

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Non-protein-coding-RNA processing in
the deep-branching protozoan parasite
Giardia intestinalis

A thesis presented in partial fulfilment of the requirements
for the degree of PhD in Molecular Genetics

at Massey University, Palmerston North
New Zealand

Xiaowei (Sylvia) Chen

2008

Acknowledgement

It has been a challenging but rewarding passage towards the completion of this thesis. I am grateful to all the people who helped during the work.

At first I would like to express my sincere gratitude to my supervisor Prof. David Penny, co-director of Allan Wilson Centre, Massey University. His wide knowledge and inspiring ideas have been great value for all his students, including me. His excellent guidance provided a good basis for this work.

I am grateful to my co-supervisor Dr. Lesley J. Collins, Allan Wilson Centre, Massey University for her continuous support and detailed comments throughout this work. Her understanding and encouragement are valuable for me, and also it has been very enjoyable to share the same office with her.

I wish to express my warm thanks to my co-supervisor Prof. Jürgen Brosius, ZMBE, University of Münster, Germany, who kindly provided me the opportunity to work with them in Münster, and to learn the valuable skills required for completion of my work.

I also would like to express my gratitude to all the lab people who helped me during different periods of my work. In particular, I am grateful to our lab manager Trish McLenachan. Her warm personality and enthusiasm have been a great help to my smooth progress. Also it has been a great time to work with the people in Münster. I wish to thank Associate Prof. Jürgen Schmitz for his friendly support, Dr. Timofey S. Rozhdestvensky and Anja Zemann for their great help of constructing the RNA library.

For the excellent help on computational side of the work, I would like to express my special thanks to my good friend Tim White. He has provided high quality technical assistance as well as great entertainment during this work.

Also I would like to thank Dr. George Ionas and Errol Kuan in MicroAquaTech, Massey University for kindly providing *Giardia* cell culture. My sincere thanks are due to Dr. Kathryn Stowell for her kind instruction on radioisotope techniques, and to Dr. Gill Norris for her instruction on protein analysis.

For the critical and useful suggestions which gave me more thoughts about this work during the thesis examination, I would like to thank my examiners, Dr. Austen Ganley, Dr. Peter K. Dearden and Dr. Paul P. Gardner.

I owe my most sincere gratitude to my Mum and Dad for their love and support throughout my life. It is always a pure joy to see them every year.

Finally, thanks to all the people in Allan Wilson Centre. Life has been great here.

This work is supported by the New Zealand Marsden fund; the Allan Wilson Centre for Molecular Ecology and Evolution; European Union (EU: LSHG-CT-2003-503022 and the Nationales Genomforschungsnetz (NGFN: 0313358A).

Content

	Page
Introduction to the thesis	1
Chapter 1: Evolution of non-protein-coding RNAs	
1.1 The origin of non-protein-coding RNAs	13
1.2 Relics from the RNA-World	
1.2.1 The history of universal house-keeping RNAs	18
1.2.2 The antiquity of catalytic RNAs	
1.2.2.1 Small catalytic RNAs	23
1.2.2.2 Self-splicing introns	25
1.2.2.3 RNase P and MRP	28
1.2.3 Ancient RNA regulators – Riboswitches	31
1.3 Evolution of ncRNAs in eukaryotes	
1.3.1 Introns and their roles in eukaryotic evolution	33
1.3.2 Divergence of regulatory RNAs	
1.3.2.1 RNAs in the nucleus – small nuclear RNAs	35
1.3.2.2 RNA editing and small nucleolar RNAs (snoRNAs)	39
1.3.2.3 Enrichment of ncRNAs in eukaryotes and systematic gene regulation	42
1.4 The present and future of ncRNA evolution	44
Chapter 2: Identification of novel non-protein-coding RNAs from <i>Giardia intestinalis</i>	
Combined experimental and computational approach to identify non-protein-RNAs in the deep-branching eukaryote <i>Giardia intestinalis</i> <i>Nucleic Acids Research</i> 35 (14): 4619-4628.....	48

Chapter 3: Analysis of the ncRNA library of *Giardia intestinalis*

3.1 Background: Molecular biology of <i>Giardia intestinalis</i> and techniques in the studies of ncRNA	
3.1.1 <i>Giardia</i> – a deep branching unicellular eukaryote	59
3.1.2 The genome of <i>Giardia</i>	62
3.1.3 Techniques in the studies of ncRNAs.....	64
3.2 Analysis of the novel ncRNA candidates	
3.2.1 Novel ncRNA candidates from the cDNA library of <i>Giardia</i>	69
3.2.2 Analysis of potential promoter elements of characterized ncRNAs from <i>Giardia</i>	71
3.2.3 Analysis of the upstream sequence elements and internal sequence elements of novel <i>Giardia</i> ncRNAs	83
3.2.4 Polymerase III transcription factors of <i>Giardia</i>	88
3.2.5 Structural analysis of uncharacterized novel ncRNA candidates	92
3.3 Conclusion	97

Chapter 4: Studies of the major spliceosomal snRNAs in *Giardia*

4.1 Introduction: Does <i>Giardia</i> have a functional spliceosome?	99
4.2 Searching for U-snRNAs in <i>Giardia</i>	
4.2.1 Prediction of <i>Giardia</i> U1-snRNA candidate	102
4.2.2 Prediction of <i>Giardia</i> U2-snRNA candidate	106
4.2.3 Prediction of <i>Giardia</i> U6 and U4 snRNA candidates	110
4.3 <i>Giardia</i> homologue of Prp8 protein – the central protein component of the spliceosome	118
4.3.1 Bioinformatical analysis of the <i>Giardia</i> homologue of Prp8 protein	119
4.3.2 Analysis of the potential RNA-recognition motif of <i>Giardia</i> Prp8 protein	
4.3.2.1 Computational analysis of the likelihood of Gp8d1	

being a protein domain	122
4.3.2.2 Cloning and <i>in vitro</i> recombinant protein expression	126
4.3.2.3 RNA-protein binding assays of Gp8d1 <i>versus Giardia</i> ncRNA candidates	127
4.4 Conclusion and overview of the major spliceosomal components in <i>Giardia</i>	131
4.5 Experimental materials and methods	
4.5.1 PCR amplification and cloning	132
4.5.2 Plasmid preparation	133
4.5.3 <i>In vitro</i> recombinant protein expression	133
4.5.4 <i>In vitro</i> RNA transcription	134
4.5.5 Affinity capturing of proteins with ability to bind U5- and U1-snRNA Candidates	134
4.5.6 Gel shift assay with radio-isotope labelled RNAs	135
4.5.7 Reverse transcription (RT) PCR	135
Chapter 5: Unusual ncRNAs in <i>Giardia</i> and the putative RNAi pathway	
5.1 Introduction: The mechanism of dsRNA-induced gene silencing and RNAi in unicellular eukaryotes	137
5.2 The unusual ncRNA repeats in <i>Giardia</i>	144
5.3 Protein components of the putative RNAi pathway in <i>Giardia</i>	152
5.4 The possible existence of a truncated <i>Giardia</i> Dicer protein	160
5.5 Conclusion and overview of the unusual RNAs found in this study....	167
5.6 Experimental material and methods	
5.6.1 Primers used in generating <i>in vitro</i> transcription template for Girep RNAs and Girep1RNA and primers used for recombinant truncated Dicer peptide	168
5.6.2 Recombinant truncated Dicer peptide expression and purification	168
5.6.3 Nuclease activity assay	169

Final Words	171
Appendix-1	175
Appendix-2	215
Appendix-3	237
Appendix-4	257
Bibliography	263

Introduction to the thesis

For decades molecular evolutionary studies have built our understanding of the natural history of life based on complexity of individual types of organisms. The complexity of a biological organism is generally determined by the number of cell types (Vogel and Chothia 2006) and degree of cellular organization. These morphological features are in turn the phenotypical representation of genomic complexity. Information within a genome can be accessed at different levels: from raw DNA sequences to structured and functional molecules, such as RNAs and proteins. Molecular evolution currently uses genetic and genomic information to understand the evolution of whole organisms, and is based on the concept that all biological functions are the result of continuous evolution from their ancestral forms. With recent advances in genome science, it has come to light that eukaryotic genomes consist of mainly non-protein-coding sequences, which were once neglected for their important roles in evolution. This thesis presents work on the non-protein-coding RNAs from the deeply diverged eukaryote: *Giardia intestinalis*, aiming towards better understanding the evolution of eukaryotes.

The importance of non-protein-coding sequences has been gradually realized since the two important paradoxes in molecular biology became evident: (1) The C-value paradox, which describes the inconsistency between cellular DNA content and biological complexity; (2) The G-value paradox which describes the inconsistency between gene numbers and biological complexity.

The C-value was confusingly defined ranging from the complete complement of DNA per nucleus (in picograms); to the amount of DNA in the haploid genome (Swift 1950). Thus, the C-value represents the crude estimation of DNA regardless of the sequence composition, and is affected by a number of variables including polyploidy*, gene duplication, repetitive sequences and experimental errors, of which polyploidy may be the main factor. Table-1

* Polyploidy: the state of having more than two full sets of homologous chromosomes.

shows some examples of DNA content (C-values) in a number of different organisms.

Table-1: DNA content (C-value) in example species

Class	Example species	Common name	C-value (pg)
Mammals	<i>Homo sapiens</i>	Human	3.50
Aves	<i>Haliaeetus leucocephalus</i>	Bald eagle	1.43
Amphibia	<i>Bolitoglossa striatula</i>	Salamander	62.70
Amphibia	<i>Bufo crucifer</i>	Crucifer toad	3.15
Secernentea	<i>Caenorhabditis elegans</i>	Nematode	0.08
Angiosperm	<i>Fritillaria assyriaca</i>	Fritillary	127.40
Algae	<i>Ostreococcus tauri</i>	-	0.01

(Bennett and Leitch 2005; Gregory 2005)

Amphibians and amoebae are two typical examples which show the lack of correlation between genome size and biological complexity (Becak and Kobashi 2004). For example, the C-value for lungfish indicates that its genome is over an order of magnitude larger than primates (Joss 2006), and it is known to be polyploid (Vervoort 1980). Groups of organisms such as crustaceans, insects and plants exhibit a wide range of C-values and are also often known to include polyploids (Otto and Whitton 2000). Polyploidy is thus an important reason for this inconsistency, because C-values are generally not corrected for polyploidy (Gregory 2005). It was later suggested that the relative complexity should be the minimum amount of information required for the operation of a biological system (Li and Vitanyi 1997). This is supported by the observation of Taft et al. that the minimum genome size increases consistently with the increase of complexity from nematode to insects to vertebrates (Taft et al. 2007). In addition to polyploidy, gene and genome duplication events also contribute to genome expansion (Ohno et al. 1968). It has been known for yeast that at least one round of whole-genome duplication has happened, followed by large-scale gene losses, and evolution of alternative gene paralogues (Scannell et al. 2006) which are often redundant. Similar evidences has been found in *Arabidopsis thaliana* (Thomas et al. 2006).

Another source of variation in genome size is transposon-derived sequences (Brosius 1991; Kidwell 2002), often referred to as “repetitive” sequences, which comprise about half of human (Lander et al. 2001) and mouse genomes (Waterston et al. 2002) and was once considered to be non-functional. However, increasing evidence (Peaston et al. 2004; Bejerano et al. 2006; Nishihara et al. 2006; Xie et al. 2006) has suggested functions for at least some of these previously neglected sequences, and it is unclear what proportion of the genome is contributing to the genetic complexity. Therefore, instead of representing measurements of biological complexity, the C-value may only reflect the quantitative amount of raw genetic material, and which is continuously subjected to sequence acquisition and loss over evolutionary time.

Compared with C-value, the G-value appears to be even more problematic, since the former can largely be explained by polyploidy and other raw evolutionary material, the latter is based on the assumption that the number of protein-coding-genes scales with biological complexity (Bird 1995). However, this assumption is not supported and hence is termed the G-value paradox (Hahn and Wray 2002). The latest estimates from genomic surveys suggest that humans have approximately 20,000 protein-coding genes (Goodstadt and Ponting 2006). The number is similar to other vertebrates such as chicken (Wallis et al. 2004) and pufferfish (Aparicio et al. 2002), and also to the nematode worm *Caenorhabditis elegans* (Stein et al. 2003) which comprises only 1,000 cells. Despite the developmental complexity, mammals do not appear to have more protein-coding genes (or according to the present estimates, even less) than plants (~26,000 for *Arabidopsis*) (Haas et al. 2005) or protists such as *Paramecium* (~40,000) (Arnaiz et al. 2007) and *Tetrahymena* (~27,000) (Eisen et al. 2006). Although part of the G-value paradox can be explained by the increase of alternative splicing in mammals (Nagasaki et al. 2005), which allows greater range of different proteins to be expressed from a single source, there is also evidence showing that complex organisms utilize a wider range of regulation mechanisms to control gene expression: including chromatin modification, RNA-modification and editing, RNA localization and stability, transcriptional and translational silencing. These regulatory networks have been suggested to override the complexity of

protein-coding genes and possibly dominate the information content of genomes (Mattick 2004; Mattick and Gagen 2005). More importantly, most of the regulatory information resides outside of the protein-coding sequences.

