH------ 261
Sc                  GKGIKLSIAEERDRRRAQQGL------ 261
Hs                  GKGIRLTIAEERDKRLAAKQSSG---- 263
Dd                  GKGVRRSRVDERNAALKRRGEKIETVA 267
At                  GKGIKLTIIEEARKRLSAQQA------ 261
Sv                  TKGIRPDIIKNRELRLRSIAK------ 260
```

(D) Ribosomal protein S12 (*Rps12*)

```
Hs                  ----------MAEEGIAAG-GVMDVN----TALQEVLKTALIHDGLARGIREAAKALD-- 43
Sc                  MS-DVEEVVEVQEETVVEQTAEVTIE----DALKVVLRTALVHDGLARGLRESTKALT-- 53
Dd                  --------MEGDAPVIAANPLEKNTDP--MVALQKVIKESLAVQGVARGLHETVKALD-- 48
At                  MSGDEAVAAPVVPPVAEAAVIPEDMDV--STALELTVRKSRAYGGVVRGLHESAKLIE-- 56
Tb                  MAEETSLVADKVPEPAVIDAVADAMPDSLEDALRIVLMKARETNGLICGLSEVTRALD-- 58
Ec                  ------------MSEMQEPMMEPEMTL--QEALSKVCKVSRTYCKLSRGAKETTKKML-- 44
Sv                  ------------------------MST---DQLKTFCKKIRVHGAMVSGVRQVVRAVENH 33
Sv (with intron)------------------------MST---QVQS*
```

```
Hs              -KRQAHLCVLASNCDEPMYVKLVEALC--AEHQINLIKVDDNKKLGEWVGLCKID---RE 97
Sc              -RGEALLVVLVSSVTEANIIKLVEGLANDPENKVPLIKVADAKQLGEWAGLGKID---RE 109
Dd              -KRTARLCVLASNCDEPNFVRLVKALA--TEHNIPLIEVPDNKALGEWAGLCKLD---KD 102
At              -KRNAQLCVLAEDCNQPDYVKLVKALC--ADESIKLLTVPSAKTLGEWAGLCKID---SE 110
Tb              -RRTAHLCVLADDCEDEEYKKLVTALAK--QNNIDLVSMDEREKLAQWAGLTRMA---AD 112
Ec              -ADKMSFVMVAENA-EPRISKLVMALAK--KKNIPVISIGSCLELGRIVGVENVS---SS 97
Sv              ATSNVKVILLANDCKEAGIKNLVKALAK--QHSIGVCEKFGAAHLGELAHQYVIKGHVTE 91

Hs              GKPRKVVGCSCVVVKDYGKESQAKDVIEEYFKCKK--- 132
Sc              GNARKVVGASVVVVKNWGAETDELSMIMEHFSQQ---- 143
Dd              LAARKVVACSTLVIKTFGKESDDYKFLMEYISKQ---- 136
At              GNARKVVGCSCLVIKDFGEETTALNIVKKHLDSN---- 144
Tb              GSVRKTLKCSCLAVRDFGERTKALDYLLSQLQ------ 144
Ec              GKVR-SKGCCVAGVQDYCEQTSEAGFVQAALLKGISSQ 134
Sv              GKIGKVRNASCMAIQNFGTLIAEDQAAFNALLQ----- 124
```

(E) Ribosomal protein S24 (RP S24)

```
Hs              ----MN-DTVTIRTRKFMTNRLLQRKQMVIDVLHPGKA-TVPKTEIREKLAKMYKTTP-D 53
Dm              ----MSGTTATIRTRKFMTNRLLARKQMVCDVLHPGLS-SVNKTEIREKLAAMYKVTP-D 54
Sc              ----MSD-AVTIRTRKVISNPLLARKQFVVDVLHPNRA-NVSKDELREKLAEVYKAEK-D 53
At              ----MAEKAVTIRTRKFMTNRLLSRKQFVIDVLHPGRA-NVSKAELKEKLARMYEVKDPN 55
Pm              ------MAEFTVRTRKFITNPLLGRKQFVVDVLHPGVG-SVSKKDLADSLAKMYKVKDAR 53
Tt              -------MTIVIRTKKILVNPLLSRRQLSLDVLHPDSP-TASKEKIREELAKQLKVDA-R 51
Tp              ----MSDQSVVVKTRKFMKNPLLARRQMIVDIIHPGRA-NVAKSELQEVVGGMHKTDS-K 54
Gl              ------MPEITVKVRKVLNNPLLQRQQCVVDVLHPGCT-YESKEAIKAKVAQQLKVADQK 53
Sv              ---------MQIKYREIVNNPILDRTQMKLKIVHPGKS-VGTIEALRELVQKDRKIKDIK 50
Sv (with intron)---------MQIKYREIVNNPILDRTQMVSLNLMYNQYQQVRSSRSSTQVSPWVPSRLSA

Hs              VIFVFGFRTHFGGGKTTGFGMIYDSLDYAKKNEPKHRLARHGLYEKKKT-SRKQRKERKN 112
Dm              VVFAFGFRTNFGGGRSTGFALIYDTLDFAKKFEPKYRLARHGLFEQKKQ-TRKQRKERRN 113
Sc              AVSVFGFRTQFGGGKSVGFGLVYNSVAEAKKFEPTYRLVRYGLAEKVEKASRQQRKQKKN 113
At              AIFVFKFRTHFGGGKSSGFGLIYDTVESAKKFEPKYRLIRNGLDTKIEK-SRKQIKERKN 114
Pm              VISLFGFKTQFGGGRSTGFGLIYDTVEKAQAFEPKHRLRRHGLAP-EFQAKRRSYKELKN 112
Tt              NVVVYGFSTQYGGGKSTGFALVYDNQQYLLKYEPNYRLRKVKILGEKPN-TRRSFKELKR 110
Tp              LVVLFGFRTKFGGGKSTGFCVIYDNEDALRKFEPKHRLVRLGLEDKKDR-SRKAMKEAKN 113
Gl              NIVLYGFKTSFGGGHTVGFCNAYQNMDALMKYEPGFRKIRCGLIEAPKPVSRKQLKNLKN 113
Sv              QVVVFDCHTKHGGNLSTASCHIYGNVETLKKVEPKYTIIRLGYIEKPKPVSRKMIKNHKN 110
Sv (with intron)SSSRRIVRSRTSSRLSSLTATPSTVVTSALLLATSTATLRP*

Hs              RMKKVRGTAKANVGAGKKPKE------------------ 133
Dm              RMKKVRGTAKAKIGTGKK---------------------- 131
Sc              RDKKIFGTGKRLAKKVARRNAD----------------- 135
At              RAKKIRGVKKTKAGDAKKK--------------------- 133
Pm              KCKKVRGTAKSKLR-SK----------------------- 128
Tt              KIKRTSGKAITKLLSEKKGDTWASVQSKKSDHLKNFVAK 149
Tp              KGKKTRGTGASVAKHKAKRAANSD---------------- 137
Gl              RRLKKRGTEKATVTLGAKK--------------------- 132
Sv              KLIRKFGTAKSKIVMSGKKN-------------------- 130
```

Figure A.3.2. ClustalW2 alignment of *S. vortens* gene alleles containing intron sequences.

Genomic trace database sequences were searched on the NCBI website using blastn and all sequences encoding *S. vortens* gene paralogs containing introns were obtained. Intron sequences and their flanking upstream and downstream exonic sequences were then aligned using ClustalW2. Unique gene paralog sequences are indicated in differential highlighting with intron sequences in red text. Only as single genomic trace for the *Rpl30* gene was identified and thus was excluded from the alignments.

A) Ribosomal Protein L7a (*Rpl7a*)

```
gnl|ti|2141614280_[429_117]       AGAAGGAGGCCAAGCTGGTCCTCATCGCCCACGACGTCGACCCAATCGAG 50
gnl|ti|2141629602_[361_673]       AGAAGGAGGCCAAGTTGGTCCTCATCGCCCACGACGTCGACCCAATCGAG 50
gnl|ti|2141515448_[326_638]       AGAAGGAGGCCAAGCTGGTCCTCATCGCCCACGACGTCGACCCAATCGAG 50
gnl|ti|2141473674_[1022_710]      GAAAGGAGGCCAAGCTGTTCTTCATCGCCCACGACGTCGACCCAATCGAG 50
                                    ************ ** ** ****************************

gnl|ti|2141614280_[429_117]       GTAAGTCTAAACTAACTGTGATCCGCGGATCGCTTGAGTTACGAAACTTT 100
gnl|ti|2141629602_[361_673]       GTAAGTCTAAACTAACTGTGATCCGCGGATCGCTTGAGTTACGAAACTTT 100
gnl|ti|2141515448_[326_638]       GTAAGTCTAAACTAACTGTGATCCGCGGATCGCTTGAGTTACGAAACTTT 100
gnl|ti|2141473674_[1022_710]      GTAAGTCTAAACTAACTGTGATCCGCGGATCGCTGAGTTTACGAAACTTT 100
                                  ********************************    ************

gnl|ti|2141614280_[429_117]       GCTAACAAACTAGCTCGTCCTGTACCTGCCAACCCTCTGCCACAAGAACA 150
gnl|ti|2141629602_[361_673]       GCTAACAAACTAGCTCGTCCTGTACCTGCCAACCCTCTGCCACAAGAACA 150
gnl|ti|2141515448_[326_638]       GCTAACAAACTAGCTCGTCCTGTACCTGCCAACCCTCTGCCACAAGAACA 150
gnl|ti|2141473674_[1022_710]      GCTAACAAACTAGCTCGTCCTGTACCTGCCAACCCTCTGCCACAAGAACA 150
                                  **************************************************

gnl|ti|2141614280_[429_117]       ACATCCCATATGCCATCGTTCGCTCCCGCACCGAGCTCGGCAAGCTGGTT 200
gnl|ti|2141629602_[361_673]       ACATCCCTTATGCCATCGTTCGCTCCCGCACCGAGCTCGGCAAGCTGGTT 200
gnl|ti|2141515448_[326_638]       ACATCCCATATGCCATCGTTCGCTCCCGCACCGAGCTCGGCAAGCTGGTT 200
gnl|ti|2141473674_[1022_710]      ACATCCCATATGCCATCGTTCGCTCCCGCACGAAGCTCGGCAAGCTGGTT 200
                                  ******* ********************  ******************

gnl|ti|2141614280_[429_117]       CACTGCACCAAGTGCACCTCCATCGCCTTCACCACCATCAAGCCGGAGGA 250
gnl|ti|2141629602_[361_673]       CACTGCACCAAGTGCACCTCCATCGCCTTCACCACCATCAAGCCGGAGGA 250
gnl|ti|2141515448_[326_638]       CACTGCACCAAGTGCACCTCCATCGCCTTCACCACCATCAAGCCGGAGGA 250
gnl|ti|2141473674_[1022_710]      CACTGCACCAAGTGCACCTCCATCGCCTTCACCACCATCAAGCCGGAGGA 250
                                  **************************************************

gnl|ti|2141614280_[429_117]       CACCGCCGCCTTCAAGTCCATCCTGGACACCGTCGCCCACGAGGTCGACT 300
gnl|ti|2141629602_[361_673]       CACCGCCGCCTTCAAGTCCATCCTGGACACCGTCGCCCACGAGGTCGACT 300
gnl|ti|2141515448_[326_638]       CACCGCCGCCTTCAAGTCCATCCTGGACACTGTTGCCCACGAGGTCGACT 300
gnl|ti|2141473674_[1022_710]      CACCGCCGCCTTCAAGTCCATCCTGGACACTGTTGCCCACGAGGTCGACT 300
                                  ****************************** ** ****************

gnl|ti|2141614280_[429_117]       ACGTCCACGCCAT 313
gnl|ti|2141629602_[361_673]       ACGTCCACGCCAT 313
gnl|ti|2141515448_[326_638]       ACGTCCACGCCAT 313
gnl|ti|2141473674_[1022_710]      ACGTCCACGCCAT 313
                                  *************
```

B) Ribosomal Protein S4 (*Rps4*)