Unlike prokaryotes, whose genomes consist of closely packed protein-coding sequences, eukaryotic genomes contain larger amount of intronic and intergenic non-protein-coding sequences, which total nearly 98% in humans (Little 2005). Both cDNA and genomic tiling array analysis of transcription have revealed that large proportions of eukaryotic genomes are transcribed (Mockler et al. 2005; Ranz and Machado 2006). At least 70% of mammalian genomes can be transcribed (Carninci et al. 2005). Similarly, majority of the *Drosophila* genome is transcribed (Manak et al. 2006). In addition, many mammalian genes have antisense transcripts which have been shown to have a regulatory role (Katayama et al. 2005). Hence it has been suggested that the biological complexity is strongly correlated with the proportion of non-protein-coding sequences (nc) within the total size (tg) of the genome (nc/tg), with corrections to complete and partial polyploidy (Mattick 2004). Using the nc/tg annotation, organisms with observed different biological complexity can roughly be clustered into comprehensive groups, with examples that the two protists *Tetrahymena* and *Paramecium*, which have unexpectedly large amount of protein-coding genes, can be clustered with other unicellular eukaryotes (Taft et al. 2007).

Consistent with the nc/tg annotation, developmentally more complex organisms contain larger number of introns, which are known, at least in vertebrate, to house most small nucleolar RNAs (Kiss 2006) and some microRNAs (Li et al. 2007). In the “intron-early” versus “intron-late” both hypotheses are based on introns being devoid of functions (de Souza 2003), and evolving neutrally (Lynch 2006). In contrast, it has been suggested that reduction and expansion of introns in complex organisms is resulted from selection of functions encoded in them (Fedorova and Fedorov 2003). Analysis of intron-distribution versus gene function from various organisms (Taft et al. 2007) has shown that intron-length and distribution are not random despite little sequence conservation. Recent studies in yeast showed that some introns

could improve transcription and translation (Juneau et al. 2006). Also, a number of studies showed that highly expressed house-keeping genes are usually compact with reduced introns (Vinogradov 2006; Pozzoli et al. 2007). Based on the hypothesis that many yet uncovered *cis*-acting and *trans*-acting non-protein-sequences may reside in introns (Taft et al. 2007), it can be understood that evolutionarily more conserved house-keeping genes generally require less tissue-specific regulation, thus contain fewer introns, compared with intron-rich genes involved in higher order functions such as development and differentiation. Also streamlining transcription and translation in highly expressed genes can produce a selective pressure to remove introns.

For many years, molecular biology has focused on proteins, whereas RNA remained as an intermediate of gene expression. But the relationship between non-protein-coding sequences and biological complexity has made a strong suggestion that non-protein-coding RNAs (ncRNAs) play important roles in evolution of eukaryotes. Sequence comparisons from different complexity levels have exhibited significant conservation of ncRNAs (Dermitzakis et al. 2003; Inada et al. 2003). Many of the ncRNAs share conserved structures (Torarinsson et al. 2006), expressional control mechanisms (Carninci et al. 2005; Katayama et al. 2005; Ravasi et al. 2006), and specific cellular locations in some cases (Prasanth et al. 2005; Ginger et al. 2006; Pollard et al. 2006), suggesting a selective evolution upon structure-function constrains. However, primary sequences of ncRNAs can evolve rapidly thus appear less conserved. For example, the *cis*-regulatory sequences in vertebrates often undergo shuffling and expansion to form new elements (Sanges et al. 2006). In addition, microarray studies indicate that natural selection on *cis*- and *trans*-acting elements leads to transcriptional variation over evolutionary time (Ranz and Machado 2006).

All evidence points to the suggestion that a large proportion of the regulatory network (e.g. transcription, translation, epigenetic control) in eukaryotes involve functions of ncRNAs, which make a major contribution to increasing genetic information in complex organisms and play important roles in all levels of genetic control. The functions of ncRNA extend from basic transcription and

translation (tRNAs, rRNAs) to complex genetic phenomena such as imprinting (Yazgan and Krebs 2007), RNA interference (Bernstein et al. 2001) and chromatin modification (Bernstein and Allis 2005). Increasingly new classes of ncRNAs are uncovered including a large number of snoRNAs which modify other RNAs (Bachellerie et al. 2002), miRNAs which regulate a wide range of developmental processes in animals and plants (Meister and Tuschl 2004), and piRNAs for which the function is not yet clear but evolve rapidly (Aravin et al. 2006). The fact that ncRNA plays a central role in eukaryotes also suggests their importance in eukaryotic evolution. The consistency between the amount of non-protein-coding sequences and biological complexity (Taft et al. 2007) indicates a strong element of ncRNAs in the evolution of complex organisms.

While recent studies have mostly focused on ncRNAs from complex multicellular eukaryotes, the facts behind the emergence and divergence of ncRNAs during early stages of eukaryotic evolution still remains unknown. Evolution of eukaryotes consists of several different stages: formation of current eukaryotic cellular structure, emergence of multicellularity, diversification of cell types, and finally formation of complex developmental mechanisms. Studies of ncRNAs in developmentally complex eukaryotes have revealed the fascinating roles of ncRNAs in the evolution of genetic controls which are crucial for the formation of new systems. However, the formation of tightly controlled systems must be a gradual process which originated from the 'basic', but perhaps no less complicated, ancestral state (Kurland et al. 2006). The presence of large number of ncRNAs in eukaryotes in contrast to the much less ncRNA content in prokaryotes (Gottesman 2005) supports the idea that the evolution of complex ncRNA-involving mechanisms is one of the key factors in the evolution from single-cellular eukaryotes to complex organisms. Therefore exploring the evolution of ncRNAs in the framework of cellular biochemistry can help understanding some fundamental questions, especially during the early stages of eukaryotic evolution. In order to answer these questions, it is necessary to focus on the ncRNA-involving mechanisms in currently available single cellular eukaryotes which still exhibit ancestral features of biochemistry and metabolism.

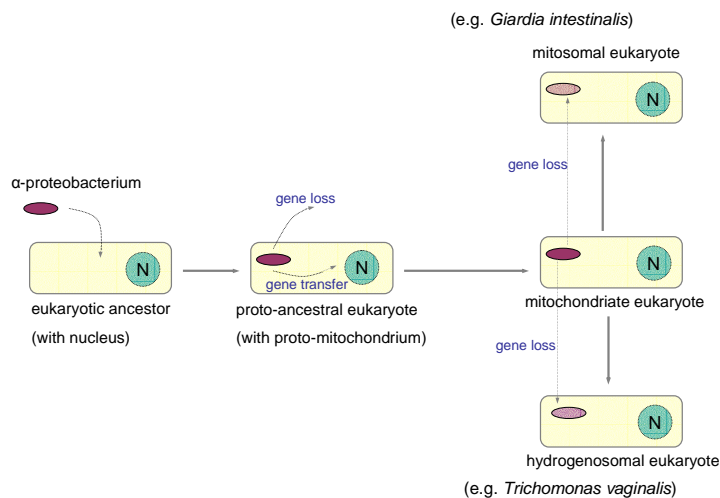
It still remains a question as what extant single cellular eukaryotes best represents the ancestral eukaryotes. The basic topology of eukaryotic tree* was constructed with the commonly used small-subunit ribosomal RNA genes (Sogin 1991). Subsequently several trees constructed by protein sequences (Keeling and Doolittle 1996; Baldauf and Doolittle 1997) showed many alternatives for several branches. More recently the studies of organelles and their interactions with other cellular components have extended the view of evolutionary biology. Reconstructing the eukaryotic phylogeny has become a process of synthesising all kinds of data including single gene trees, multi-gene trees, and structural characteristics of cellular architecture. A recent eukaryotic tree (Keeling et al. 2005) has grouped all the eukaryotic lineages into five large groups, where the group of excavates (Simpson 2003) contains a diverse range of parasitic and/or anaerobic protists, including ultrastructurally simple organisms such as *Giardia* and *Trichomonas*. These protists were once considered to be the closest remnants of ancestral eukaryotes due to lacking key organelles for generating energy such as mitochondria and hydrogenosomes. New biochemical and ultrastructural methods have been used to reinvestigate the processes involved in energy generation of these protists, and results indicate that instead of being “primitive”, they possess organelles which can be regarded as the reduced form of mitochondria, such as mitosomes identified in *Giardia* (Tovar et al. 2003) and *E. histolytica* (Tovar et al. 1999). Therefore these protists cannot be classified as the most ancestral eukaryotes, although there is not yet a defined root for the eukaryotic tree. Here in this study, the protists within the group of excavates are termed “deep-branching” eukaryotes due to their complex evolutionary heritage, because no matter where on the tree they fall they are distantly related to everything else.

Although it is now certain that excavates are not ancestors of modern eukaryotes (Tovar et al. 2003; van der Giezen and Tovar 2005), they can still reflect certain ancestral features of ancient eukaryotes. It is now widely accepted that all modern eukaryotes are evolved from an ancestor possessing mitochondria (Embley and Martin 2006), because there has not been a modern

* The tree in evolutionary field is a tree showing the evolutionary relationships among various biological species or other entities that are thought to have a common ancestor.

eukaryote found to be lacking all mitochondrial functions. Protists such as *Giardia*, *Trichomonas*, *Entamoeba*, anaerobic fungi and ciliates are very similar in a sense that they possess either reduced form of mitochondria (Tovar et al. 1999; Tovar et al. 2003) or hydrogenosomes that produce ATP (Dyall et al. 2004). It has also been shown that mitochondria and hydrogenosomes are two forms of the same fundamental organelles that share a common origin (Embley et al. 2003). Biological and phylogenetic data favours the hypothesis that ancestral eukaryotes were evolved from ancient phagocytotic cells which had ability to capture food through endocytosis, and later acquired mitochondria by engulfing the ancestral α -proteobacteria (de Duve 2007; Poole and Penny 2007). However, the relatively anaerobic living environment for protozoan* parasites, such as *Giardia* and *Trichomonas*, have gone through reductive evolution (Figure-1) and resulted in reduced organelles and genomes compared to higher eukaryotes.

Figure-1: Reductive evolution of some deep-branching eukaryotes



Most experimental data available is consistent with mitochondria-related organelles: (mitosomes and hydrogenosomes) are vertical descendants from proto-mitochondria, which were taken by the eukaryotic ancestor through endosymbiosis (van der Giezen and Tovar 2005). This evolutionary pathway involves multiple rounds of gene transfer to the nucleus and also gene loss.

* Protozoa (in Greek *proto* = first and *zoa* = animals) are one-celled eukaryotes that commonly grouped in the kingdom Protista together with the plant-like algae and fungus-like water molds and slime molds.

Although the currently available protozoan genomes have not been thoroughly analysed for ncRNA content, it is likely that the ncRNA contents in these organisms are also reduced based on their reduced genome size (<http://www.sanger.ac.uk/Projects/Protozoa/>). The deep phylogeny of eukaryotes is not yet known, but looking for common features within all deep-lineages of eukaryotes will possibly reveal features of ancestral eukaryotes (Collins and Penny 2005). In addition, deep-branching eukaryotes are biologically much less complex than higher eukaryotes, and are good modern models for looking into true basal eukaryotes in evolutionary biology. Extracting genetic information from these organisms can provide important insights into the evolution of eukaryotes.

In this thesis, one of the deepest-branching eukaryotes, *Giardia intestinalis* is used as the main model organism for the study of ncRNAs with an aim to better understand the early evolution of eukaryotes. The thesis is divided into five chapters, which are summarized as below.

Chapter-1: Evolution of non-protein-coding RNAs – A review of current literature

To date, genetics and molecular biology have shown that RNA, being one of the most important molecules in biology, is involved in almost all key mechanisms of house-keeping, regulation of gene expression and development in all living organisms (Mattick and Makunin 2006). The antiquity of RNA-driven biological machinery can be traced back to the hypothetical RNA world (Gilbert 1986; Brosius 2005). Modern organisms at different degrees have inherited the basic features of ancestral RNAs and a wide range of RNA-involving genetic pathways have evolved during eukaryotic evolution: These include diversification of class and processing mechanism of ncRNAs. This chapter reviews up-to-date literature on various classes of eukaryotic ncRNAs from their current functions to their history of evolution, building up a comprehensive background of the network of functional ncRNAs currently existing in eukaryotes. This review helps to select directions of study and also provides a standard which can be compared with the information collected from the model organism *Giardia*.