```
gnl|ti|2141479638_[515_772] TTTTTTACATAAAAGTGTGAAAAATATA-C---AGTT------TTGTACA 40
gnl|ti|2141550682_[26_283]  -TTTTTACATAAAAGTGTGAAAAATATNAC---AGTT------TTGTACA 40
gnl|ti|2141495195_[88_345]  -TTATAATGCAGAAATATGAAAA-TATAACTAGAATCATTATATCATACA 48
gnl|ti|2141536103_[796_539] -TTATAATGCAGAAATATGAAAA-TATAACTAGAATCATTAGATCATACA 48
                             ** * *   * ** * ****** ***  *   * *       *  ****

gnl|ti|2141479638_[515_772] AATAAAATATTTATATAAATACTATATGTTTCACACCTTTTCTTTTGTTG 90
gnl|ti|2141550682_[26_283]  AATAAAATATTTATATAAATACTATATGTTTCACACCTTTTCTTTTGTTG 90
gnl|ti|2141495195_[88_345]  TAGAAAAT-TTCATAT---CACGAT---TTCCACACCTTTTCTTTTGATA 91
gnl|ti|2141536103_[796_539] TAGATAAT-TTCACGT---CACGAT---TTCCACACCTTTTCTTTTGATA 91
                             * * *** ** *  *    ** **   ** ***************** *
```

```
gnl|ti|2141479638_[515_772] TGATAACATGGTAAGTCTAAAATGTGTGCGCACGGCGCATCATCTATTTG 140
gnl|ti|2141550682_[26_283]  TGATAACATGGTAAGTCTAAAATGTGTGCGCACGGCGCATCATCTATTTG 140
gnl|ti|2141495195_[88_345]  TC-TATTATGGTAAGTCTAAAATGTGTGCGCACGCCGCATCATATATTTT 140
gnl|ti|2141536103_[796_539] TC-TATTATGGTAAGTCTAAAATGTGTGCGCACGCCGCATCATATATTTT 140
                            *  **  ************************** ******** *****

gnl|ti|2141479638_[515_772] CTTGAACTAACAAACTAGGCTCGTGGTCCAAAACTTCATATGAAACGTCT 190
gnl|ti|2141550682_[26_283]  CTTGAACTAACAAACTAGGCTCGTGGTCCAAAACTTCATATGAAACGTCT 190
gnl|ti|2141495195_[88_345]  CTTGAACTAACAAGCTAGGCTCGTGGTCCAAAACTTCATATGAAACGCCT 190
gnl|ti|2141536103_[796_539] CTTGAACTAACAAGCTAGGCTCGTGGTCCAAAACTTCATATGAAACGCCT 190
                            ************* ********************************** **

gnl|ti|2141479638_[515_772] TAACGCTCCATCCCACTGGCAGCAGGACAAGCTTGGCGGCATCTACTCCA 240
gnl|ti|2141550682_[26_283]  TAACGCTCCATCCCACTGGCAGCAGGACAAGCTTGGCGGCATCTACTCCA 240
gnl|ti|2141495195_[88_345]  TAACGCTCCATCCCACTGGTAGCAGGACAAGCTTGGTGGCATTTACTCCA 240
gnl|ti|2141536103_[796_539] TAACGCTCCATCCCACTGGTAGCAGGACAAGCTTGGTGGCATTTACTCCA 240
                            ****************** ***************** ***** *******

gnl|ti|2141479638_[515_772] CCAAGTGCAACCTCTCCA 258
gnl|ti|2141550682_[26_283]  CCAAGTGCAACCTCTCCA 258
gnl|ti|2141495195_[88_345]  CCAAGTGCAACCTCTCCA 258
gnl|ti|2141536103_[796_539] CCAAGTGCAACCTCTCCA 258
                            ******************
```

C) Ribosomal Protein S12 (*Rps12*)

```
gnl|ti|2141614914_[152_392] TCCACATCCTGAGACAGCATTTATCGAACTATAATTTTGTCTAAATTTAA 50
gnl|ti|2141634934_[500_260] TCCACATCCTGAGACAGCATTTATCGAACTATAATTTTGTCTAAATTTAA 50
gnl|ti|2141503180_[211_451] --CCCATCCTGAGACAGCATTTATCGAACTATAATTTTGTCTAAATTTAA 48
gnl|ti|2141498925_[552_792] -CCACATCCTGAGACAGCATTTATCGAACTATAATTTTGTCTAAATTTAA 49
gnl|ti|2141599697_[595_355] -CCACATCCTGAGACAGCATTTATCGAACTATAATTTTGTCTAAATTTAA 49
                              * ********************************************

gnl|ti|2141614914_[152_392] -TGAATATCC-AAATAATTTCCATACCTTTTCTTGTGAAGATACATGTCA 98
gnl|ti|2141634934_[500_260] -TGAATATCC-AAATAATTTCCATACCTTTTCTTGTGAAGATACATGTCA 98
gnl|ti|2141503180_[211_451] ATGAATATCCCAAATAATTTCCATACCTTTTCTTGTGAAGATACATGTCA 98
gnl|ti|2141498925_[552_792] ATGAATATCC-AAATAATTTCCATACCTTTTCTTGTGAAGATACATGTCA 98
gnl|ti|2141599697_[595_355] ATGAATATCC-AAATAATTTCCATACCTTTTCTTGTGAAGATACATGTCA 98
                             ********* ***************************************

gnl|ti|2141614914_[152_392] ACGTAAGTCTAGAGCTGAGCAGTCAACTTTACTAACAAAATAGTGACCAA 148
gnl|ti|2141634934_[500_260] ACGTAAGTCTAGAGCTGAGCAGTCAACTTTACTAACAAAATAGTGACCAA 148
gnl|ti|2141503180_[211_451] ACGTAAGTTTAGAGCTGAGCAGTCAACTTTACTAACAAAATAGTGACCAA 148
gnl|ti|2141498925_[552_792] ACGTAAGTCTAGAGCTGAGCAGTCAACTTTACTAACAAAATAGTGACCAA 148
gnl|ti|2141599697_[595_355] ACGTAAGTCTAGAGCTGAGCAGTCAACTTTACTAACAAAATAGTGACCAA 148
                            ******** ***************************************

gnl|ti|2141614914_[152_392] CTGAAGACTTTCTGCAAGAAGATCCGCGTCCACGGCGCGATGGTCTCCGG 198
gnl|ti|2141634934_[500_260] CTGAAGACTTTCTGCAAGAAGATCCGCGTCCACGGCGCGATGGTCTCCGG 198
gnl|ti|2141503180_[211_451] CTGAAGACTTTCTGCAAGAAGATCCGCGTCCACGGCGCGATGGTCTCCGG 198
gnl|ti|2141498925_[552_792] CTGAAGACTTTCTGTAAGAAGATCCGCGTCCACGGTGCTATGGTCTCCGG 198
gnl|ti|2141599697_[595_355] CTGAAGACTTTCTGTAAGAAGATCCGCGTCCACGGTGCTATGGTCTCCGG 198
                            ************** ******************** ** ***********

gnl|ti|2141614914_[152_392] CGTCCGCCAGGTCGTGCGCGCCGTCGAGAACCACGCCACCTCC 241
gnl|ti|2141634934_[500_260] CGTCCGCCAGGTCGTGCGCGCCGTCGAGAACCACGCCACCTCC 241
gnl|ti|2141503180_[211_451] CGTCCGCCAGGTCGTGCGCGCCGTCGAGAACCACGCCACCTCC 241
gnl|ti|2141498925_[552_792] CGTCCGCCAGGTCGTGCGCGCCGTCGAGAACCACGCCACCTCC 241
gnl|ti|2141599697_[595_355] CGTCCGCCAGGTCGTGCGCGCCGTCGAGAACCACGCCACCTCC 241
                            *******************************************
```

D) Ribosomal Protein S24 (*Rps24*)

```
gnl|ti|2141541737_[73_313]  ATTATTCAAACGAATGTTTTCTTATCTTTTCTTTTGTGGCCTAATGCAGA 50
gnl|ti|2141586865_[578_818] ATTATTCAAACGAATGTTTTCTTATCTTTTCTTTTGTGGCCTAATGCAGA 50
gnl|ti|2141602726_[494_254] ATTATTCAAACGAATGTTTTCTTATCTTTTCTTTTGTGGCCTAATGCAGA 50
gnl|ti|2141597815_[404_164] ATTATTCAAACGAATGTTTTCTTATCTTTTCTTTTGTGGCCTAATGCAGA 50
                            **************************************************

gnl|ti|2141541737_[73_313]  TCAAGTATCGCGAAATTGTCAACAACCCGATCCTCGATCGTACTCAAATG 100
gnl|ti|2141586865_[578_818] TCAAGTATCGCGAAATTGTCAACAACCCGATCCTCGATCGTACTCAAATG 100
gnl|ti|2141602726_[494_254] TCAAGTATCGCGAAATTGTCAACAACCCGATCCTCGATCGTACTCAAATG 100
gnl|ti|2141597815_[404_164] TCAAGTATCGCGAAATTGTCAACAACCCGATCCTCGATCGTACTCAAATG 100
                            **************************************************

gnl|ti|2141541737_[73_313]  GTAAGTCTAAATCTCATGTATAACTAATACTAACAAGTTAGAAGCTCAAG 150
gnl|ti|2141586865_[578_818] GTAAGTCTAAATCTCATGTATAACTAATACTAACAAGTTAGAAGCTCAAG 150
gnl|ti|2141602726_[494_254] GTAAGTCTAAATCTCATGTATAACTAATACTAACAAGTTAGAAGCTCAAG 150
gnl|ti|2141597815_[404_164] GTAAGTCTAAATCTCATGTATAACTAATACTAACAAGTTAGAAGCTCAAG 150
                            **************************************************

gnl|ti|2141541737_[73_313]  ATCGTCCACCCAGGTAAGTCCGTGGGTACCATCGAGGCTCTCCGCGAGCT 200
gnl|ti|2141586865_[578_818] ATCGTCCACCCAGGTAAGTCCGTGGGTACCATCGAGGCTCTCCGCGAGCT 200
gnl|ti|2141602726_[494_254] ATCGTCCACCCAGGTAAGTCCGTGGGTACCATCGAGGCTCTCCGCGAGCT 200
gnl|ti|2141597815_[404_164] ATCGTCCACCCAGGTAAGTCCGTGGGTACCATCGAGGCTCTCCGCGAGCT 200
                            **************************************************