Chapter-2: Identification of novel ncRNAs from *Giardia intestinalis*

A cDNA library was constructed from RNA sized 70 to 600nt purified from total *Giardia* RNA. Sequencing and structural analysis identified a number of typical eukaryotic small ncRNAs while most of the ncRNAs identified from this library did not exhibit any conservation with known ncRNAs from other model organisms studied to date. Following computational predictions using a modified Snoscan programme (Lowe and Eddy 1999) I have identified putative candidates of C/D-box snoRNAs from the *Giardia* genome. In addition, unusual dsRNAs were found in *Giardia*. Results from this project suggest that the genetic information encoded in ncRNAs of *Giardia* may differ considerably from the standard context of ncRNAs in higher eukaryotes, though the key characteristic ncRNAs of eukaryotes such as snoRNAs and RNase P are present.

Chapter-3: Analysis of the ncRNA library of *Giardia intestinalis*

This chapter extends the analysis on the *Giardia* cDNA library. After more clones were sequenced, the collective data agrees with my first observation that *Giardia* possesses many currently uncharacterized novel ncRNAs. The reason behind the phenomenon is not yet understood, but is likely to have resulted from long evolutionary deviation of *Giardia* from the major groups of eukaryotes. Nonetheless, analysis of expressional patterns of various classes of ncRNAs in *Giardia* reveals conservation of certain upstream sequence motifs within proposed promoter regions. The transcription apparatus in *Giardia* is known to be highly reduced (Best et al. 2004). New information obtained from my studies about the potential features of ncRNA transcription in *Giardia* may lead to further investigation of the transcriptional systems in distant eukaryotes. In addition, potential new protein candidates of *Giardia* RNA polymerase system are presented here. Finally, detailed structural analysis of the novel ncRNA candidates has been performed using specialized RNA structural alignment tools. Results indicate a number of conserved structures within these novel ncRNAs of *Giardia*.

Chapter-4: Studies of the major spliceosomal snRNAs in *Giardia*

Messenger RNA splicing is one of the best studied RNA-processing pathways in eukaryotes. The key macromolecular machinery – the spliceosome, which catalyses the splicing reaction, is composed of five snRNAs and over 200 proteins in human (Nilsen 2003). Because introns are major genetic elements in eukaryotes, the spliceosome is usually highly active and the components of the spliceosome are highly conserved. Interestingly, in *Giardia* there have been only three spliceosomal introns published to date (Nixon et al. 2002; Russell et al. 2005), though others may be present (personal communication with Scott Roy, NIH). This gives rise to the question that whether *Giardia* possesses a complete spliceosome. It is puzzling if a complex spliceosome evolved in this organism just for splicing three introns. However on the other hand, the presence of introns strongly suggests the existence of a spliceosome. A number of studies have identified *Giardia* protein homologues involved in mRNA splicing (Nixon et al. 2002; Collins and Penny 2005), but the presence of the five snRNAs (apart from one candidate: the U5-snRNA), which are believed to be the key catalytic components of the spliceosome (Valadkhan 2005; Valadkhan et al. 2007), is not certain. In this chapter, computational predictions for four spliceosomal snRNAs are carried out, followed by analysis of expression and one central protein component of the spliceosome (Prp8) is studied using biochemical methods.

Chapter-5: Unusual ncRNAs in *Giardia* and the putative RNAi pathway

A number of transcribed dsRNAs (double-stranded RNAs) in *Giardia* has raised my interest to further look into their unusual features. dsRNAs are known to be largely involved with gene silencing mechanisms in various eukaryotic organisms. The pathways of dsRNA triggered gene silencing are reviewed in this chapter. Recent biochemical studies have characterised Dicer: the key protein component (Bernstein et al. 2001) of the RNAi mechanism from *Giardia* (Macrae et al. 2006). This finding reinforces the earlier suggestions that *Giardia* uses RNAi to regulate gene expression (Ullu et al. 2004), however the *Giardia* endogenous RNAs which are possibly involved in gene silencing are not yet discovered. Several long tandem repeats of dsRNAs have been observed to be highly transcribed, and some of them undergo self-

cleavage in the presence of divalent metal ions. The transcriptional patterns and sequences of these novel dsRNAs are analysed in this section. Results show that they are likely to be candidates of Dicer protein substrates, although further verification is still needed. In addition, an earlier study discovered a truncated transcript of Dicer mRNA, which led to investigations of the individual RNase III domain of *Giardia* Dicer protein. The possibility of RNA-induced silencing in *Giardia* is also reviewed.

In all, studies conducted in this thesis provide a systematic view of the ncRNA-encoded genetic information in *Giardia*, which is chosen as a model organism representing evolutionarily reduced, deep-branching eukaryotes. By exploring some key RNA-processing pathways in *Giardia*, various lines of evidence are collected to enable constructing the framework of ncRNA-regulated mechanisms in a deep-branching eukaryotic model. A number of novel ncRNAs with no obvious homology to currently known ncRNAs are studied from structural and expressional pattern points of view. Results have extended the current vision of ncRNAs in model eukaryotes. The analysis of knowledge obtained from *Giardia* is done comparatively, based on the current understanding of ncRNA functions in complex biological pathways. It is hoped that the study of ncRNAs in a deep-branching model organisms can unearth previously unknown information about the relationship between ncRNAs and biological evolution and aid in a better understanding of eukaryotic evolution.

Chapter-One: Evolution of non-protein-coding RNAs

Abstract

To date genetics and molecular biology have shown that RNA, being one of the most fundamental molecules in biology, is involved in almost all key mechanisms of house-keeping, regulation of gene expression and development in all living organisms. The antiquity of RNA-driven biological machinery can be traced back to the hypothetical RNA world. Modern organisms at different degrees have inherited the basic features of ancestral RNAs and a wide range of RNA-involving genetic pathways have evolved during eukaryotic evolution: These include diversification of classes and processing mechanisms of ncRNAs. This chapter reviews the up-to-date literatures on various classes of eukaryotic ncRNAs, from their current functions to their history of evolution. It builds up a comprehensive background of the network of functional ncRNAs currently existing in eukaryotes. This review helps in selecting directions of study and also provides a standard which can be compared with the information collected from the model organism *Giardia*.

This chapter outlines many of the ideas about non-coding (nc) RNAs and describes several different classes. The earlier parts of this chapter, where the proposed RNA and RNP worlds are described, is more hypothetical and speculative. Then the different classes of ncRNAs are discussed based on experimental and computational data. Nevertheless in order to well understand the evolution of ncRNAs it is appropriate to discuss the RNA-world first.

1.1 The origin of non-protein-coding RNAs

Modern life on earth is based on a cellular form of reproduction system maintained by a network of biochemical pathways, which use energy and generate material required for the continuous operation of the living system. The total information of any living organism is encoded in the form of DNA, which is transcribed into RNA and usually translated into proteins, which in turn join into the complex construction of cellular structures and metabolism. The information flow from DNA to RNA to protein has been described as the “Central Dogma” of molecular biology (Crick 1970). It was once generally accepted that the functions of cellular machines were determined by proteins. Therefore during early days of biological study, it was thought that the complexity of an organism, which was encoded within its genome, was

correlated with the amount of protein-coding DNA. However, more recent experimental data does not agree with this assumption. Latest surveys of sequenced genomes suggest a relatively stable number of protein-coding genes across eukaryotic organisms with wide range of biological complexity (Adam 2000; Aparicio et al. 2002; Waterston et al. 2002; Stein et al. 2003; Haas et al. 2005; Little 2005; Arnaiz et al. 2007). Recent data collected from a large number of genome-wide transcriptional studies suggests that the large quantity of non-protein-coding sequences in eukaryotes is strongly correlated with biological and developmental complexity (Taft et al. 2007), as well as the evolution of modified gene functions (Dermitzakis et al. 2003; Fedorova and Fedorov 2003; Ranz and Machado 2006).

Nowadays, as the importance of non-protein-coding sequences of genomes has been widely accepted, increasing number of studies have been carried out for the discovery of novel non-protein-coding RNAs (ncRNAs) (Aspegren et al. 2004; Inagaki et al. 2005; Pang et al. 2005; Tang et al. 2005; He et al. 2006; Huttenhofer and Vogel 2006; Mattick and Makunin 2006). To date various classes of ncRNAs have been characterised with wide range of functions including regulation of gene expression, modification of chromatin structures, and editing other RNAs. The functions of ncRNAs appear to diversify during eukaryotic evolution but some fundamental features remain conserved. That is to say, all the ncRNAs found today can be divided into two categories: catalytic and regulatory.

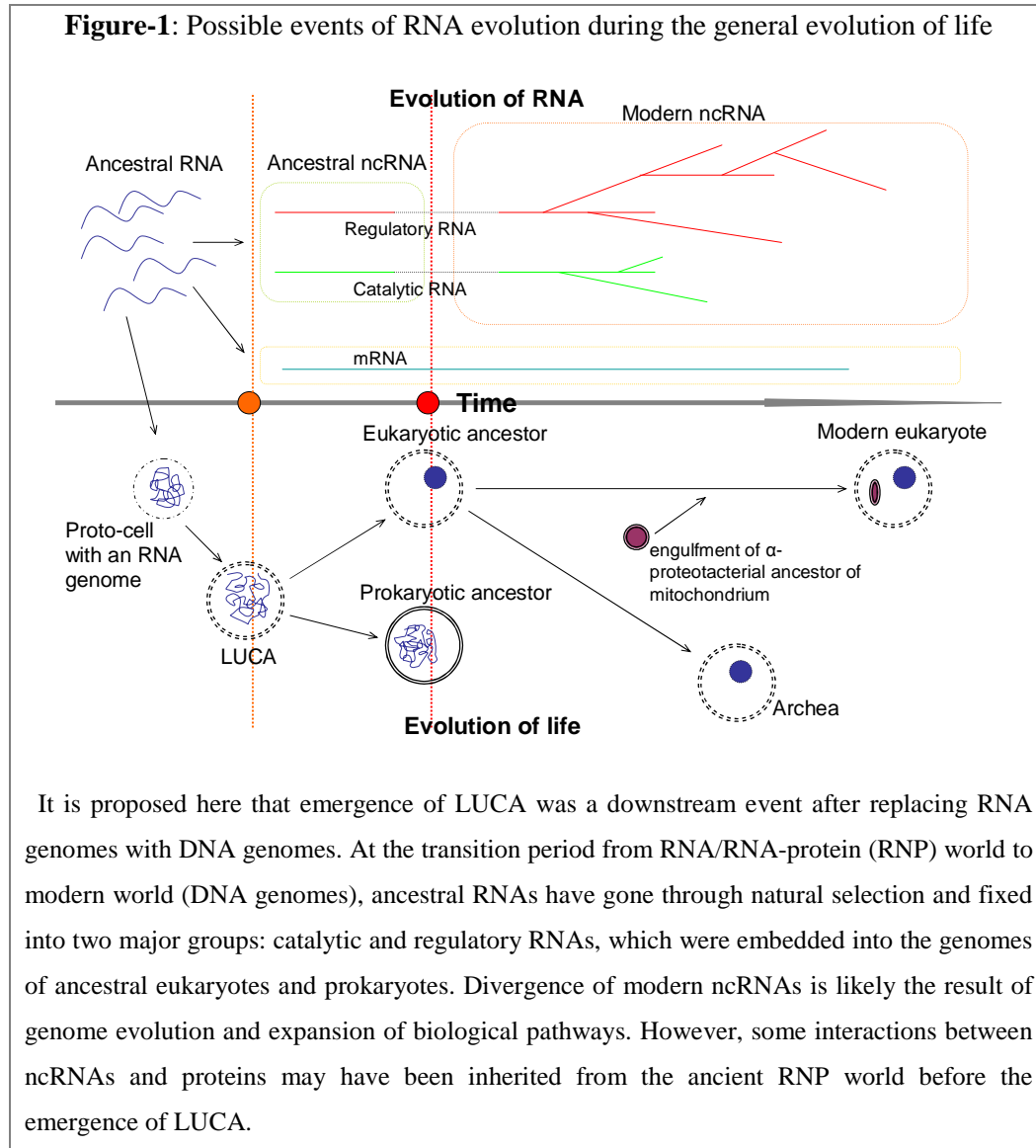
The origin of ncRNA is perhaps one of the earliest events when life emerged on earth. The theory of “RNA-world” and “RNA-protein (RNP)-world” (Gilbert 1986; Brosius 2005) suggests that self-replicating RNAs were one of the first forms of “life”. The versatile features of RNA molecules support the hypothesis of an RNA-constructed self replicable system. First, RNA stores information in the same way as DNA. Second, single-stranded RNA molecules are highly flexible in forming secondary and tertiary structures, like proteins, they can form enclosed reactive centres and behave as enzymes in an aqueous environment. Although the catalytic ability of modern RNA enzymes (ribozymes) is limited to ligation and nucleotide-cleavage (Doherty and

Doudna 2000), these basic reactions should have been the key components of an ancient RNA-constructed self-replicating system, as suggested in simulation studies (Lehmann 2002). The catalytic potential of ancient RNAs is expected to be more diverse than present (Huang et al. 2000), however the enzymatic features of ancient RNAs were gradually lost when proteins emerged and took over the catalytic roles, while a few catalytic features remained due to the fact that they are not limited by the rate of catalysis (Jeffares et al. 1998). The remnants of RNA catalysis might be seen in coenzymes, such as FAD, NAD and NADP, which are dinucleotides with a ribonucleotide structure (Huang et al. 2000).