gnl|ti|2141541737_[73_313]  CGTCCAGAAGGATCGTAAGATCAAGGACATCAAGCAGGTTG 241
gnl|ti|2141586865_[578_818] CGTCCAGAAGGATCGTAAGATCAAGGACATCAAGCAGGTTG 241
gnl|ti|2141602726_[494_254] CGTCCAGAAGGATCGTAAGATCAAGGACATCAAGCAGGTTG 241
gnl|ti|2141597815_[404_164] CGTCCAGAAGGATCGTAAGATCAAGGACATCAAGCAGGTTG 241
                            *****************************************
```

E) Bifunctional folylpolyglutamate synthase-like gene (*FolC-like*)

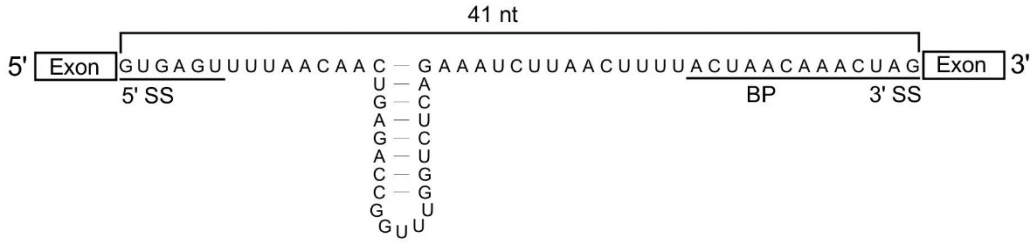
```
gnl|ti|2141479887_[559_319] ACTGATAAGAAAAATTAATATGTTACTATTTAACTTTATTACGCATATAA 50
gnl|ti|2141588169_[695_455] ACTGATAAGAAAAATTAATATGTTACTATTTAACTTTATTACGCATATAA 50
gnl|ti|2141538557_[649_889] ACTGATAAGAAAAATTAATATGTTACTATTTAACTTTATTACGCATATAA 50
gnl|ti|2141671053_[509_749] ACTGATAAGAAAAATAAATATGTTGCTATTTAACTTTATTATGGATATAA 50
gnl|ti|2141610787_[847_607] CNTGATAAGAAAAATAAATATGTTACTATTTAACTTTATAATGTATATAA 50
                              ************* ******** ************** * * ******

gnl|ti|2141479887_[559_319] ATTAAGAATTTCACTATTCGACATTAACG**ATG**TAGTACCCTCAAGTGCTT 100
gnl|ti|2141588169_[695_455] ATTAAGAATTTCACTATTCGACATTAACGATGTAGTACCCTCAAGTGCTT 100
gnl|ti|2141538557_[649_889] ATTAAGAATTTCACTATTCGACATTAACGATGTAGTACCCTCAAGTGCTT 100
gnl|ti|2141671053_[509_749] ATTGAACAATTCACTATTCGACATTCACGATGTAGTACCCTCAAGTCCTT 100
gnl|ti|2141610787_[847_607] ATTGAAAAATTCACTATTCGACATTAACGATGTAGTACCCTCAAGTGCTT 100
                             *** * * ***************** ******************** ***

gnl|ti|2141479887_[559_319] GTAAGTCAACTTTTGCCATCAAACTTTTGCTAACAAATTAGGATTCCCTC 150
gnl|ti|2141588169_[695_455] GTAAGTCAACTTTTGCCATCAAACTTTTGCTAACAAATTAGGATTCCCTC 150
gnl|ti|2141538557_[649_889] GTAAGTCAACTTTTGCCATCAAACTTTTGCTAACAAATTAGGATTCCCTC 150
gnl|ti|2141671053_[509_749] GTAAGTCAACTTTTGTCATCAAACTTTTGCTAACAAATTAGGATTCCCTC 150
gnl|ti|2141610787_[847_607] GTAAGTCAACTTTTGTCATCAAACTTTTGCTAACAAATTAGGATTCCATC 150
                            *************** ******************************** **

gnl|ti|2141479887_[559_319] TCAAAAGTATCATAAACTGTCAAGTCGCATTGGTCATAGCTACCAGATCT 200
gnl|ti|2141588169_[695_455] TCAAAAGTATCATAAACTGTCAAGTCGCATTGGTCATAGCTACCAGATCT 200
gnl|ti|2141538557_[649_889] TCAAAAGTATCATAAACTGTCAAGTCGCATTGNTCATAGCTACCAGATCT 200
gnl|ti|2141671053_[509_749] TCANAAGTATCATAAACTGTTAAGTCGCATTGGTCATAGCTACCAGATCT 200
gnl|ti|2141610787_[847_607] TCAAAAGTATCATAAACTGTCAAGTCGCATTGGTCATAGCTACCAGATTT 200
                            *** ***************** *********** *************** *

gnl|ti|2141479887_[559_319] ACTAAATAAGTAATTTCGACCGGAATTGCTGTTTCACGTCA 241
gnl|ti|2141588169_[695_455] ACTAAATAAGTAATTTCGACCGGAATTGCTGTTTCACGTCA 241
gnl|ti|2141538557_[649_889] ACTANATAAGTAATTTCGACCGGGATTGCTGTTTCACGTCA 241
gnl|ti|2141671053_[509_749] ACTAAATAAGTAATTTCGACCGGAATTGCTGTTTCACGTCA 241
gnl|ti|2141610787_[847_607] ACTAAATAAGTAATTTCGACCGGAATTGCTGTTTCACGTCA 241
                            **** ****************** *****************
```

F) Hypothetical ORF #1

```
gnl|ti|2141628303_[520_739]      -TTGTATTGGTAATGATACAATATATTGAAATATTTTCGTTTCCTCCGAA 49
gnl|ti|2141664662_[809_509]      -TTGTATTGGTAATGATACAATATATTGAAATATNTTCGTTTCCTCCGAA 49
gnl|ti|2141612529_[78_297]       -TTGTATTGGTAATGATACAATATATTGAAATATTTTCGTTTCCTCCGAA 49
gnl|ti|2141491491_[416_197]      -TTGTATTGGTAATGATACAATATATTGAAATATTTTCGTTTCCTCCGAA 49
gnl|ti|2141541942_[483_702]      -TTGCATTGGTAATGATACAATATATTGTAATATTTTCGTTTCCTCCGAA 49
gnl|ti|2141510109_[430_211]      -TTGCATTGGTAATGATACAATATATTGTAATATTTTCGTTTCCTCCGAA 49
gnl|ti|2141636533_[463_682]      -TTGCATTGGTAATGATACAATATATTGTAATATTTTCGTTTCCTCCGAA 49
gnl|ti|2141657616_[782_945]      ATTGCATTGGTAATGATACCATATATTGTAATATTTTCGTTTCCTC-GAA 49
gnl|ti|2141510109_[228_9]        -TTGCATTGGTAATGATACAATATATTGTAATATTTTCGTTTCCTCCGAA 49
                                  *** ************** ******** ***** *********** ***

gnl|ti|2141628303_[520_739]      TGTCCGAAACCTCGTCCTCCAGCGACGCTGGAGACGCTTTCGAGTAATAT 99
gnl|ti|2141664662_[809_509]      TGTCCGAAACCTCGTCCTCCAGCGACGCTGGAGACGCTTTCGAGTAATAT 99
gnl|ti|2141612529_[78_297]       TGTCCGAAACCTCGTCCTCCAGCGACGCTGGAGACGCTTTCGAGTAATAT 99
gnl|ti|2141491491_[416_197]      TGTCCGAAACCTCGTCCTCCAGCGACGCTGGAGACGCTTTCGAGTAATAT 99
gnl|ti|2141541942_[483_702]      TGTCCGAAACCTCGTCCTCCAGCGACGCTGGAGACGCTTTCGAGTAATAT 99
gnl|ti|2141510109_[430_211]      TGTCCGAAACCTCGTCCTCCAGCGACGCTGGAGACGCTTTCGAGTAATAT 99
gnl|ti|2141636533_[463_682]      TGTCCGAAACCTCGTCCTCCAGCGACGCTGGAGACGCTTTCGAGTAATAT 99
gnl|ti|2141657616_[782_945]      TGTCCGAAACCTCGTCCTCCAGCGACGCTGGAGACGCTTTCGAGTAATAT 99
gnl|ti|2141510109_[228_9]        TGTCCGAAACCTCGTCCTCCAGCGACGCTGGAGACGCTTTCGAGTAATAT 99
                                 *************************************************

gnl|ti|2141628303_[520_739]      CGTAGGTCTAATTGATATTGATAACTTTACTAACAAACTAGTGCAATAGC 149
gnl|ti|2141664662_[809_509]      CGTAGGTCTAATTGATATTGATAACTTTACTAACAAACTAGTGCAATAGC 149
gnl|ti|2141612529_[78_297]       CGTAGGTCTAATTGATATTGATAACTTTACTAACAAACTAGTGCAATAGC 149
gnl|ti|2141491491_[416_197]      CGTAGGTCTAATTGATATTGATAACTTTACTAACAAACTAGTGCAATAGC 149
gnl|ti|2141541942_[483_702]      CGTAGGTCTAATTGATATGGATAACTTTACTAACAAACTAGTGCAATAGC 149
gnl|ti|2141510109_[430_211]      CGTAGGTCTAATTGATATGGATAACTTTACTAACAAACTAGTGCAATAGC 149
gnl|ti|2141636533_[463_682]      CGTAGGTCTAATTGATATGGATAACTTTACTAACAAACTAGTGCAATAGC 149
gnl|ti|2141657616_[782_945]      CGTAGGTCTAAATGATATGGATAACTTTACTAACAAACTAGTGCAATAGC 149
gnl|ti|2141510109_[228_9]        CGTAGGTCTAATTGATATGGATAACTTTACTAACAAACTAGTGCAATAGC 149
                                 ********** ****** ******************************

gnl|ti|2141628303_[520_739]      GCCAGAAAAAGAACCAAGAAATCGAAGAATTAAAGAATCAAACGCAGTAA 199
gnl|ti|2141664662_[809_509]      GCCAGAAAAAGAACCAAGAAATCGAAGAATTAAAGAATCAAACGCAGTAA 199
gnl|ti|2141612529_[78_297]       GCCAGAAAAAGAACCAAGAAATCGAAGAATTAAAGAATCAAACGCAGTAA 199
gnl|ti|2141491491_[416_197]      GCCAGAAAAAGAACCAAGAAATCGAAGAATTAAAGAATCAAACGCAGTAA 199
gnl|ti|2141541942_[483_702]      GCCAGAAAAAGAACCAAGAAATCGAAGAATTAAAGAATCAAACGCAGTAA 199
gnl|ti|2141510109_[430_211]      GCCAGAAAAAGAACCAAGAAATCGAAGAATTAAAGAATCAAACGCAGTAA 199
gnl|ti|2141636533_[463_682]      GCCAGAAAAAGAACCAAGAAATCGAAGAATTAAAGAATCAAACGCAGTAA 199
gnl|ti|2141657616_[782_945]      GCCCAGAAAAGAACC---------------------------------- 164
gnl|ti|2141510109_[228_9]        GCCAGAAAAAGAACCAAGAAATCGAAGAATTAAAGAATCAAACGCAG-AA 198
                                 ***   *********
```

(A)

S. salmonicida - Rpl30



(B)