In contrast to catalytic RNAs, regulatory RNAs cover much a wider range on the biological landscape. The variety of regulatory RNAs is associated with the complexity of metabolism and developmental stages. Collective information from the main ncRNA database Rfam (Griffiths-Jones et al. 2003) shows that eukaryotes have a larger variety of regulatory RNAs than prokaryotes, and also eukaryotes and prokaryotes do not usually share common types of regulatory RNAs. It may be difficult to trace the origins of different regulatory RNAs to particular ancestors from the RNA-World. However, it is clear that the evolution and diversification of regulatory RNAs are continuous processes closely associated with the evolution of cellular life. In addition, regulatory RNAs have made better adaptation in eukaryotes, suggesting that the positive natural selection of ncRNA functions (Taft et al. 2007) is one of the important elements in the evolution of complex biological mechanisms.

All living organisms can be grouped into the three kingdoms of life (Woese et al. 1990). However the passage that leads to the first cellular life with DNA genome from the theoretical RNA-World still remains unclear. Combining evolutionary data provides evidence of a possible “Last Universal Common Ancestor (LUCA)” of modern life (Doolittle 1999). It has also been suggested that at least a number of basic catalytic RNAs and regulatory RNAs were present in LUCA (Jeffares et al. 1998). Although it is not certain how long it took for the first cellular life to evolve, the emergence of LUCA could be a landmark that separated two phases of RNA evolution: the ancient and modern.

Clearly there can be no consensus yet on what such RNA or RNP-World would have looked like. Therefore ideas are tentative and Figure-1 shows the proposed events of RNA evolution in parallel with the evolution of cellular life from the RNA-World to modern times.



(A) Ancient RNA evolution: This covers the earliest stages of chemical transformation from non-reproducible materials to biological systems with genetic inheritance. This process could be random until the initial sequence pool expanded up to a saturation point, which was restricted by the replication rate, accuracy, and catalytic ability, of ancestral RNA molecules. Eventually, faster replicating RNAs gained evolutionary advantage and dominated the sequence pool. These catalytic RNAs could then evolve into the equivalent of

modern ribozymes. Another possible feature of ancestral ribozymes was likely to be the ability to utilize amino acids during catalysis, and hence amino acids could be absorbed into the replication process of RNA. The consequence of amino-acid incorporation was the emergence of protein which, enabled production of longer RNA molecules, formation of cellular structures, reduction of RNA to DNA, and eventually emergence of LUCA (given proteins have high catalytic rate and accuracy).

(B) Modern RNA evolution: In contrast to ancient RNA evolution, the principle of modern RNA evolution is no longer ‘information expansion’, but ‘optimising regulation of biological mechanisms’. Compared to proteins, the size of RNA is small and easy to fit into any catalytic centre. When the size of ancient genomes were not large enough to accommodate as many protein-coding genes for the requirement of biochemical pathways, it is thought that many RNAs were continuously used as regulatory tools since small size and structural flexibility of RNA molecules were advantageous for the evolution of new regulatory functions at relatively low cost. Genetic studies of modern organisms have shown that ncRNAs have a positive influence on evolution of genetic regulation in eukaryotes (Brown et al. 2003; Fedorova and Fedorov 2003; Pozzoli et al. 2007). Therefore, studying modern RNA evolution can lead to a better understanding of the versatile features of ncRNAs and their close relations to the evolution of modern lives.

In general, it appears that RNA evolution has helped the formation of coding system and accelerated the evolution of biochemical pathways. However, there seems to be a “missing age” between ancient and modern RNA evolution, when active information gain and loss took part in the shaping of current ncRNA landscape. The effect has been noticed that, while some ncRNA functions may be traced back to ancestral states, many modern ncRNAs appear to have emerged *de novo* within eukaryotes, thus missing a direct link to any possible ancestral features. This situation is common in higher eukaryotes such as mammals. We can envisage different forms of RNA continuity back to an RNA-world. The first would be direct continuity, and would probably include the ribosome and RNaseP. However intermediate forms could be when the

protein processing machinery for small RNAs were required during eukaryotic evolution and RNAs with new functions could be recruited, such as the Xist RNA in X-chromosome inactivation in mammals. Certainly it is not a favoured observation that the evolution of RNAs appears to be partly discontinuous. But considering the complex evolutionary passage from single-cellular organisms to developmental complex higher eukaryotes, numerous evolutionary incidents could happen so that one function could find a new application with small modifications. Being highly adaptable, RNAs could have gone through multiple steps of adaptation and lost the original identity originally formed in the lost RNA-world.

1.2 Relics from the RNA-world

1.2.1 The history of universal house-keeping RNAs

The “DNA-RNA-protein” coding system is a universal feature of all living organisms. rRNAs and tRNAs are two types of ncRNA that play key role in this system and are absolutely required for any organism, hence they can be considered as “house-keeping RNAs”. rRNAs and tRNAs from all species share conserved sequences and structures, and therefore are likely to be the oldest ncRNAs according to current knowledge of evolution.

Throughout the early 1970s, much evidence suggested rRNA played functional roles in translation, from the studies of site-directed modification of nucleic acid residues in rRNAs (Bowman et al. 1971; Senior and Holland 1971; Helser et al. 1972; Noller and Chaires 1972; Lai et al. 1973). The discovery of catalytic RNA in 1982 (Kruger et al. 1982) strengthened the possibility of rRNA taking part in the active catalytic mechanism of protein synthesis. It remained a question until the structure of ribosome was explored to 3.2 Å resolution (Ban et al. 2000), and it became clear that rRNA plays fundamental roles in at least two basic mechanisms of translation: tRNA binding and catalysis of peptide bond formation. The atomic structure of the peptidyl transferase active site shows that the RNA that surrounds the substrate

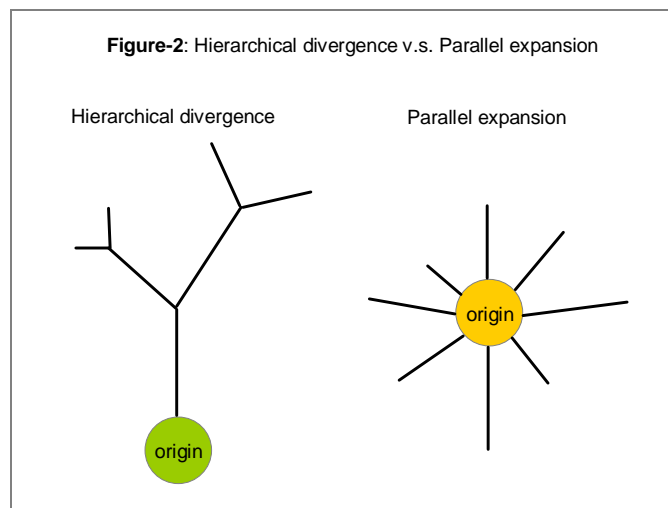
analogues is closely packed (like the active site of a protein enzyme), and the nucleotides in contact with an inhibitor molecule are >95% conserved in all three kingdoms of life (Nissen et al. 2000). Therefore it is clear that the ribosome is a ribozyme, and in its earliest form proteins might only have had structural/stabilizing roles.

Ribosomal RNAs have well defined secondary structures with extensive three dimensional helical structures formed from double-stranded segments, thus they are stereochemically constrained from many random mutations and recombination. According to the theory of the origin of genetic information (Eigen 1993), assembling information from poly-nucleotides was random and the initial sequence pool consisted of a vast number of similar but not identical sequences. Mutation and natural selection then drove the functional competent sequences through an evolutionary bottleneck. Study of the mutation rates of modern genomes found that DNA-based genomes are relatively stable compared to RNA-based genomes such as in lytic viruses and retroviruses (Drake 1999). The overall complex and mature structure of rRNA suggests that ancestral rRNA has reached an evolutionarily competent level before the emergence of first DNA-based organism. The length of modern rRNA does not permit a spontaneous origin from genomic recombination; instead rRNA may only evolved from naturally selected functional ancestral RNA motifs before LUCA.

tRNAs are another type of house-keeping RNA likely to have emerged from the RNA-world (Maizels and Weiner 1994). The cloverleaf shaped structure of tRNA is highly conserved in all living organisms. It has been discovered that the energy used to drive translocation during translation is stored in the tRNA-mRNA-ribosome complex after peptide-bond formation, thus translocation is a function inherent to the ribosome (Fredrick and Noller 2003). Since tRNA is much smaller than rRNA, it is likely that tRNA evolved to a modern tRNA-like mature form far earlier than rRNA. The fact that tRNA had been selected as a basal component of the translation system does not assume that tRNA evolved for this purpose. tRNAs from various species do not share significant sequence similarity, and the transcriptional patterns of tRNAs vary largely. Group I and

Group II self-splicing introns are both found in bacterial tRNA genes (Reinhold-Hurek and Shub 1992). A special tRNA-type intron is found in eukaryotes and archaea (Phizicky and Greer 1993). And in isolated cases such as the deep-branching *Nanoarchaeum equitans*, the tRNA genes exist in two halves, transcribed separately and ligated by an unknown mechanism (Randau et al. 2005). Therefore the ways by which ancestral tRNAs embedded into early genomes were likely to be highly variable.

Studies on the origin of tRNA lead to the hypothesis that tRNAs were formed by ligation of primordial hairpin RNAs (Di Giulio 1999; Tanaka and Kikuchi 2000). It has been proposed that the tRNA family started from a sequence distribution of neutral mutants with a subsequent parallel expansion rather than hierarchically successive divergence, and historically this infers to the period

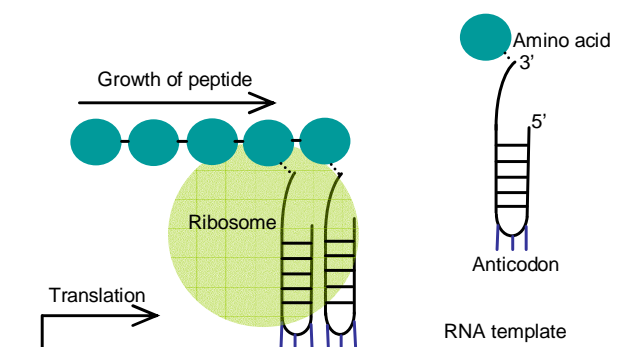


when the genetic information was not yet integrated into one replicable unit like in modern genomes (Eigen and Winkler-Oswatitsch 1981a). However, to retain the conserved structures against the highly dynamic nature of ancient evolution, they may have had a function which was retained as a basic constituent of the evolution apparatus during the first stages of evolution (Eigen and Winkler-Oswatitsch 1981b).

The anticodon structure of modern tRNA strongly suggests that the “tri-nucleotide” coding system retained its ancient form at least during peptide synthesis in the late RNA-world. In the theoretical model of primordial aminoacylation described by Lehmann (Lehmann 2002) as shown in Figure-3,

the translated region of the RNA template is composed of tri-nucleotide repeats, which are recognized by primordial tRNAs containing the tri-nucleotide anticodons and amino-acid docking sites, so that aminoacylation proceeds on the ends of the tRNA-like molecules assisted by the ancient ribosome. This model applies the anticodons to evolve towards “RNY” triplets, which are a later form of codons of “GNC” triplets (Eigen and Schuster 1977), and encourage the evolution of the first peptide replicator. This model was reinvestigated by computational approach that non-trinucleotide anticodons disappeared during simulation (Lehmann et al. 2004). It is likely that the origin of the first peptide replicator and the prevalence of “RNY” triplet were crucial steps during the transition from the RNA World to DNA/Protein World, and thus the origin of tRNA-like molecules is the most important step during this period. Figure-3 shows a proposed model of ancient translation.

Figure-3: Lehman model of ancient translation directed by primordial tRNAs



Translation starts where the RNA template begin to have successive “RNY” triplets, which enable the primordial tRNAs to bind one after another, thus the peptide grows at the 3’ ends of the primordial tRNAs under the assistance of an ancient ribosome. (Lehmann, 2002)

1.2.2 The antiquity of catalytic RNAs

Perhaps most ncRNAs seen today cannot be traced directly back to ancestral RNAs. However, relics from the RNA World can still be found. Although most biological catalysis is performed by proteins, eight classes of natural ribozymes have been discovered to take part in several fundamental biochemical reactions such as nucleotide cleavage and ligation (summarized in Table-1). Unlike many other modern ncRNAs, which function in protein-RNA complexes, these ribozymes are able to catalyse as RNAs.

Table-1: Summary of natural ribozymes

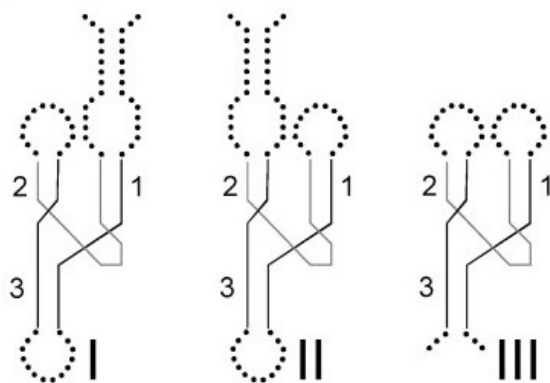
Classification	Functions	Expression in host organisms
Hammerhead ribozyme	Self-cleavage of rolling circle replication products into genome-length units	Satellite RNA in plant viruses
Hepatitis delta virus (HDV) ribozyme	Metal ion independent self-cleavage of rolling circle replication products at high temperature	<i>Hepatitis delta</i> viruses
Hairpin ribozyme	Reversible self-cleavage of rolling circle replication products	Satellite RNA in plant viruses
<i>Neurospora Varkud</i> satellite (VS) ribozyme	Self-cleavage to generate monomeric VS RNA	Satellite RNA in <i>Neurospora Varkud</i>
Group I self-splicing intron	Mg ²⁺ dependent self-splicing of mRNA, tRNA and rRNA <i>in vitro</i> , but protein machinery is required <i>in vivo</i> .	rRNA, tRNA and mRNA of organelles in fungi, plants and protists; tRNA mRNA of bacteria/bacteriophage; rRNA of protists, fungi and isolated cases in animal mitochondria (e.g. Sea anemone)
Group II self-splicing intron	Self-splicing occurs <i>in vitro</i> but protein machinery is required for splicing <i>in vivo</i> .	rRNA, tRNA and mRNA of organelles of fungi, plants, protists and mRNA of bacteria.
Ribonuclease P and MRP	Catalyse endonucleotide cleavage reactions on pre-tRNA or pre-rRNA.	RNaseP is present in all three kingdoms of life, while MRP is only present in eukaryotes.
Riboswitch	Binding to small metabolites in the absence of protein factors. Some riboswtiches can directly cleave the associated mRNA	mRNA of both gram-positive and gram-negative bacteria, plants and fungi

(Diener 1989; Saville and Collins 1991; Schmitt et al. 1993; Pley et al. 1994a; Scott et al. 1995; Earnshaw et al. 1997; Ferre-D'Amare et al. 1998; Pannucci et al. 1999; Doherty and Doudna 2000; Ikawa et al. 2000; Bonen and Vogel 2001; Blount and Uhlenbeck 2002; Hartmann and Hartmann 2003; Vitreschak et al. 2004; Haugen et al. 2005)

1.2.2.1 Small catalytic RNAs

The hammerhead, hairpin, HDV and NVS ribozymes are small RNA molecules (50-150nt) which can self-cleave independently of protein cofactors both *in vitro* and *in vivo* (Fedor and Williamson 2005). The hammerhead ribozymes are a type of simplest ribozyme. Studies using *in vitro* selection techniques, starting from a random pool of 60nt long RNA molecules, obtained predominantly hammerhead RNAs after 16 rounds of replication and selection (Salehi-Ashtiani and Szostak 2001). As repeated evolutionary selection tends to achieve the most common solution to a biological problem, the hammerhead ribozyme which has the simplest structure is likely to represent the ancestral form of these ribozymes originated from the RNA-world.

After the first crystal structure of hammerhead appeared (Pley et al. 1994b; Scott et al. 1995), it soon became clear that functional groups involved in cleavage reactions either protrude into the solvent, or interact with other groups with less importance (McKay 1996). A number of studies further analyzed the inconsistency between observed structures and mechanistic understanding of hammerhead (Horton et al. 1998; Kisseleva et al. 2005; Vogt et al. 2006), and all data implies that the active form of hammerhead is not the ground-state conformation observed by crystallography. Figure-4 shows three typical stem-loops structures of hammerhead for different sequences. Unexpectedly, it has been shown that peripheral structural elements in stem I and stem II, which share little sequence homology across various hammerhead ribozymes, greatly contributed to catalysis under physiological conditions (De la Pena et al. 2003; Khvorova et al. 2003). A more recent study using a full-length “tertiary-stabilized” hammerhead has explained the previously irreconcilable sets of experimental data by showing an intricate network of interactions between the loop regions of stems I and II, so that the functional nucleotides could closely approach each other (Martick and Scott 2006). Despite lacking of homology between different small self-cleaving RNAs, similar impact of long-range interaction between peripheral domains in ribozyme catalysis has also been observed for the HDV ribozyme (Tinsley and Walter 2007).

Figure-4: Schematic diagrams of hammerhead stem-loop structures

Types I, II and III hammerhead ribozymes are shown here. Peripheral regions shown as dotted lines contain the tertiary stabilizing motifs and can be of arbitrary sizes. Stems 1, 2 and 3 are indicated (Burke and Greathouse 2005).

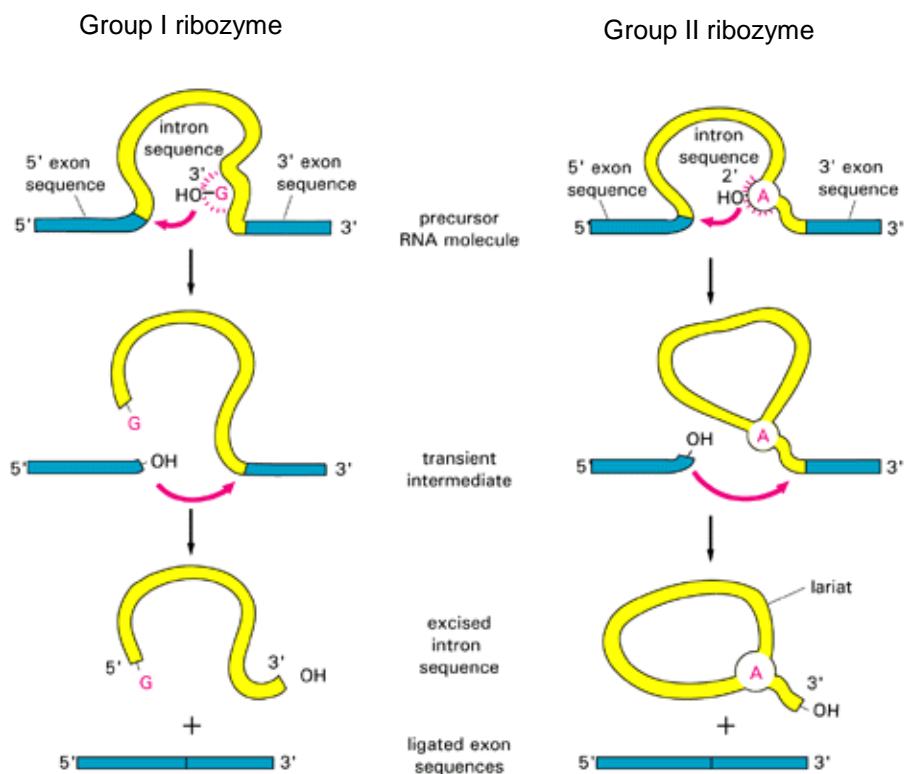
The hammerhead may be called a simple ribozyme due to its small size, but the mechanism of its catalysis is by no means simple. Under standard *in vitro* assaying conditions with elevated divalent metal ions (10mM Mg^{2+}), a minimum hammerhead enzyme can catalyse self-cleavage at a high rate. However an active-site metal ion was not evident in the native hammerhead RNA from the blood-fluke *Schistosoma* (Martick and Scott 2006), which exhibited a single-step folding process at low concentration of Mg^{2+} (Penedo et al. 2004). In contrast, the folding behaviour of *Schistosoma* hammerhead in Na^+ solution showed a two step folding process similar to the minimal hammerhead lacking the tertiary loop-loop interactions (Penedo et al. 2004). While *Schistosoma* hammerhead requires inner-sphere interactions with divalent ions, a recent work has investigated the ability of an artificial hammerhead ribozyme to self-cleave in the presence of either monovalent, poly-amine or exchange-inert trivalent cations, and results indicate this ribozyme can use an alternative folding pathway in the presence of non-divalent cations (Roychowdhury-Saha and Burke 2007). In addition, monovalent cation promoted catalysis of HDV ribozyme (Ke et al. 2007) and hairpin ribozyme (Young et al. 1997; Murray et al. 1998) has also been reported.

It has become clear from recent studies of small self-cleaving ribozymes that tertiary structural interactions are crucial for the activity of ribozymes, and peripheral stem-loops are important for the stabilization of the active structure. The divergence of peripheral sequences, the different folding processes and varying requirements of cations indicate a high tolerance of structural changes during the evolution of these ribozymes, which are likely to have evolved from an ancestral minimal hammerhead ribozyme. The strong correlations between structure and function during evolution of ribozymes (Hoogstraten and Sumita 2007) are also observed in other major classes of natural ribozymes.

1.2.2.2 Self-splicing introns

The variable dependence of cations for competent catalysis of small self-cleaving ribozymes indicates direct participation of RNA groups in the chemistry of catalysis. The observations from studies of hammerhead and hairpin ribozymes (Pyle 1993) contradict the early concept that RNA was relatively inert in catalytic terms and serving mainly to correctly position catalytically-active metal ions. Unlike small ribozymes, the relatively large self-splicing ribozymes from all studies to date have been shown to be obligate metalloenzymes. There has been no evidence for RNA groups participating in the acid-base catalysis.

There are two classes of self-splicing ribozymes, named Group I and Group II introns. Group I introns, including the first discovered *Tetrahymena* ribozyme, are widely distributed in protist nuclear rRNA genes, fungal mitochondria, bacteria and bacteriophages, and they are self-spliced by a distinctive two-step mechanism relying on an external guanosine as cofactor (Haugen et al. 2005). Group II introns are found in bacterial and organellar genomes, and are spliced through a different mechanism, instead of a guanosine, the 2'-OH group within the intron acts as nucleophile (Bonen and Vogel 2001). Both groups of ribozymes can act as pure RNA catalyst *in vitro*, but usually require protein cofactors *in vivo*. Figure-5 compares the splicing mechanisms of Group I and II introns.

Figure-5: Splicing mechanisms of Group I and Group II ribozymes

The group I intron binds a free guanosine (G) to a specific site to initiate splicing, while the group II intron uses a specially reactive adenosine (A) in the intron sequence itself for the same purpose. Both reactions are normally aided by proteins that speed up the reaction, but the catalysis is nevertheless mediated by the RNA in the intron sequence. This figure is from (Alberts et al. 2002).

Group I introns were the first example of RNA catalysis discovered (Kruger et al. 1982). Various studies have shown the critical role of a precise core of multiple divalent cations at the active site. Substitution experiments with sulphur replacing phosphate oxygen, thus disrupting Mg^{2+} binding, followed by metal-rescue experiments have provided a clear view of the constellation of cations at the transition state of *Tetrahymena* ribozyme self-splicing reaction (Shan et al. 1999; Shan et al. 2001), and identified functional binding sites for the catalytic divalent ions within the intronic active site (Szewczak et al. 2002; Hougland et al. 2005). These data provides a concise model of the chemistry of transition state, and contributes to the analysis of atomic-level structures of the ribozyme.

High resolution crystallography of the active Group I ribozyme has identified two divalent metal ions in the active site (Adams et al. 2004). The crystal structures of Group I ribozymes from *Tetrahymena* (Guo et al. 2004), *Twort* (Golden et al. 2005) and *Azoarcus* (Stahley et al. 2007) support a conserved core structure stabilized by peripheral elements which are variable between organisms. This feature of peripheral structures stabilizing a conserved core structure is similar to the structure-function relationships seen in small ribozymes. Similarly, Group II introns also require the interaction of peripheral stem-loops to stabilize the tertiary structure (Fedorova et al. 2003) and two divalent metal ions in the reaction centre (Gordon et al. 2007).