```
SS5-377_16979(Rpl30) GUGAGUUUUAACAACUGAGACCGGUUUGGUCUCAGAAAUCUUAACUUUUACUAACAAACUAG 67 nt
SS5-377_16134        GUAUGUUUUAA------------------CAAUUAAAAAAUAACUUAUACUAACAAACUAG 43 nt
SS5-377_18398        GUAUGUUUUAA------------------CUCAAUAAAUACAACUUUUACUAACAAACUAG 43 nt
SS5-377_17358        GUAUGUCUAAA------------------CUUUUUUAAUGUAACUUAUACUAACAAACUAG 43 nt
                     **  ** * **                  *      *    ***** **************
```

Figure A.3.3. Base pairing potential the in *S. salmonicida Rpl30* intron.
(A) MFOLD secondary structural prediction for the *Rpl30* intron from *S. salmonicida* (Xu *et al.* 2014) is shown as described in Figure 3.2, with the predicted single stranded length indicated. **(B)** ClustalW2 alignment of *S. salmonicida* introns (modified from Xu *et al.* 2014), with stem loop forming nucleotide from the *Rpl30* intron in red and total intron lengths indicated in nucleotides (nt).

Table A.3.2. Structural potential of *cis*-spliceosomal introns in *Trichomonas vaginalis*. Spliceosomal introns sequences from *T. vaginalis* genes are shown with red text indicating intron regions predicted to form secondary structure by MFOLD software (Zuker 2003). Total intron lengths and the predicted single stranded (SS) length of folded pre-mRNA introns are indicated.

Gene ID	Genomic Location	Gene Description	Intron Sequence	Intron Length	Intron SS Length
TVAG_014960	DS113774: 39,976 - 41,345 (+)	TATA binding protein associated factor, putative	gtacgtatttctttggtttctggcttgttttattttaaaataaaccaggaaccaaatttttttcagaaactaacacacag	81	36
TVAG_020880	DS113200: 11,125 - 13,510 (-)	AGC family protein kinase	gtatgtatttttattttttcatcgattacaaaatttttctgaaaatataagatttgagatatatatttcgaataccgaaaatttgatgatttttttatgaaatttttggtatttttttatatctcaaaaattatttcagaaaaaaagaaaggtggttgatgaaataattttagatcattactaacacacag	196	37
TVAG_043580	DS113505: 33,566 - 34,321 (-)	maintenance of ploidy protein mob2, putative	gtacagttttaattctaacaaacag	25	25
TVAG_053820	DS113785: 33,670 - 35,267 (+)	CAMK family protein kinase	gtatgtttttaatgaaatttttaataacaaaaaaaaattttcaaaaattccaaatttttttgttattaaaatttcatatttttttttatttttaaaaatactaacacacag	110	24
TVAG_056030	DS113419: 42,709 - 43,113 (+)	conserved hypothetical protein	gttctatttaatttctaacaaacag	25	25
TVAG_065500	DS115094: 3,745 - 5,333 (-)	CAMK family protein kinase	gtatgtatttttttggtaaacctcaatttttttcaaatgacttttcttaacctttcgaaatatactttcaaaagtgttagaaatgtcatattgtaaataagaggtgccatttttgttaacatttactaacacacag	134	39
TVAG_085780	DS114439: 10,258 - 11,686 (+)	conserved hypothetical protein	gtatgtacttttgagctggtaacattatttaccagctccttttacttagaaaatactaacacacag	67	41
TVAG_087980	DS113624: 20,938 - 22,047 (-)	STE family protein kinase	gtatgtatactttgatgtcaatttttatttttttaattgacatcattttctttttttatttttagatactaacacacag	78	47
TVAG_089630	DS114221: 15,715 - 17,109 (-)	AGC family protein kinase	gttctttttttatttctaacaaacag	25	25
TVAG_110020	DS113198: 45,552 - 46,947 (+)	TATA binding protein associated factor, putative	gtacgtatttttaggcaaagtatttctttttttaattttttaaaaaattgaaatattttgccaaaatttaaacaaaatatactaacacacag	91	43
TVAG_110580	DS113198: 188,777 - 191,296 (+)	centaurin gamma, putative	gtatgtatattttctggcgtaaaaagaaagatataaattaactttttacccctcatttgctttttaaatgaaagggaatcgtatttttatgctttctttttgcgctatttttttacatttactaacacacag	127	38
TVAG_125100	DS113398: 99,086 - 100,336 (+)	CMGC family protein kinase	gtatgttttcgagttcttcgtacataaagaaacagacattttagtctgttttctttttggaaaaaaaaacgaagaatttttttaaaaaaaatttactaacacacag	105	36
TVAG_126240	DS113357: 56,024 - 57,470 (-)	CAMK family protein kinase	gtatgtttctttttattggttatcactattaatagaaaatctcaaaatttttctattaatgtggaaccaaaaacaattttttataagttactaacacacag	99	44
TVAG_130170	DS113203: 191,775 - 193,935 (+)	conserved hypothetical protein	gtatgtatttatttctaaccaacag	25	25
TVAG_134480	DS114086: 5,751 - 6,702 (+)	ribosomal protein S6 kinase, 90kD, polypeptide, putative	gttctttttttatttctaacaaacag	25	25
TVAG_147850	DS114056: 23,260 - 24,449 (-)	CAMK family protein kinase	gtatgtatttttaatttttggagtgttatattgatcattccaatgtttatcatataactaacacacag	68	42
TVAG_148640	DS113755: 29,291 - 30,330 (+)	CAMK family protein kinase	gtatgtactatttttattttttgcctaatttacaaatattttttatttgataaatttcgcaaaaattttttatttttttcaaaatactaacacacag	93	44
TVAG_176980	DS113680: 30,321 - 31,501 (+)	CMGC family protein kinase	gtacgtatttcatctatgcataatttatgcatagatattttttcaaaactaacacacag	59	34
TVAG_198230	DS113190: 59,033 - 59,932 (+)	conserved hypothetical protein	gtatgtattttattttcataagttgccaatttttctttttagtataaccggcaacttatgagttttatacattactaacacacag	84	37
TVAG_217460	DS113550: 71,345 - 72,565 (-)	ankyrin repeat domain protein, putative	gtatgtacctatttataagcaaattggcgtaacaatacgctatttgcattatttttcaatgctaacacacag	72	34
TVAG_225200	DS113224: 7,991 - 9,110 (-)	nuclear lim interactor-interacting factor,	gtatgtatatttttgtttcatatttcctatttggaaatttgaaaacatttttggaaaaaatttactaacacacag	76	42

		putative			
TVAG_242770	DS113657: 12,422 - 13,817 (+)	conserved hypothetical protein	gtactgctttatttctaaccaacag	25	25
TVAG_306990	DS114021: 8,464 - 9,903 (-)	CMGC family protein kinase	gtatgtttcttacaagcatgtgttcttcgcaatatcg aaattttgcgaagaaaacgtgtttgaaaaaaaaaa ttaaattttactaacacacag	93	43
TVAG_324910	DS113569: 18,626 - 19,252 (-)	RNA recognition motif (rrm) domain containing protein, putative	gtacattttttatttctaacaaacag	25	25
TVAG_350500	DS113477: 60,512 - 62,110 (-)	CAMK family protein kinase	gtatgtacctatttaccttcgataatttcatttaattac gggctaatttagctttttaccgtaattttatgaaatttt atcgaaatttttgaaaaataatttttactaacacacag	114	33
TVAG_360840	DS113782: 26,243 - 27,435 (-)	ankyrin repeat-cotaining protein, putative	gtatgcattcgaaattcgacttttcagtcgaatgttt tttgaatactaacacacag	56	25
TVAG_383350	DS113985: 30,460 - 31,220 (+)	RAB-2,4,14, putative	gtataatttaatttctaacaaacag	25	25
TVAG_388620	DS113269: 100,189 - 101,767 (+)	poly(A) polymerase gamma, putative	gtatgtacaatttttttgattaatattattgtttcttgca ttttcatgcttgaaacaattatattaattgttttattcat tcactaacacacag	94	42
TVAG_390460	DS113480: 48,186 - 49,227 (-)	nuclear lim interactor-interacting factor, putative	gtatgtataatattacaaatatcttcaatatattttga agatatttaaattttcaaatactaacacacag	70	40
TVAG_413420	DS113675: 23,577 - 24,757 (-)	CMGC family protein kinase	gtatgtttcattacacgtctagatttttacgtctagac tgtaaatttttttgaattactaacacacag	68	37

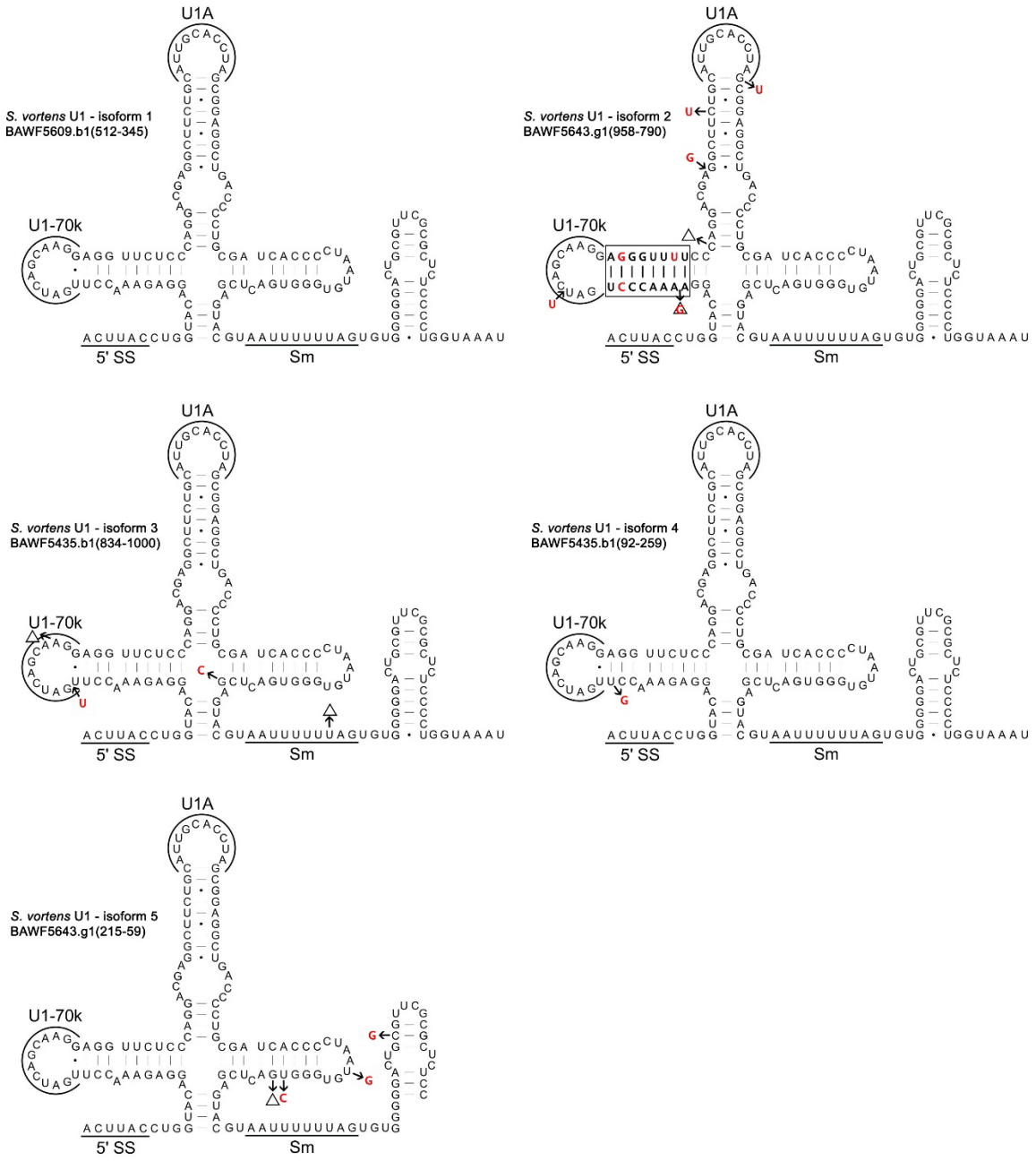


Figure A.3.4. *S. vortens* U1 snRNA isoforms.
Predicted secondary structures (MFOLD) for *S. vortens* U1 snRNA isoform 1 (also shown in Figure 4.3A) are shown with arrows and/or red text indicating nucleotide changes observed in the various U1 snRNA gene copies. The boxed region in U1 'isoform 2' contains numerous nucleotide substitutions that result in altered base pairing of the SL I. The genomic sequence read for 'isoform 5′ is truncated at the RNA 3′ end.

Table A.3.3. Evolutionary conservation of *Rps4* gene introns in eukaryotes

	Intron Present?	Intron Length (nt)
Archaeaplastida		
Green Plants (green algae, prasinophytes and land plants)		
Land plants		
Eudicots		
Arabidopsis thaliana	1	307
Medicago truncatula	1	1398
Ricinus communis	1	682
Fragaria vesca	1	77
Vitis vinifera	1	937
Monocots		
Setaria italica	1	845
Brachypodium distachyon	1	740
Oryza sativa	1	670
Club moss		
Selaginella moellendorffii	1	49
Moss		
Physcomitrella patens	1	220
Green algae		
Ostreococcus lucimarinus	0	
Chlamydomonas reinhardtii	0	
Volvox carteri f. Nagariensis	0	
Micromonas pusilla	0	
Rhodophyta (red algae)		
Porphyridium purpureum	0	
Glaucophytes (Cyanophora)		
Cyanophora paradoxa	0	
Total	10	

	Intron Present?	Intron Length (nt)
Unikonts		
Opisthokonts		
Animals (Metazoa)		
Chordates		
Homo sapiens	1	958
Gallus gallus	0	
Oreochromis niloticus	1	563
Xenopus (Silurana) tropicalis	1	675
Ciona intestinalis	1	223
Arthropods		
Culex quinquefasciatus	0	
Drosophila melanogaster	1	95
Bombus impatiens	1	455
Tribolium castaneum	1	338
Bombyx mori	1	316
Nematodes		
Brugia malayi	0	

Trichinella spiralis	0	
Loa loa	0	
Caenorhabditis elegans	0	
Cnidarians		
Hydra magnipapillata	1	133
Sponge		
Amphimedon queenslandica	1	80
Mollusc		
Aplysia californica	1	3183
Choanoflagellates		
Salpingoeca sp. ATCC 50818	0	
Monosiga brevicollis	0	
Filastera		
Capsaspora owczarzaki ATCC 30864	1	430
Ichthyosporea		
Fungi		
Ascomycetes		
Candida albicans	0	
Ashbya gossypii	0	
Aspergillus fumigatus	0	
Coccidioides posadasii	0	
Neurospora crassa	0	
Schizosaccharomyces pombe	0	
Pyrenophora tritici-repentis	0	
Botryotinia fuckeliana	0	
Nectria haematococca	0	
Magnaporthe oryzae	0	
Verticillium albo-atrum VaMs.102	0	
Zymoseptoria tritici	0	
Basidiomycetes		
Coprinopsis cinerea okayama7#130	1	162
Schizophyllum commune H4-8	1	143
Cryptococcus neoformans var. neoformans B-3501A	1	151
Postia placenta Mad-698-R	1	210
Puccinia graminis f. sp. tritici CRL 75-36-700-3	1	147
Microsporidians		
Encephalitozoon hellem	0	
Encephalitozoon intestinalis	0	
Amoebozoa		
Physarum polycephalum	1	74
Entamoeba histolytica	0	
Entamoeba dispar	0	
Entamoeba invadens	0	
Acanthamoeba castellanii	0	
Dictyostelium discoideum	1	430
Dictyostelium fasciculatum	0	
Total	19	

	Intron Present?	Intron Length (nt)
Excavates		
Malawimonads		
Euglenozoa		
Kinetoplastids		
Trypanosoma brucei	0	
Trypanosoma cruzi	0	
Leishmania braziliensis	0	
Leishmania major	0	
Leishmania infantum	0	
Heterolobosea		
Naegleria gruberi	1	144
Jakobida		
Parabasalids		
Trichomonas vaginalis	0	
Fornicata		
Giardia lamblia	0	
Preaxostyla		
Total	1	

	Intron Present?	Intron Length (nt)
Chromalveolates		
Cryptomonas paramecium	0	
Guillardia theta	0	
Hemiselmis andersenii	0	
Rhizaria		
Cercozoa		
Bigelowiella natans	0	
Foraminifera		
Radiolaria		
Alveolates		
Perkinsus marinus	0	
Apicomplexans		
Theileria parva strain Muguga	1	103
Cryptosporidium muris	1	62
Plasmodium knowlesi	1	675
Babesia equi	1	104
Plasmodium falciparum	1	715
Ciliates		
Ichthyophthirius multifiliis	1	61
Stramenopiles		
Diatoms		
Thalassiosira pseudonana	0	
Phaeodactylum tricornutum	0	
Oomycetes		
Phytophthora infestans	0	
Total	6	

Table A.3.4. Evolutionary conservation of *Rps24* gene introns in eukaryotes

	Intron Present?	Intron Length (nt)
Archaeaplastida		
Green Plants (green algae, prasinophytes and land plants)		
Land Plants		
Eudicot		
Glycine max	1	432
Arabidopsis thaliana	1	299
Ricinus communis	1	713
Solanum lycopersicum	1	117
Cucumis sativus	1	321
Fragaria vesca	1	112
Vitis vinifera	1	667
Monocot		
Sorghum bicolor	0	
Setaria italica	1	92
Club-mosses		
Selaginella moellendorffii	1	73
Mosses		
Physcomitrella patens	1	137
Green Algae		
Chlamydomonas reinhardtii	1	79
Ostreococcus tauri	0	
Ostreococcus lucimarinus	1	168
Volvox carteri	1	71
Micromonas pusilla	1	168
		0
Rhodophyta (red algae)		
Porphyridium purpureum	0	
Glaucophytes (Cyanophora)		
Cyanophora paradoxa	1	79
Total	15	

	Intron Present?	Intron Length (nt)
Unikonts		
Opisthokonts		
Animals (Metazoa)		
Chordates		
Homo sapiens	1	93
Danio rerio	1	106
Gallus gallus	1	327
Anolis carolinensis	1	951
Arthropods		
Drosophila melanogaster	0	
Aedes aegypti	0	
Culex quinquefasciatus	0	
Apis mellifera	1	167
Nasonia vitripennis	1	75
Pediculus humanus corporis	1	83

Nematodes		
Caenorhabditis elegans	1	116
Brugia malayi	1	392
Cnidarians		
Hydra magnipapillata	0	
Nematostella vectensis	1	644
Echinodems		
Strongylocentrotus purpuratus	1	457
Molluscs		
Aplysia californica	1	2386
Placozoans		
Trichoplax adhaerens	1	600
Choanoflagellates		
Monosiga brevicollis	0	
Salpingoeca	1	279
Filastera		
Ichthyosporea		
Capsaspora owczarzaki	1	243
Fungi		
Ascomycetes		
Saccharomyces cerevisiae	0	
Naumovozyma dairenensis	0	
Ashbya gossypii	0	
Tetrapisispora phaffii	0	
Vanderwaltozyma polyspora	0	
Candida albicans	0	
Yarrowia lipolytica	0	
Basidiomycetes		
Schizophyllum commune	0	
Cryptococcus neoformans	0	
Postia placenta	0	
Microsporidians		
Encephalitozoon hellem	0	
Encephalitozoon intestinalis	0	
Nucleariidae		
Amoebozoa		
Entamoeba histolytica	0	
Entamoeba dispar	0	
Entamoeba invadens	0	
Dictyostelium discoideum	0	
Dictyostelium fasciculatum	0	
Dictyostelium purpureum	0	
Acanthamoeba castellanii	1	117
Physarum polycephalum	1	?
Total	17	

	Intron Present?	Intron Length (nt)
Chromalveolates		
Alveolata		
Apicomplexan		
Plasmodium falciparum	0	
Plasmodium vivax	0	
Toxoplasma gondii	0	
Cryptosporidium parvum	0	
Babesia bovis	0	
Theileria parva	0	
Neospora caninum	0	
Cilliates		
Paramecium tetraurelia	1	23
Ichthyophthirius multifiliis	0	
Perkinsus		
Perkinsus marinus	1	48
Stramenopiles		
Thalassiosira pseudonana	0	
Phaeodactylum tricornutum	0	
Phytophthora infestans	1	79
Rhizaria		
Cercozoa		
Bigelowiella natans	1	135
Foraminifera		
Radiolaria		
Hacrobia		
Cryptomonads		
Cryptomonas paramecium	0	
Guillardia theta	0	
Hemiselmis andersenii	0	
Total	4	

	Intron Present?	Intron Length (nt)
Excavates		
Malawimonads		
Euglenozoa		
Kinetoplastids		
Trypanosoma brucei	0	
Trypanosoma cruzi	0	
Leishmania major	0	
Leishmania donovani	0	

Heterolobosea		
Naegleria gruberi	0	
Jakobida		
Parabasalids		
Trichomonas vaginalis	0	
Fornicata		
Diplomonads		
Giardia lamblia	0	
Spironucleus vortens	1	41
Preaxostyla		
Total	1	

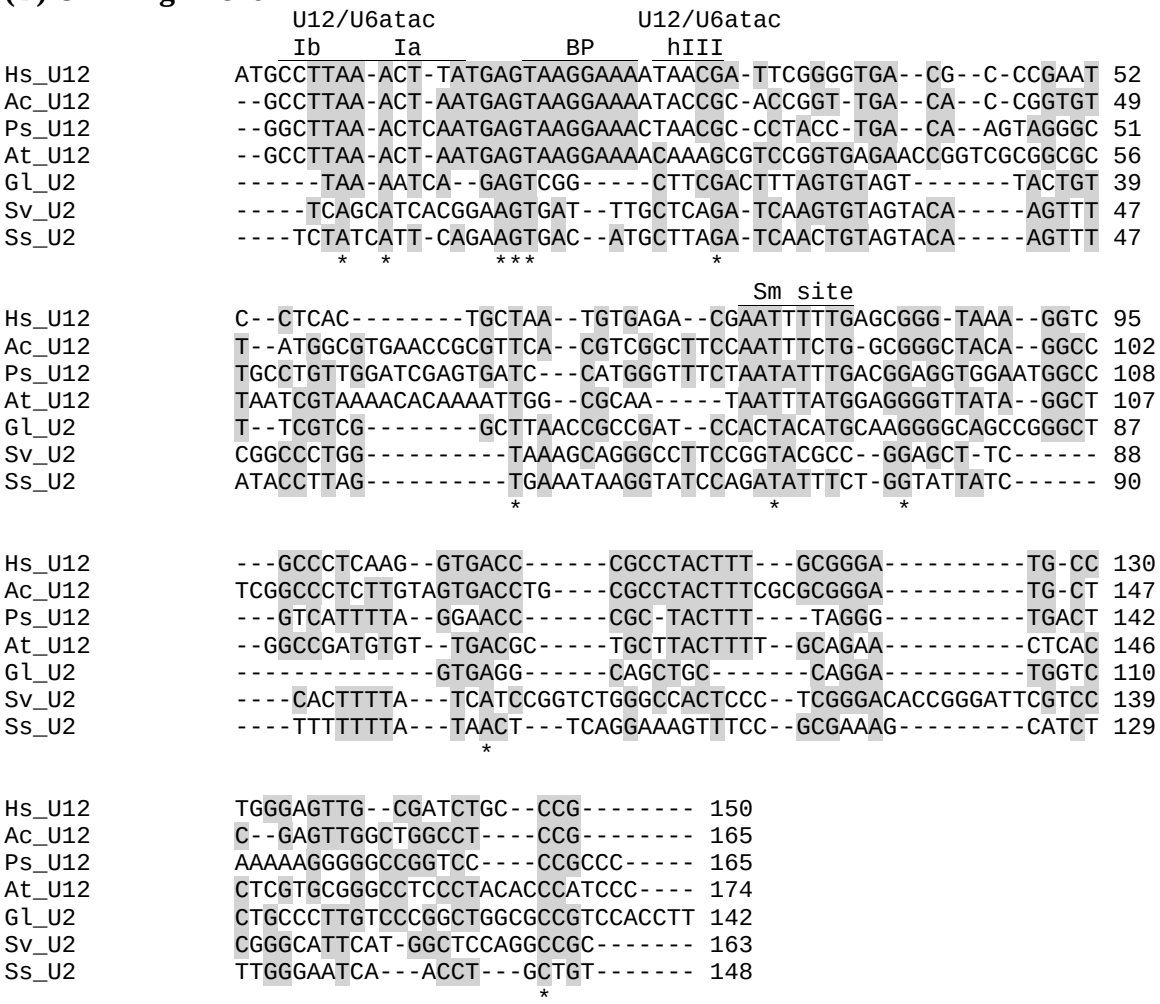
Figure A.3.5. Primary sequence comparison of *Spironucleus* U2 snRNA candidates with U2 and U12 snRNAs from representative eukaryotes.

Primary sequences of *S. vortens* and *S. salmonicida* U2 snRNAs were aligned with (**A**) U2 or (**B**) U12 snRNAs from representative eukaryotes using ClustalW2 software (Larkin *et al.* 2007). Regions of U2/12 snRNA predicted for form intermolecular base pairing with U6/U6atac or the intron branch point (BP) are indicated above the alignments. Nucleotides conserved in at least four eukaryotes are highlighted in grey. Alignments were constructed using: *S. vortens* (Sv) U2 snRNA (NCBI trace archive ti|2141663608: nucleotide positions 84-246); *S. salmonicida* (Ss) U2 snRNA (GenBank AUWU01000434:68649-68502); *Giardia lamblia* (Gl) U2 [GenBank Accession JX416862];, *Acanthamoeba castellanii* (Ac) U2 [GenBank CW933695:787-579], U12 [CW917526:369-205]; *Phytophthora spp*. (Pr) U2 [*Phythophthora ramorum* genome release V1.0 scaffold_1672:234-416], U12 (Ps) [AAQY02000248:644532-644696]; *Arabidopsis thaliana* (At) U2 [X06478:200-359], U12 [CP002684:22603122-22603295]; *Homo sapiens* (Hs) U2 [NR_002716:1-187], U12 [L43846:331-480].

(A) U2 Alignment