Despite well studied biochemistry of self-splicing ribozymes, the evolution of both Group I and Group II introns still remains a topic of extensive discussion. Group I introns have sporadic distribution on the tree of life. Most of them are found in the organelles of fungi, plants and red algae with the rest found in bacteria and isolated cases in animals (Haugen et al. 2005). Viruses and phage can possess Group I introns as well (Nishida et al. 1998; Sandegren and Sjoberg 2004). The distribution of Group II introns is very similar to that of the Group I introns, with the majority found in organellar genomes of plants, fungi and algae, the minority found in bacteria and archaea, but none in animals so far (Bonen and Vogel 2001; Toro 2003). Both groups of ribozymes are mobile genetic elements (Goddard and Burt 1999; Cousineau et al. 2000), both have gone through extensive horizontal gene transfer (Belfort and Roberts 1997) or act similarly as retrotransposons (Bonen and Vogel 2001). The enrichment of self-splicing ribozymes in organellar genomes suggests vertical inheritance of the ribozymes in cyanobacterial ancestors of organelles, but their association with different types of genes (Belfort and Roberts 1997; Bonen and Vogel 2001) and distinct mechanism of gene transfer suggest an early divergence of these ribozymes before the origin of cyanobacteria.

The autocatalytic feature of self-splicing ribozymes *in vitro* is an evidence of their possible early origin in the RNA World. *In vitro* experiment showed that converting a self-splicing Group I intron into a protein-binding ribonucleoprotein (RNP) complex only required small structural change in the

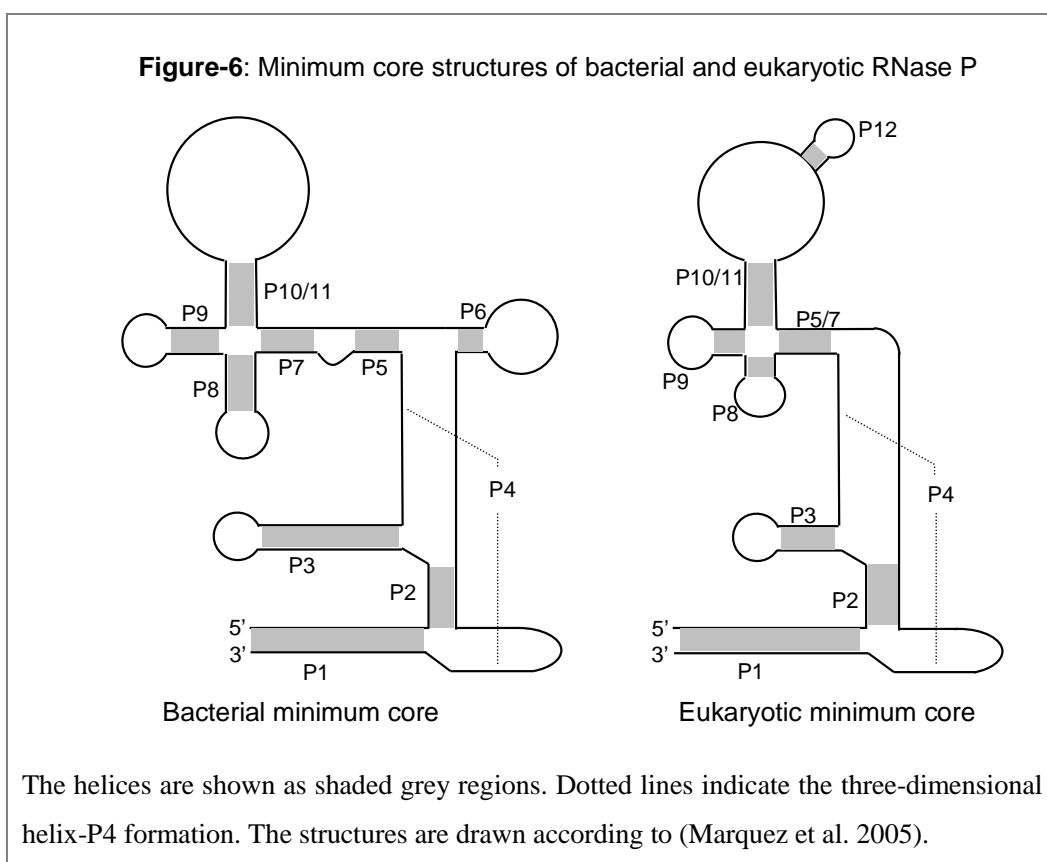
peripheral domains, and the active core structure was stabilized by the protein cofactor (Garcia and Weeks 2003). This suggests that evolution from catalytic RNAs to catalytic RNPs has low cost in RNA molecules, because the later incorporated proteins have functions mainly in protecting and stabilizing the RNA cores in more complex cellular environments. The same role of protein as structural support is also seen in the classical ribozyme RNase P, which has universal existence across three kingdoms of life.

1.2.2.3 RNase P and MRP

The ribonucleoprotein enzyme RNase P catalyses the endonucleolytic reaction at 5'- end of primary tRNA transcripts to produce mature tRNAs (Frank and Pace 1998). To date, RNase P is the only ribozyme, besides the ribosome, required in all species of prokaryotes and eukaryotes. Eukaryotes have another RNase P related endonuclease MRP, which processes the precursor of rRNA. The bacterial RNase P ribozyme contains one catalytic RNA subunit of 350-450 nucleotides and a single small protein (Brown 1998). Archaeal and eukaryotic RNase P, and eukaryotic MRP, contain an RNA subunit of the similar size, however, they have multiple protein subunits. The RNAs in RNase P and MRP from different organisms share the same structural architecture around the catalytic core (see Figure-6) although the overall sequence and structure differ significantly (Marquez et al. 2005).

Similar to other large ribozymes, the activity of RNase P is dependent on divalent metal ions. The RNase-P RNA is associated with a large number of Mg^{2+} (Beebe et al. 1996). The folding of individual domains is a cooperative process in the presence of Mg^{2+} (Kent et al. 2000), and it has been suggested that changes in intracellular concentrations of divalent ions (e.g. Mg^{2+} and Ca^{2+}) regulates the activity of RNase P (Brannvall and Kirsebom 2001). Two groups of metal ions are likely to be involved with RNase P catalysed cleavage: one promotes the folding and the other aids catalysis (Kirsebom 2007).

Both prokaryotic and eukaryotic RNase P RNAs are active as protein-free ribozymes *in vitro*, although eukaryotic RNase P has a much lower rate of cleavage (Kikovska et al. 2007). Sequence alignment of 30 eukaryotic RNase-P RNAs showed <80% similarity among them, the non-homologous regions are eukaryote-specific and highly variable in both sequence and length (Marquez et al. 2005). Although the sequence homology is low, the highly similar core structure (Figure-6) between prokaryotic and eukaryotic RNase P RNAs suggests an early common origin of this RNA structure, and the catalytic core has remain conserved by evolutionary constraint while the peripheral domains evolve more freely.



In contrast to the RNA subunit, RNase-P proteins have high degrees of diversity between the kingdoms of bacteria, archaea and eukaryotes. The single bacterial RNase-P protein has no detectable homologues in eukaryotes, and most archaeal RNase P enzymes only contain 4 orthologs with over 10 eukaryotic RNase-P proteins (Hartmann and Hartmann 2003). The consensus core of RNA subunits versus the divergence of protein subunits suggests that the RNase P may have evolved from an ancestral ribozyme before the

divergence of bacteria, archaea and eukaryotes. This possibility is reinforced by the fact that tRNAs is likely to have the most ancient origin, therefore RNase P which processes tRNAs may have an ancient origin too.

Unlike RNase P, RNase MRP is only found in eukaryotes. It was first identified as an RNase that cleaves RNA primers for the initiation of mitochondrial DNA replication in mouse and yeast, and later found to be mainly involved in rRNA processing in the nucleus (Morrissey and Tollervey 1995). The overall structure of the RNA subunit in RNase MRP is similar to that of RNase P, and the MRP enzyme shares a number of protein subunits with RNase P (Chamberlain et al. 1998). MRP RNAs from different eukaryotes show a conserved core structure, but the peripheral domains differ significantly among different species (Woodhams et al. 2007). An early duplication event could have separated the functions of these two enzymes, and the presence of MRP RNA in a number of deep-branching eukaryotes suggests that this duplication event likely happened during early eukaryotic evolution (Woodhams et al. 2007)

The evolution of RNase P and MRP raises the interesting question of how an early RNA ribozyme gradually recruited proteins to adapt to the new cellular environment. While the conserved secondary structures of the RNA subunits in RNase P and MRP correspond to the conserved functional requirements, changing cellular compartmentalization may exert great influence on the variations such as difference in protein association. It has been proposed that RNase P enzyme had persisted as an RNA-protein ribozyme for a certain time during an evolutionary period where the Protein World was already dominating and a cellular compartmentalization already existed at that stage (Hartmann and Hartmann 2003). As a consequence, protein binding is likely to lead to sequence change of the ancestral RNA subunit, and change in sequence can lead to association with new protein partners. Therefore, changing in sequence and associating with increasing protein subunits may be mutually advantageous before optimal catalytic potential has been achieved in complex organisms.

In all, the catalytic ability of RNAs has persisted a long time since its origin from the RNA World. Ribozymes of different level of complexity share similar features such as the requirements of metal ions and important structure-function relationships. It has been seen from various ribozymes that folding of RNA sequences is the key determinant of function. Therefore functionally ncRNAs are like protein enzymes such that the specific folding and interactions between sub-domains of an RNA sequence can promote interactions with substrates. This feature is also seen in ncRNAs that function as signal transduction molecules regulating gene expression by binding to small ligands.

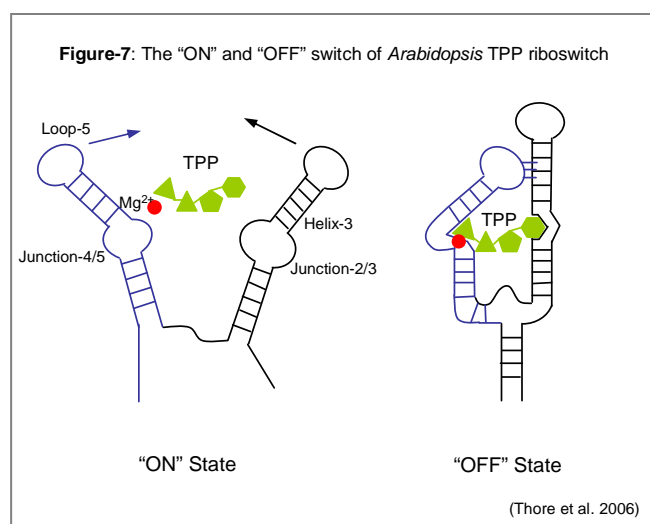
1.2.3 Ancient RNA regulators – Riboswitches

A type of RNA sequence within mRNAs can fold into structural domains and regulate adjacent gene expression by binding to small molecules such as metabolites. These RNA structures are called riboswitches, which are broadly distributed in bacteria (Winkler and Breaker 2005) and use a number of different mechanisms to regulate gene expression: such as preventing ribosome binding (Winkler et al. 2002), formation of hairpin structures which terminate transcription (Mandal and Breaker 2004), or acting as ribozymes for direct cleavage of mRNAs (Doudna and Lorsch 2005).

Riboswitches have been well adapted into various metabolic pathways in nature. Studies of conserved regulons in bacterial genomes often lead to discoveries of conserved riboswitches. *Vice versa*, comparative genomic studies using riboswitches can lead to discoveries of new metabolic pathways too (Rodionov et al. 2002). Riboswitches are suggested to have an ancient origin (Vitreschak et al. 2004). NMR studies showed that binding of guanine or adenine to the 5'-UTR of the G-switch or A-switch RNA in *B. subtilis* is specified by intermolecular Watson-Crick-type base pair between the ligand and the riboswitch (Noeske et al. 2005). This evidence of an extremely small ligand-switch interface supports a very flexible evolutionary passage of

riboswitches, because only small changes in sequence or structure are required for changes of ligand.

Riboswitches have also been discovered in eukaryotes. Production of thiamine pyrophosphate (TPP) – an essential cofactor in bacteria, archaea and eukaryotes is tightly regulated by TPP-binding riboswitches, which share the same conserved structure and undergo similar conformational changes during gene regulation (Sudarsan et al. 2003). Fungal and plants' TPP-binding riboswitches are found either within introns or the 5'- or 3'- untranslated regions of the regulated genes (Kubodera et al. 2003; Sudarsan et al. 2003). They have recently been found to regulate some alternative-splicing (Cheah et al. 2007).