```
                                          U2/U6                  U2/U6
                        U2/U6 hII          Ib  Ia        BP      hIII
At_U2          ---ATACCTTTCTCGGCCTTT--TGGCTAAGA-TCAAGTGTAGTATCTGTTCTTATCAGT 54
Hs_U2          ---AT-CGCTTCTCGGCCTTT--TGGCTAAGA-TCAAGTGTAGTATCTGTTCTTATCAGT 53
Ac_U2          -ACAT---CTTCTCGGCCCAAG-TGGCTAAGA-TCATGTGAAGTATCTGTTCTTATCAGC 54
Pr_U2          ---AC---CTTCTCGGCCTTT--TGGCTAAGA-TCAAGTGTAGTATCTGTTCTAATCAGT 51
Gl_U2          TAAAA-----TCAGAGTC------GGCTTCGACTTTAGTGTAGTTACTGTT-TCGTCGGC 48
Sv_U2          -TCAG---CATCACGGAAGTGATTTGCTCAGA-TCAAGTGTAGTACAAGTTTCGGCCCTG 55
Ss_U2          TCTAT---CATT-CAGAAGTGACATGCTTAGA-TCAACTGTAGTACAAGTTTATACCTTA 55
                  *     *    *         ***  ** *    ** ***    ***     *

                                                                 Sm site
At_U2          TTAATATCTGAT----ATGTGGGCCATCGGCCCACACGATATTAACTCTATTTTTTAAGG 110
Hs_U2          TTAATATCTGAT----ACGTCCTCTATCCGAGGACAATATATTAAATGGATTTTTGGAGC 109
Ac_U2          TTAATCTCTGGTAGTGAGGCCTCCTGTGCCTCACCTCAAGGTTAGACTTATTTTTCTTGT 114
Pr_U2          GTGAAAACTGGTTCCGACGTTTTTCGTTGGTCTTTT-----TCACATTCATTTTT----- 101
Gl_U2          TTAACCGCCGAT-----------CCAC--------------TACATGCA---------A 73
Sv_U2          GTAAAGCAGGGCCTTCCGGTACGCC-GGAGCT-TC-------CACTTTTATCATCCGGTC 106
Ss_U2          GTGAAATAAGGTATCCAGATATTTCTGGTATTATC-------TTTTTTTATAACT---TC 105
                * *    *                                      *

At_U2          GAGAAAGCCCGT-TAAGAT--------AGCT-TGCT------------------AT--C 139
Hs_U2          AGGGAGATGGAA-TAGGA----------GCT-TGCTCCGTCCACTCCACGCATCGAC--C 155
Ac_U2          TGGGC-TCCTGG-CACCATGCCCTTCCAGCTATGCTGTGGGCAGTCCAGAGAGCAGTGAT 172
Pr_U2          -GGGCATCCCGA-TGTCGCGC------AGCT-TGCTGTGCGAGGTC---GGGGCGGTTTC 149
Gl_U2          GGGGCAGCCGGGCTGTGAGGC------AGCT--GCC------------AGGATGGT--C 110
Sv_U2          TGGGCCACTCCCTCGGGACACCG----GGATTCGTC----------------------C 139
Ss_U2          AGGAAAGTTTCCGCGAAAG------------CATC----------------------T 129
                 *