Crystal structure of *Arabidopsis* TPP-binding riboswitch (Thore et al. 2006) revealed the TPP-induced conformational change which determined the "on" or "off" state of mRNA translation (Figure-7). The highly conserved TPP structure consists of

five helices. Upon TPP binding with the conserved sequences at helix junctions 4/5 and 2/3, the loop 5 interacts with helix 3 and brings the two parallel helices together forming the "off" structure.

Recent study of TPP-binding riboswitches in the fungus *Neurospora crassa* has found that this riboswitch can regulate gene expression through alternative mRNA splicing. Binding of TPP induces structural rearrangement which either blocks or reveals key intron sequences and therefore determines the different mRNAs resulted from alternative splicing. (Cheah et al. 2007). The role of alternative splicing in eukaryotic gene control has become increasingly

apparent (Moroy and Heyd 2007; Sorek 2007), and the involvement of riboswitches in regulating alternative splicing is rather certain as more evidence is being discovered (Borsuk et al. 2007). Given the extreme flexibility of RNA folding, it is likely that riboswitches are widely used for gene control in eukaryotes. Although similar riboswitches have not yet been seen in animals, an artificial riboswitch has been made to regulate splicing in mammalian cells (Kim et al. 2005).

1.3 Evolution of ncRNAs in eukaryotes

With advanced computational strategies for searching conserved ncRNA sequences from sequence genomes, a vast variety of ncRNAs have been found in eukaryotes (Griffiths-Jones et al. 2005; Mattick and Makunin 2006; Yazgan and Krebs 2007). High throughput biochemical methods for ncRNA identification also have revealed that eukaryotes possess many more regulatory ncRNAs than prokaryotes, as seen in the Rfam database (Griffiths-Jones et al. 2005). The evolution of ncRNA in eukaryotes has reached the stage of modern diversification involved with the complex genetic control and development.

1.3.1 Introns and their roles in eukaryotic evolution

Most eukaryotic mRNAs include intron sequences which are spliced out by a large ribonucleoprotein complex – the spliceosome (Nilsen 2003) during a coupled transcription-translation cascade. The discovery of introns (Williamson 1977) soon provoked different theories on intron evolution. The “intron-late” theory suggests that introns are transposable elements inserted into previously un-split genes (Cavalier-Smith 1985). In contrast, the “exon theory of genes” suggests that introns allow shuffling of protein-coding sequences and increase complexity by recombination and supports the “intron-early” hypothesis (Gilbert 1978; Roy 2003) which suggests that introns are ancient border exons. In addition, the “intron-early” hypothesis suggests that some introns encoding small RNAs are older than the surrounding exons (Poole et al. 1998).

Spliceosomal introns are only found in eukaryotes, and they are spliced with the same general mechanism as Group II introns (Jurica and Moore 2003). It was once suggested that the self-splicing introns, tRNA introns and spliceosomal introns have unequal antiquity (Cavalier-Smith 1991). However, the completely different protein apparatuses associated with the self-splicing introns and spliceosomal RNAs suggests an early divergence of these RNAs, possibly even before the RNA-protein interactions were evolved. Several studies have uncovered spliceosomal introns in phylogenetically basal unicellular amitochondriate parasitic eukaryotes (Nixon et al. 2002; Russell et al. 2005; Vanacova et al. 2005), which were thought to be ancient eukaryotes before spliceosomal introns originated. However they are now classified as a group known as excavates, which generally have, a low intron density. However recently, remarkable high densities of ancient spliceosomal introns have been found in the oxymonad excavate *Streblomastix strix* (Slamovits and Keeling 2006). These results suggest that spliceosomal introns were likely to be abundant in ancestral eukaryotes, but subsequently lost in some lineages of excavates. Therefore, the spliceosomal introns in eukaryotes are probably as ancient as the self-splicing introns. Extensive biochemical studies have been done on the mechanism of splicing, and information on RNA folding and catalysis has provided insights into the evolution of introns and splicing.

Spliceosomal introns are constantly undergoing extensive loss and gain process (Roy and Gilbert 2006). Exonization of introns has been realized as a frequent process, which leads to the formation of thousands of new exons in vertebrates (Wang and Kirkness 2005; Wang et al. 2005; Alekseyenko et al. 2007; Krull et al. 2007). Alternative splicing of introns allows expression of new genes with recently integrated exons while old genes still remain intact. It has been shown that new exons are more frequently spliced out than old exons (Alekseyenko et al. 2007), therefore the effect of exon insertion is probably, on average, mildly deleterious or neutral (Xing and Lee 2006). In addition to protein-coding genes, alternatively spliced ncRNAs have also been discovered (Feng et al. 2006; Royo et al. 2007). Together, the dynamic pattern of intron evolution has remarkably increased the information content of genomes but yet maintains the original transcriptome.

1.3.2 Divergence of regulatory RNAs

Compartmentalization of cellular structures and divergence of cell types in eukaryotes has increased divergence of regulatory ncRNAs. Extensive studies have uncovered a large number of novel ncRNAs including small RNAs functioning in the nucleus and nucleolus as well as abundant large ncRNAs which function at various levels. These will be discussed next.

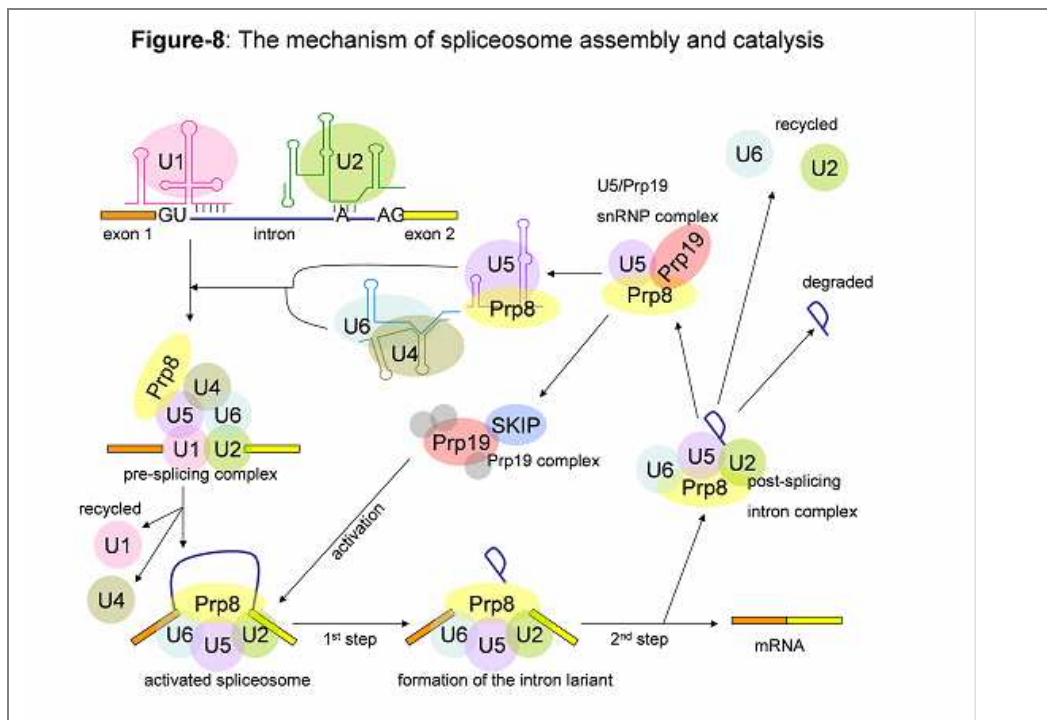
1.3.2.1 RNAs in the nucleus – small nuclear RNAs

A number of key processes in eukaryotes are compartmentalized in the nucleus, such as replication of the chromosome, transcription and RNA-editing. An increasing amount of novel ncRNAs are being found to be located in the nucleus, including small-nuclear RNAs (snRNAs) functioning in mRNA splicing, small nucleolar RNAs (snoRNAs) involved in RNA methylation and pseudouridylation, and numerous large ncRNAs over the size of 10kb, with functions yet to be characterised (Furuno et al. 2006).

Intron splicing is one of the major and well-studied processes which take place in the nucleus. Three types of spliceosome-mediated splicing exist in nature: major splicing, minor splicing, and trans-splicing. Almost all the eukaryotic nuclear pre-mRNAs are spliced by the major spliceosome, which appears highly conserved in eukaryotes. The major spliceosome is a large multi-subunit macromolecule formed by five uridine-rich small nuclear ribonucleoprotein particles (U-snRNPs): U1, U2, U4, U5, U6 and over two hundred non-snRNP splicing factors (Kramer 1996; Jurica and Moore 2003; Nilsen 2003).

Pre-mRNA introns are spliced in the same general way as Group II ribozymes (Figure-5). The splicing process (as shown in Figure-8) involves sequential association of UsnRNPs with the conserved 5'- and 3'- intron sites, and the formation of a catalytically competent spliceosome. Most introns contain canonical 5'- (GU) and 3'- (AG) sites, and a conserved branch point sequence followed by a polypyrimidine tract. In the early stage of spliceosome assembly, U1 binds at the 5'- splice site, and U2 snRNP binds loosely near the 3'- splice

site through complementary nucleotide sequences (Das et al. 2000). In the presence of ATP, the pre-assembled U5/U6.U4 tri-snRNP complex binds to the 5' splice site with assistance from the DExH-box RNA helicase family protein Prp8, and U2 snRNA firmly base pairs to the conserved adenine on the branch site of the intron (Maroney et al. 2000). Upon initial assembly, the spliceosome undergoes a series of structural rearrangements: the extensive base pairing between U4 and U6 snRNAs is unwound, followed by U6 base pairing with U5 and 5'- splice site, whereas U1 and U4 are released for recycling (Collins and Guthrie 2000).

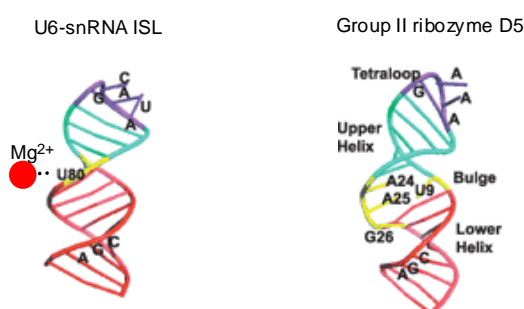


The overview of the spliceosome-mediated intron splicing is drawn according to (Grainger and Beggs 2005; Chen et al. 2006) with modifications. Much evidence points to the possibility that snRNPs are responsible for the catalytic activity of spliceosome (Valadkhan 2005). U1 snRNA functions early to specify the 5'-intron site through base pairing, but is released before the actual catalysis proceeds. Similarly, U4 snRNA is released before the first step of splicing reaction, and is dispensable in the catalytically active spliceosome (Yean and Lin 1991). U2, U6 and U5 snRNAs remain at the catalytic core throughout the splicing reaction. Three interactions between U2 and U6 were identified from studies of mammalian and yeast systems, and were shown to be required for splicing (Hausner et al. 1990; Datta and Weiner 1991; Madhani and Guthrie 1992; Sun and Manley 1995) U5 appears to act as a scaffold RNA to hold the two exon-intron junction sites at appropriate orientation by its invariant loop (Collins and Guthrie 2000).

Both structural and biochemical data suggest that the major spliceosome is a ribozyme. *In vitro* transcribed human U2 and U6 snRNAs can bind and position a small RNA fragment containing the sequence of the branch-point nucleotide in the presence of Mg^{2+} , and the RNA-RNA interacting complex is structurally very similar to the RNA complex during the first step of splicing (Valadkhan and Manley 2001). A ribozyme model for U6 snRNA has been suggested; U6 possesses an intramolecular stem-loop structure which is able to form in the active spliceosome when bound to U2 to form a C-A wobble base pair (Huppler et al. 2002). The sequence of U6 snRNA is highly conserved in phylogenetically diverse eukaryotes (Tani and Ohshima 1991), suggesting its early origin in the eukaryotes.

U6 snRNA shares extensive structural and functional similarities with the catalytic domain of a Group II ribozyme, which was shown to be able to substitute the U6 snRNA in an active spliceosome (Shukla and Padgett 2002). NMR structures of the catalytic motif of U6 snRNA and Group II ribozyme show three critical regions: the tetraloop, bulge and conserved “AGC” sequence (as shown in figure-9).