At_U2          TGG---------GCTTT-------CGCGA-GTCGCCCA---- 160
Hs_U2          TGGTATTGCAGTACCTC-------CAGGA-ACGGTGCACC-- 187
Ac_U2          CAG---CTTTGTACTGCACCACCCTGCAAAGTTCTTCAAAT- 210
Pr_U2          CGGGGGCTTT-CACCTCTCC--CCCGCGAGGC----CAAC-- 182
Gl_U2          CTG----CCCTTGTCCCGGC------TGGCGCCGTCCACCTT 142
Sv_U2          CGGGCATTCATGGCTCC-----------AGGCCGC------- 163
Ss_U2          TTGGGAATCA--ACCT-------------GCTGT------- 148
                 *
```

(B) U12 Alignment



Appendix 4 – Oligonucleotide Primers Used in this Study

Table A.4.1. Oligonucleotides Used in Chapter 2: Numerous Fragmented Spliceosomal Introns, AT–AC Splicing, and an Unusual Dynein Gene Expression Pathway in *Giardia lamblia*

Name	Sequence (5′ to 3′)	Description
oAH1	GAT CCT CTT CAT CCC CAA GCG C	Forward primer for RT-PCR used in combination with oAH3 to confirm splicing at the Hsp90 intron boundary. Primer sequence corresponds to position +954 to +976 of Hsp90 exon 1.
oAH2	GAT GAG GTG GAT TTT TTG CCC GG	Reverse primer used with oAH1 to PCR-amplify the 5′ exon-intron boundary of Hsp90. Complementary to the intronic region +190 to +213 nt. downstream of the 5′ splice junction.
oAH3	GAT TCT TCT GGA GCA TCT CAC GG	Forward primer used with oAH4 for genomic amplification of the 3′ intron-exon boundary of Hsp90. Primer sequence is the region -116 to -93 nt. upstream of the 3′ splice site.
oAH4	CCA ACG AAC CAG TGC CAA GCG	Reverse primer used to generate the cDNA to confirm the single Hsp90 splice boundary. Sequence is complementary to positions +94 to +114 of the second Hsp90 exon.
oAH5	CGC TCA ACA AGG AGC TTC TCA GC	Forward primer for RT-PCR in combination with oAH8 to verify splicing of DHC beta exons 1 and 2. Primer sequence anneals to +7073 to +7095 region of DHC beta exon 1.
oAH6	AAG CAG AGT ATA CCG GTA CCT CG	Reverse primer used with oAH5 to amplify the 3′ region of DHC beta exon 1 until an annotated downstream ORF. Primer is complementary to a region +223 to +245 nt. downstream of the DHC beta exon 1 termination codon.

Name	Sequence (5′ to 3′)	Description
oAH7	GGC CGC CAG AGC GAA TGT TGG	Forward primer used with oAH8 for genomic amplification of the 5′ terminal region of DHC beta exon 2. Primer anneals to region -263 to -283 upstream of the exon 2 start codon.
oAH8	GGA GGG ACT TCT TGA TGA GAG CC	Reverse primer used to generate the cDNA to confirm splicing of DHC beta exon 1 and 2. Primer is complementary to region +177 to +199 of the DHC beta exon 2.
oAH9	GGC CGC GAT GCA GCT CGA AGC	Forward primer used with oAH12 for RT-PCR verification of splicing of DHC beta exon 2 and 3. Primer sequence spans +3924 to +3944 coding sequence of DHC beta exon 2.
oAH10	GGG CCC CTC TCT TCC TCT CTT CC	Reverse primer used with oAH9 for genomic amplification of the 5′ splice donor of DHC beta exon 2. Primer is complementary to the region +197 to +219 nt. downstream of the 5′ splice site
oAH11	GCC AAG CCC ATC CCT TGG TCC	Forward primer used with oAH12 for genomic amplification of the 3′ splice acceptor of the second DHC beta intron. Primer sequence corresponds to -313 to -333 upstream from the 3′ splice site on DHC beta exon 3.
oAH12	GCC TTC GTG GCA TTC GCC TGC CG	Reverse primer used for cDNA synthesis for confirming splicing of DHC beta exon 2 and 3. Sequence is complementary to the region +27 to +49 downstream of the 3′ splice junction on DHC beta exon 3.
oAH13	CGT TTG AAA TGT GCT CCA AGG G	Forward primer used with oAH16 for RT-PCR confirmation of splicing of DHC beta exons 3 and 4. Sequence is coding sequence -82 to -103 nt. upstream of the 5′ donor splice site of the second DHC beta intron.
oAH14	GTT GGC AGA TAG ATT GGT AGG C	Reverse primer used with oAH13 for genomic amplification of the region flanking the second DHC beta 5′ donor. Primer is complementary to +236 to +257 downstream of the 5′ splice site.
oAH15	CTA ATC GAA CCG CGA CGC TTG C	Forward primer used with oAH16 for genomic amplification of the 3′ splice acceptor for DHC beta intron 2. Sequence anneals to -194 to -215 nt. upstream of the intron 2 3′ splice site.
oAH16	GGT CGT CCC AGG GGA TTT TGG	Reverse primer for generating the cDNA for verifying splicing of DHC beta exon 3 and 4. Primer is complementary to +81 to +101 downstream of the 3′ splice site of the

Name	Sequence (5′ to 3′)	Description
		second DHC beta intron.
oAH32	GTT ATT ACC CTC ATC CCC TCT TGC	Forward primer used with oAH35 for RT-PCR verification of splicing of DHC gamma exon 1 and 2. Primer anneals -88 to -111 nt. upstream of the 5′ splice donor site.
oAH33	GAG CTG ACC TGG ACA TAA AGA GC	Reverse primer used with oAH32 for genomic amplification of the 5′ DHC gamma splice boundary. Sequence is complementary to +125 to +147 downstream of the 5′ splice site.
oAH34	GAC GTG CAA AGG CAA CTA CAG G	Forward primer used with oAH35 for genomic amplification of the DHC gamma 3′ acceptor splice site. Sequence corresponds to -251 to -272 upstream of the 3′ splice site.
oAH35	GT GCT TGG AAA GCT GCT CCT C	Reverse primer to generate cDNA to verify splicing of DHC gamma exon 1 and 2. Primer sequence is complementary to +106 to +127 nt. downstream of the 3′ splice acceptor.

Name **Sequence (5′ to 3′)** **Description**

Table A.4.2. Oligonucleotides Used in Chapter 3: Evolutionarily divergent spliceosomal snRNAs and a conserved non-coding RNA processing motif in *Giardia lamblia*.

Name	Sequence (5′ to 3′)	Description
p-94	AAT AAA GCG GCC GCG GAT CCA ATT TTT TTT TTT TTT V	Reverse primer for 3′ RACE of polyA-tailed *Giardia* total RNAs. 'V' indicates any nucleotide except for 'T'.
RML-5′ RACE Linker	rGrCrU rGrArU rGrGrC rGrArU rGrArA rUrGrA rArCrA rCrUrG rCrGrU rUrUrG rCrUrG rGrCrU rUrUrG rArUrG rArArA	RNA oligo for addition to 5′ ends of RNA samples for 5′ RACE analysis.
5′ RACE Outer Primer	GCT GAT GGC GAT GAA TGA ACA CTG	Forward primer for PCR step during 5′ RACE. Primer anneals to position +1 to +24 of RML-5′ RACE linker.
5′ RACE Inner Primer	CGC GGA TCC GAA CAC TGC GTT TGC TGG CTT TGA TG	Alternate forward primer for PCR step during 5′ RACE. Underlined region anneals to position + 17 to + 42 of RML-5′ RACE linker.
oAH1	GAT CCT CTT CAT CCC CAA GCG C	Forward primer for RT-PCR used in combination with oAH2 to detect Hsp90 and Replication Factor C dicistronic transcript. Primer sequence corresponds to position +954 to +976 of Hsp90 exon 1.
oAH2	GAT GAG GTG GAT TTT TTG CCC GG	Reverse primer for RT-PCR used in combination with oAH2 to detect Hsp90 and Replication Factor C dicistronic transcript. Complementary to the coding sequence +84 to +106 nt. downstream of Replication Factor C start codon.
oAH9	GGC CGC GAT GCA GCT CGA AGC	Forward primer used for RT-PCR detection of the expression of DHC beta exon 2 and downstream flanking region precursor transcript. Primer sequence spans +3924 to +3944 coding sequence of DHC beta exon 2.
oAH10	GGG CCC CTC TCT TCC TCT CTT CC	Reverse primer used with oAH9 for RT-PCR detection of the expression of DHC beta exon 2 and downstream flanking region precursor transcript. Primer is complementary to the region +197 to +219 nt. downstream of the 5′ splice site
oAH13	CGT TTG AAA TGT GCT CCA AGG G	Forward primer used with oAH14 for RT-PCR detection of DHC beta exons 3 and downstream flanking region precursor transcript. Primer anneals in exonic region -82 to -103 nt. upstream of the *trans*-intron 5′ splice site.

Name	Sequence (5′ to 3′)	Description
oAH14	GTT GGC AGA TAG ATT GGT AGG C	Reverse primer used with oAH13 for RT-PCR detection of DHC beta exons 3 and downstream flanking region precursor transcript. Primer is complementary to +236 to +257 nt. downstream of the *trans*-intron 5′ splice site.
oAH32	GTT ATT ACC CTC ATC CCC TCT TGC	Forward primer used with oAH33 for RT-PCR detection of DHC gamma exon 1 and Hypothetical Protein dicistronic transcript. Primer anneals within exonic region -88 to -111 nt. upstream of the *trans*-intron 5′ splice site.
oAH33	GAG CTG ACC TGG ACA TAA AGA GC	Reverse primer used with oAH32 for RT-PCR detection of DHC gamma exon 1 and Hypothetical Protein dicistronic transcript. Sequence is complementary to +24 to +46 nt. dow`nstream of the Hypothetical Protein start codon.
oAH60	GGC CGG CAT AAC CGA AAT CG	Reverse primer for primer extension and 5′ RACE of *G. lamblia* **GlsR26**. Primer anneals -9 to -28 nt. upstream of 3′ processing motif sequence.
oAH62	GAG ATC ACA AAT GTG CTC CGG CCA GG	Reverse primer for primer extension and 5′ RACE of *G. lamblia* **GlsR27**. Primer anneals from -6 to -31 nt. upstream of 3′ processing motif sequence.
oAH70	GGA TGG GAT CCT TCC CCT TGC TTC TGG	Reverse primer for primer extension and 5′ RACE of *G. lamblia* **GlsR28**. Primer anneals from -8 to -34 nt. upstream of 3′ processing motif sequence.
oAH72	GGT GCA AGC ACG CGC CAA CGG GC	Reverse primer for primer extension and 5′ RACE of *G. lamblia* **U6 snRNA candidate**. Primer anneals from -6 to -28 nt. upstream of 3′ processing motif sequence.
oAH73	CGG GCC CGG ATT GAG GAT GGA CG	Reverse primer for RT-PCR detection of GlsR17 + GlsR18 polycistronic transcript precursor. Anneals within GlsR18 mature sequence, -7 to -29 nt. upstream of the GlsR18 3′ processing motif.
oAH74	ATA ATG CGC TTC TTT GAG CCG CGG G	Forward primer to be used with oAH73 for RT-PCR detection of GlsR17 + GlsR18 polycistronic transcript precursor. Anneals within the mature GlsR17 sequence, -101 to -125 nt. upstream of the GlsR17 3′ processing motif.
oAH75	GAG GCT GCT AAA ACA CAG GGC	Forward primer to be used with oAH60 for RT-PCR detection of GlsR25 + GlsR26 polycistronic transcript precursor. Anneals within GlsR25 mature sequence, -94 to -

Name	Sequence (5′ to 3′)	Description
		115 nt. upstream of the GlsR25 3′ processing motif.
oAH76	ATA AGC TGG AAT TCC ACG TCT TCC TCG	Reverse primer to be used with oAH77 for RT-PCR detection of Candidate-23 + DNA polymerase delta catalytic subunit polycistronic transcript. Primer is antisense to region +54 to +80 downstream of predicted first "AUG" codon of DNA pol mRNA (protein CDS in region is conserved with *Entamoeba* and *Culex Spp.*)
oAH77	GGC ATG GAG AAG AGC AGA CTT GAG G	Forward primer to be used with oAH76 for RT-PCR detection of Candidate-23 + DNA polymerase delta subunit polycistronic transcript. Anneals within Candidate-23 mature sequence to region -7 to -31 upstream of 3′ processing motif.
oAH78	GGA GCT CGA CCA TTT TCA CAT CCC	Reverse primer for use with oAH79 for RT-PCR mediated verification of Candidate-5 + Ser/Thr kinase CDS polycistronic transcript. Is complementary to +146 to +169 downstream of predicted first AUG codon of Ser/Thr mRNA. Anneals in portion of CDS which encodes an ankyrin domain.
oAH79	CCT TGC CCA GTC TGC CTC CAT AC	Forward primer to be used with oAH78 for RT-PCR amplification of Candidate-5 + Ser/Thr kinase polycistronic transcript. Anneals within Candidate-5 sequence -9 to -31 upstream of its 3′ processing motif.
oAH95	GGG TAC GGT AGC AGG TCT GAG AGC	Reverse primer for primer extension and 5′ RACE of *G. lamblia* U1 snRNA candidate. Anneals to region -12 to -35 upstream of the 3′ motif sequence.
oAH102	GTT ATG TTT GTA TGC TGT ATG TGT GC	Forward primer for 3′ RACE of *Giardia* HSP90 *trans* intron 5′ half. Primer anneals to region +5 to +31 downstream of exon 1-intron splice site and upstream of the 3′ motif sequence.
oAH103	GTA TGT TAC TGG GTG AAA CGC TAC	Forward primer for 3′ RACE of *Giardia* DHC Beta *trans* intron #1 - 5′ half. Primer anneals to region +1 to +24 downstream of exon 2 - intron 1 splice site and upstream of the 3′ motif sequence.
oAH104	GTA ATC TGT GTA GTC GCA GTA TGC C	Forward primer for 3′ RACE of *Giardia* DHC Beta *trans* intron #2 - 5′ half. Primer anneals to region +9 to +33 downstream of exon 3 - intron 2 splice site and upstream of the 3′ motif sequence.
oAH105	CAC AGG TGG TTT GGT GTG	Forward primer for 3′ RACE of *Giardia*

Name	Sequence (5′ to 3′)	Description
	TAT GC	DHC Gamma *trans* intron 5′ half. Primer anneals to region +8 to +30 downstream of exon 1 - intron splice site and upstream of the 3′ motif sequence.
oAH112	CAC TCA AGT ATG TTC TTG CG	Forward primer for 3′ RACE of Protein 21.1 CDS (GL50803_25296) to detect processing near its 3′ motif sequence. Primer anneals within protein coding region -13 to -32 nt. upstream of the predicted 3′ motif sequence.
oAH113	AGC TTT CTA AAA CCA CTC CC	Forward primer for 3′ RACE of Hypothetical Protein CDS (GL50803_7350) for detection of processing near its 3′ motif sequence. Primer anneals within protein coding region -20 to -39 nt. upstream of the predicted 3′ motif sequence.
oAH114	CTG TAT TGT ATG CTT CAA TGG	Forward primer for 3′ RACE of U5 Helicase Protein CDS (GL50803_9352) for detection of processing near its 3′ motif sequence. Primer anneals within protein coding region -37 to -47 nt. upstream of the predicted 3′ motif sequence.
oAH117	CGT GTG GTT GTT CTT TGG TG	Reverse primer for 5′ RACE of *G. lamblia* U4 snRNA candidate (Candidate-11). Primer is antisense to region -17 to -36 upstream of 3′ motif sequence.
oAH118	CAC CAA AGA ACA ACC ACA CG	Forward primer for 3′ RACE of *G. lamblia* U4 snRNA candidate (Candidate-11). Primer corresponds to -17 to -36 upstream of 3′ motif sequence.
oAH119	CCT GGC AGC TGC CTC ACA GC	Reverse primer for 5′ RACE of *G. lamblia* U2 snRNA candidate (Candidate-14). Primer is antisense to region -35 to -54 upstream of 3′ motif sequence.
oAH120	GCT GTG AGG CAG CTG CCA GG	Forward primer for 3′ RACE of *G. lamblia* U2 snRNA candidate (Candidate-14). Primer corresponds to region -35 to -54 upstream of 3′ motif sequence.
oAH123	GCT GTA ATA CGA CTC ACT ATA GGC TAG GCT GAA GCT GCC AAG GTG CG	Forward primer for *G. lamblia* U4 snRNA (Candidate-11) *in vitro* transcription. Primer anneals to region +1 to +24 nt from predicted mature 5′ end. A T7 promoter sequence is underlined.
oAH124	AAT GAG AGA TTC GGC TGT GCC	Reverse primer for *G. lamblia* U4 snRNA (Candidate-11) *in vitro* transcription. Primer is antisense to region -21 to -1 nt. from mature 3′ end.
oAH125	GCT GTA ATA CGA CTC ACT ATA GGT AAC AAA AAC AGA	Forward primer for *G. lamblia* U6 snRNA candidate *in vitro* transcription. Primer

Name	Sequence (5′ to 3′)	Description
	GAC AGT TAG CAC C	anneals to region +1 to +26 nt from predicted mature 5′ end. A T7 promoter sequence is underlined.
oAH126	AAG GAG CGG GGT GCA AGC ACG	Reverse primer for *G. lamblia* U6 snRNA candidate *in vitro* transcription. Primer is antisense to region -20 to -1 nt. from mature 3′ end.
oAH133	CAG TGC TCT CAG ACC TGC TAC C	Forward primer for 3′ RACE of U1 snRNA candidate. Primer anneals to region -39 to -17 upstream of 3′ motif sequence.
oAH136	TCT CTG TTT TTG TTA CC	Giardia U6 snRNA candidate antisense primer to be used during *in vitro* U4/U6 complex formation. Is complementary to region +1 to +17 from mature 5′ end.
oAH137	GAA ACC AGA GGT CCC CCA GC	Giardia U6 snRNA candidate antisense primer to be used during *in vitro* U4/U6 complex formation. Is complementary to region +46 to +65 from mature 5′ end.
oAH139	GGC GAT TCG GCT GTG CCG TGT GG	Reverse primer for 5′ RACE of *G. lamblia* U4 snRNA candidate (Candidate-11). Primer is antisense to region -4 to -23 upstream of 3′ end motif sequence.
oAH140	GGC CCT TGC ATG TAG TGG ATC GG	Reverse primer for 5′ RACE of *G. lamblia* U2 snRNA candidate (Candidate-14). Primer anneals to region -63 to -83 nt. upstream of 3′ end motif sequence.
oAH141	GGG CTT TGA CAG AGA CCA CTG CGA G	Reverse primer for 5′ RACE of GlsR28. Primer anneals to region -51 to -72 nt. upstream of 3′ motif sequence.
oAH165	GAA TTT TCG TGA GAA GGA CC	Reverse primer to be used with oAH167 for RT-PCR detection of spliced Hsp90 trans intron 5′ half which has not been cleaved at downstream motif sequence. Primer is antisense to region +20 to +39 nt. downstream of motif sequence in Hsp90 intron 5′ half.
oAH166	GTC TCG CAC ACA TAC AGC	Reverse primer to be used with oAH167 as an RT-PCR positive control to demonstrate the ability to generate a PCR amplicon after RT using a forward primer (oAH167) which anneals upstream of branch point sequence of the *Giardia* Hsp90 trans-intron 3′ half. Primer is antisense to region upstream of motif site and +18 to +35 nt. downstream of the 5′ splice site.
oAH167	GAG TTC TCG CAC ACA TAC	Forward primer for PCR step after RT

Name	Sequence (5′ to 3′)	Description
	AG	using reverse primers oAH165 and oAH166. Primer is sense to region –20 to –39 nt. upstream of Hsp90 trans-intron branch point sequence in the intron 3′ half (overlaps nucleotides predicted for intermolecular association of intron halves)
oAH168	CGG TGG TCT GAC TTC TCG	Reverse primer to be used with oAH170 for RT-PCR detection of spliced DHC beta trans intron #1 5′ half which has not been cleaved at downstream motif sequence. Primer is antisense to region +28 to +45 nt. downstream of motif sequence in DHC beta intron #1 5′ half.
oAH169	GAG CGC GAA GAC ATA TAA GC	Reverse primer to be used with oAH170 as an RT-PCR positive control to demonstrate the ability to generate a PCR amplicon after RT using a forward primer (oAH170) which anneals upstream of branch point sequence of the *Giardia* DHC beta trans-intron #1 3′ half. Primer is antisense to region upstream of motif site and +37 to +56 nt. downstream of the 5′ splice site.
oAH170	CAC CGA CAG AAG CCT TTC C	Forward primer for PCR step after RT using reverse primers oAH168 and oAH169. Primer is sense to region –8 to –26 nt. upstream of DHC beta trans-intron #1 branch point sequence in the intron 3′ half (portion overlaps nucleotides predicted for intermolecular association of intron halves)
oAH171	AAT AGA ACA GAG CAA CTG CAC	Reverse primer to be used with oAH173 for RT-PCR detection of spliced DHC beta trans intron #2 5′ half which has not been cleaved at downstream motif sequence. Primer is antisense to region +6 to +26 nt. downstream of motif sequence in DHC beta intron #2 5′ half.
oAH172	CTG GCA TAC TGA CAT AAT GAC	Reverse primer to be used with oAH173 as an RT-PCR positive control to demonstrate the ability to

Name	Sequence (5′ to 3′)	Description
		generate a PCR amplicon after RT using a forward primer (oAH173) which anneals upstream of branch point sequence of the *Giardia* DHC beta trans-intron #2 3′ half. Primer is antisense to region upstream of motif site and +46 to +66 nt. downstream of the 5′ splice site.
oAH173	CTC ACA GCT TTA CTG ACC AG	Forward primer for PCR step after RT using reverse primers oAH171 and oAH172. Primer is sense to region –7 to –26 nt. upstream of DHC beta trans-intron #2 branch point sequence in the intron 3′ half (portion overlaps nucleotides predicted for intermolecular association of intron halves)
oAH174	TCA TTT TCG GCC CAA GTA TTG	Reverse primer to be used with oAH176 for RT-PCR detection of spliced DHC gamma trans intron 5′ half which has not been cleaved at downstream motif sequence. Primer is antisense to region overlapping nt. 10 to 12 of motif sequence and +1 to +18 nt. downstream of motif sequence in DHC gamma intron 5′ half.
oAH175	CAC ATA CAC GCC AAG CAT AC	Reverse primer to be used with oAH176 as an RT-PCR positive control to demonstrate the ability to generate a PCR amplicon after RT using a forward primer (oAH176) which anneals upstream of branch point sequence of the *Giardia* DHC gamma trans-intron 3′ half. Primer is antisense to region upstream of motif site and +18 to +37 nt. downstream of the 5′ splice site sequence.
oAH176	CCG AGG GAC ACA CAC CTC	Forward primer for PCR step after RT using reverse primers oAH174 and oAH175. Primer is sense to region –23 to –40 nt. upstream of DHC gamma trans-intron branch point sequence in the intron 3′ half (overlaps nucleotides predicted for intermolecular association of intron halves)
oAH182	CTA ATA GGA GTT CTT GCT	Reverse primer for 5′ RACE of *Giardia*

Name	Sequence (5′ to 3′)	Description
	CTG	DNA polymerase catalytic subunit (GL50803_35094) encoded downstream of ‘Candidate-23′ ncRNA. Primer anneals to antisense region +1013 to +1033 downstream of the predicted ‘AUG’ start codon.
oAH183	GAC CTA ACA CGT CAA ATA GCC	Reverse primer for 5′ RACE of *Giardia* Ser/Thr Kinase (GL50803_4329) encoded downstream of ‘Candidate-5′ ncRNA. Primer anneals to antisense region +620 to +640 downstream of the predicted ‘AUG’ start codon.
oAH184	CAT TCT GGA ACA AGA GAT TGG	Reverse primer for 5′ RACE of *Giardia* Replication factor C, subunit 5 (GL50803_16127) encoded downstream of Hsp90 5′ trans-spliced exon. Primer anneals to antisense region +729 to +749 downstream of the predicted ‘AUG’ start codon.
oAH219	AAG TAA GTT GTG TGC GAT GC	Reverse primer for 5′ RACE of transcripts corresponding to the intergenic region between the Candidate-23 ncRNA (upstream) and DNA polymerase delta subunit (downstream) genes. Primer is antisense to region +8 to +27 nt downstream of Candidate-23 motif sequence.
oAH220	CTG CTA AGA CAC ATC ATT CAC	Reverse primer for 5′ RACE of transcripts corresponding to the intergenic region between the Candidate-5 ncRNA (upstream) and Ser/Thr Kinase (downstream) genes. Primer is antisense to region +21 to +41 nt downstream of Candidate-5 motif sequence.
oAH221	GAA TTT TCG TGA GAA GGA CC	Reverse primer for 5′ RACE of transcripts corresponding to the intergenic region between the Hsp90 5′ trans exon 1 (upstream) and Replication Factor C (downstream) genes. Primer is antisense to region +20 to +39 nt downstream of Hsp90 motif sequence.

Name	Sequence (5′ to 3′)	Description
oAH231	GTA TAC ACG ACT TCC CTA TAA C	Forward primer to be used with oAH165 or oAH166 during PCR after first strand cDNA synthesis. Primer anneals -25 to -4 nt upstream of the Hsp90 intron branch point A.
oAH232	GAA GCC TTT CCA GAC AAT AC	Forward primer to be used with oAH168 or oAH169 during PCR after first strand cDNA synthesis. Primer anneals -23 to -4 nt upstream of the DHC beta intron #1 branch point A.
oAH233	CTT TAC TGA CCA GAC ACA TAC	Forward primer to be used with oAH171or oAH172 during PCR after first strand cDNA synthesis. Primer anneals -24 to -4 nt upstream of the DHC beta intron #2 branch point A.
oAH234	GGG CCA CCT CTC CCT ACA GC	Forward primer to be used with oAH174 or oAH175 during PCR after first strand cDNA synthesis. Primer anneals -27 to -8 nt upstream of the DHC gamma intron branch point A.
oDE1	TTC TGA TGC GGA TAC CTT GC	Forward primer for the 3′-RACE of Candidate-17. Primer anneals -102 to -82 relative to the 3′ processing motif.
oDE3	CTT GCG TGC GCA TAT CTC C	Forward primer for the 3′-RACE of GlsR26. Primer anneals -73 to -54 relative to the 3′ processing motif.
oDE4	CTA CGT GTT ATG GGC AGC G	Forward primer for the 3′-RACE of GlsR27. Primer anneals -82 to -63 relative to the 3′ processing motif.
oDE5	TTC AAC TCA GCC GGA CAG C	Forward primer for the 3′-RACE of Candidate-15. Primer anneals -87 to -68 relative to the 3′ processing motif.
oDE7	TAG GTA GGG CCG ATG AGC	Forward primer for the 3′-RACE of Candidate-23. Primer anneals -86 to -68 relative to the 3′ processing motif.
oDE8	ACG AGG AAA CGA GTG TTT CG	Forward primer for the 3′-RACE of Candidate-5. Primer anneals -76 to -56 relative to the 3′ processing motif.
oDE9	GTA GTT ACT GTT TCG TCG GC	Forward primer for the 3′-RACE of U2 snRNA candidate (Candidate-14). Primer anneals -110 to -90 relative to the 3′ end processing motif.
oDE17	CTA TCC AGC CAA GAG CCG	Forward primer for RT-PCR and 3′ RACE of processing motif-containing CDS for DNA repair protein Rhp26p (GL50803_87205) transcript. Primer anneals +1005 to +1023 downstream of the predicted start codon and upstream of the predicted motif sequence.

Name	Sequence (5′ to 3′)	Description
oDE18	GGC GTC GTA GTA TGT AAG G	Reverse primer for RT-PCR detection of motif-containing CDS for DNA repair protein Rhp26p (GL50803_87205) transcript. Primer anneals +138 to +157 downstream of the predicted 3′ processing motif.
oDE19	GAT GCC TTA GAG AAC TTG CG	Forward primer for RT-PCR and 3′ RACE of motif-containing CDS for U5 200kDa Helicase (GL50803_9352) transcript. Primer anneals +2566 to +2586 downstream of the ‘AUG’ start codon and upstream of the predicted 3′ end motif sequence.
oDE20	CGT AGA TGA GAT AGC TGC C	Reverse primer for RT-PCR detection of motif-containing CDS for U5 200kDa Helicase (GL50803_9352) transcript. Primer anneals +146 to +165 downstream of the predicted 3′ end processing motif.
oDE21	ATG GGT CTG ACA CGT TGG	Forward primer for RT-PCR and 3′ RACE of process motif-containing Protein 21.1 (GL50803_25296) transcript. Primer anneals to region +538 downstream to +556 of the ‘AUG’ start codon and upstream of the predicted 3′ end processing motif.
oDE22	TAG ATG AAT CAG CGA TGT GC	Reverse primer for RT-PCR detection of process motif-containing Protein 21.1 (GL50803_25296) transcript. Primer anneals to region +991 to +1011 downstream of the ‘AUG’ start codon and downstream of the predicted 3′ processing motif.
oDE23	GTG ACC GTC CTC TGT ACG	Forward primer for RT-PCR and 3′ RACE of process motif-containing hypothetical protein (GL50803_7350) transcript. Primer anneals to region +3156 to +3174 downstream of the ‘AUG’ start codon and upstream of the predicted 3′ motif sequence.
oDE24	AAG ATG AAG GGC TCA AAA CG	Reverse primer for RT-PCR detection of process motif-containing Hypothetical Protein (GL50803_7350) transcript. Primer anneals to region +3618 to +3638 downstream of the ‘AUG’ start codon and downstream of the predicted 3′ end motif.