Figure-9: The catalytic motif of U6-snRNA and Group II ribozyme



This figure shows the comparison between the - U6-snRNA intramolecular stem-loop (ISL) and Domain-5 (D5) of Group II ribozyme, two RNA structures are highly similar. The hypothesis that U6 possesses catalytic

function came from the discovery that yeast U6 snRNA coordinates a Mg^{2+} ion, and *in vitro* splicing and mutagenesis using synthetic U6 snRNA demonstrated that two non-bridging oxygens of the uridine-80 residue of U6 was necessary for the catalytic activity of the spliceosome (Yean et al. 2000).

The presence of spliceosomal introns in all the eukaryotes known to date suggests that spliceosomal introns were present in the common ancestor of eukaryotes, and evolution of the spliceosome is likely to be the result of

selection upon its splicing function. The shared structural and functional features between spliceosomal RNAs and Group II ribozymes suggest a common origin of these ncRNAs. Distribution of self-splicing ribozymes shows little evidence that they had ever been transferred into the nuclear genomes of eukaryotes. Ancient spliceosomal introns in deep-branching eukaryotes (Nixon et al. 2002; Russell et al. 2005; Vanacova et al. 2005; Slamovits and Keeling 2006) suggest that spliceosomal introns and the spliceosome were present in the common ancestor of eukaryotes.

The evolution of eukaryotes involves formation of the nucleus, acquisition of organelles and eventually the origin of multi-cellularity. Each of these processes is accompanied by emergence of new genes to accomplish new functions, thus extensive genome rearrangements (e.g. gene duplication, recombination, translocation etc.) must have occurred more than once. The origin of spliceosomal introns was not certain unless the “intron-first” hypothesis (Poole et al. 1998) applies, however it was not random, for there are only two types of spliceosomal introns: the major spliceosomal introns, which usually start with “GT” and end with “AG”, and the U12 snRNP-dependent minor spliceosomal introns, which are present in rare mRNAs and many start with “AT” and end with “AC”. But some spliceosomal introns in basal eukaryotic lineages do not always obey the consensus sequences. For example, an extremely short spliceosomal intron in the mitochondrial ferredoxin gene in *Giardia intestinalis* has non-canonical 5'- splice site starting with “GC” (Nixon et al. 2002); and an intron discovered in *Trichomonas vaginalis* is the same type of the *Giardia* intron (Vanacova et al. 2005). Both organisms are evolutionarily deep-branching eukaryotes, with highly reduced cellular architecture, and exhibit archaea-like features. It is likely that the spliceosomal intron arose very early during evolution before the emergence of cellular life, but the rising of snRNAs and spliceosome happened during a transition period from eukaryotic ancestors to eukaryotes, when the cellular structures and functions were not yet fully evolved, but extensive gene rearrangements occurred and required many introns to be spliced. Therefore, study of the

splicing mechanism in basal eukaryotes could gain much insight into some important changes that occurred during the evolution of eukaryotes.

1.3.2.2 RNA editing and small nucleolar RNAs (snoRNAs)

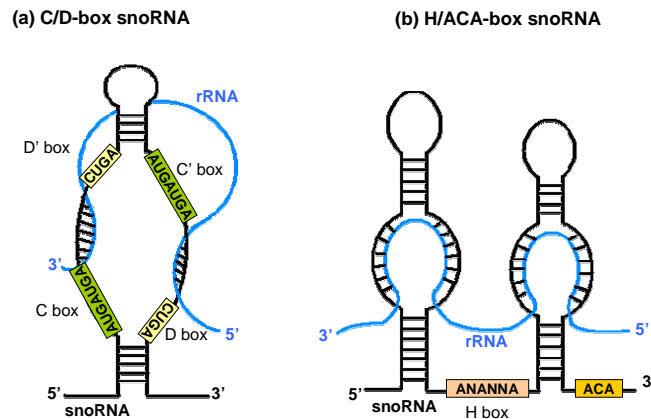
Both prokaryotic and eukaryotic rRNAs undergo extensive post-transcriptional modification. Three types of base modification occur during ribosome biogenesis, these include nucleotide base methylation, 2'-O-methylation of the hydroxyl groups of ribose residues, and pseudouridylation. Methylation of nucleotides is frequently seen in bacteria, but 2'-O-methylation and pseudouridylation are more frequent in eukaryotes (Maden 1990).

In eukaryotes, 2'-O-methylation and pseudouridylation are directed by two groups of small nucleolar RNAs: C/D box snoRNAs and H/ACA box snoRNAs respectively. In both cases, the snoRNA binds near the site of modification through antisense binding and guides the protein enzyme to the correct site. snoRNAs are widely distributed in eukaryotes, including animals, plants, yeasts, metazoans and protists. Some are also found in archaea. The two major classes of snoRNAs are characterized by consensus sequence motifs (Balakin et al. 1996). In addition, there is a third category of snoRNAs, named orphan snoRNAs, which do not have identified targets. Table-2 summarises the three types of snoRNAs. The common structures of C/D-box and H/ACA-box snoRNAs are shown in Figure-10.

Table-2: Types of snoRNAs

Class	Targets
C/D box snoRNAs	Methylation sites on rRNAs and snRNAs
H/ACA box snoRNAs	Pseudouridylation sites on rRNAs, some tRNAs and snRNAs, with possible sites on mRNAs as well
Orphan snoRNAs	No identified targets

Figure-10: Common structures of snoRNAs in eukaryotes



All C/D box snoRNAs contain conserved motifs termed C-box and D-box with antisense elements of 10 to 20 nucleotides immediately upstream to the D-box. The antisense elements are crucial for the specificity of snoRNAs (Cavaille et al. 1996) some snoRNAs also have additional less conserved C'-box and D'-box. H/ACA snoRNAs have a common secondary structure consisting of two parallel hairpins linked by a hinge. Two conserved motifs box H (ANANNA) and box ACA are located at the hinge and the 3' tail respectively (Ni et al. 1997). However, the antisense elements of box H/ACA snoRNAs are very short and bipartite.

Increasing number of novel snoRNAs are being identified in all kinds of eukaryotic species (Bachelierie et al. 2002; Mattick and Makunin 2006). Remarkably, a significant number of snoRNAs have also been found in archaea, where sno-like RNAs are typically shorter than eukaryotic snoRNAs, but contain well defined C, D, C' and D' motifs (Omer et al. 2002). A significant amount of snoRNAs are found in the deep-branching protist *Giardia intestinalis* (Yang et al. 2005; Luo et al. 2006; Chen et al. 2007), which does not appear to have a nucleolus structure (Niu et al. 1994) but does have nucleolus functions (Narcisi et al. 1998; Xin et al. 2005). The presence of snoRNAs in deep-branching eukaryotes and archaea, and the antisense mechanism of target RNA recognition are consistent with an ancient origin of these RNAs and a role in rRNA biogenesis. The presence of a type of guide RNAs in eukaryotes but not in prokaryotes could mean that the ancestral guide RNAs were likely to have evolved before the divergence of eukaryotes and prokaryotes but not inherited in the prokaryotic lineage.

The genomic organization of snoRNA genes gives clue to the evolutionary history of snoRNAs. Vertebrate snoRNAs are mostly encoded within intronic sequences, and in most cases, they are processed as debranched lariats spliced out from pre-mRNAs (Kiss 2006). In yeasts, most introns are either in monoforms, or arranged in polycistronic patterns, transcribed under the control of shared control elements, and subsequently cleaved by both endo- and exonucleases (Qu et al. 1999). A few yeast snoRNAs are encoded in introns, and locations of these snoRNAs within the host introns are important for snoRNA biogenesis (Vincenti et al. 2007). Polycistronic snoRNA genes are also common in plants and *Trypanosoma* (Leader et al. 1997). Archaeal sno-like RNAs are encoded on both strands of the DNA and distributed around the entire circular chromosome, and in most cases they are located within short spacer regions between protein-coding ORFs (Dennis et al. 2001).

The coordinated transcription of snoRNA in eukaryotes represents the feature of a regulation cascade: where the expression of snoRNAs can be finely adjusted with the expression of genes required for ribosome biogenesis and translation, hence the genomic location of snoRNAs in eukaryotes ensures that maturation of snoRNAs, modification of rRNAs and translation are tightly coupled.

Besides rRNA modification, snoRNAs also target modification to a wide range of other cellular RNAs. According to one hypothesis (Gerbi 1995), the snoRNAs were evolved as RNA chaperones to lock the structure of rRNAs to a “dead end”, and nucleotide modifications were made to indicate the completion of structural arrangement. Increasing numbers of snoRNA targeted RNAs have been found, these include tRNAs, snRNAs and mRNA targets. In model eukaryotic organisms, the major spliceosomal snRNAs are modified with 2'-O-methylation and pseudouridylation guided by snoRNAs. These modifications are often located in the region of intermolecular RNA-RNA interactions. For example, a conserved pseudouridine in yeast U2 snRNA induces a change in structure and stability of the branch-site sequence, thus facilitates in binding to the intron during splicing (Newby and Greenbaum 2001). The chaperone-like

feature of snoRNAs is also supported by the finding that box C/D snoRNPs share a common core structure with the spliceosomal U4 snRNP (Watkins et al. 2000), which functions as a chaperone and deliver the catalytic snRNPs to the centre of the spliceosome (Staley and Guthrie 1998). A number of novel organ-specific snoRNAs have been found in vertebrates (Cavaille et al. 2000).

The functional diversity of snoRNA beyond ribosomal RNA processing suggests an adaptive evolution of these small RNAs, in which case the non-ribosomal functions do not directly relate to the original functions of ancestors, from which the identity of functional motifs are derived. It is unclear at this stage whether the wide roles of snoRNAs are ancestral to eukaryotes, or expanded within eukaryotes. Part of the work of this thesis is to help understand the distribution of snoRNAs in eukaryotes.

During the evolution of eukaryotes, ncRNAs with modified functions may have been recruited to the increasing number of RNA-processing pathways, and thus the increasing genomic and cellular complexity brings about divergence of ncRNAs and their specific functions.

1.3.2.3 Enrichment of ncRNA in eukaryotes and systematic gene regulation

In addition to ribozymes, single-gene regulators and guide RNAs, some eukaryotic-specific ncRNAs are involved in tightly coupled gene regulatory pathways and chromatin modification. These ncRNAs have been discovered relatively recently but they have wide range of functions. In addition, large-scale cDNA cloning and genome tiling arrays have revealed that large proportions of eukaryotic genomes are transcribed and the number of ncRNAs has far exceeded previous thoughts. These are described below.

Eukaryotes use a variety of small ncRNAs to regulate gene expression. These small RNAs are parts of large ribonucleoprotein complexes that function in almost all aspects of gene control. In addition to the snRNAs and snoRNAs discussed above, a large class of small RNAs with size ranging from 21 to 25

nucleotides are found in most eukaryotes. These are microRNAs (miRNAs) and small-interfering RNAs (siRNAs). miRNAs are partially complementary to mRNAs and function by antisense-binding to the 3'-UTR regions of mRNAs and inhibit translation (Pasquinelli et al. 2005). siRNAs lead to degradation of complementary mRNAs (Morris 2005). si- and mi- RNAs are classified by their slightly different structures and precursors, but they function through a similar general mechanism referred to collectively as: RNA interference (RNAi). The miRNAs in animals are usually transcribed as long and often polycistronic precursors, and then processed into small hairpin intermediates, which are cleaved by a conserved protein Dicer (Bernstein et al. 2001) into mature miRNAs. miRNAs are double-stranded and functioning through activating the RNA induced silencing complex (RISC), upon activation, the dsRNAs are unwound and base-paired to complementary mRNA sequences, followed by mRNA degradation (Hammond et al. 2001). In animals, many miRNAs are encoded in gene clusters and homologous miRNAs have been found in different vertebrates (Lagos-Quintana et al. 2003). A study on human *mir17* clusters suggests a complex duplication and loss of miRNA genes from a *de novo* precursor similar to the vertebrate *Hox* gene cluster, and they have undergone positive selection as well as random drift (Tanzer and Stadler 2004). Identification of precursor-like miRNA genes in early diverged eukaryotic lineages will help to understand how the function of miRNA arose and generalize.

In addition to well studied small ncRNAs, there are also thousands of large ncRNAs whose functions are not yet well understood. Studies from classical eukaryotic models such as yeast and human have suggested that large ncRNAs have important roles in controlling gene expression. Large ncRNAs are generally transcribed as introns, antisense RNAs and also as separate RNAs.

It has been observed in different eukaryotes that antisense transcription is a common phenomenon. *S. cerevisiae* transcriptome studies showed that 85% of the genome could be expressed with many transcripts overlapping known genes in the antisense direction (David et al. 2006). Genome-wide screening of

Arabidopsis antisense transcripts revealed a large number of dsRNAs paired by *cis*- and *trans*- transcripts, which were likely to be involved with gene-regulatory networks (Wang et al. 2006). Abundant sterile* antisense transcripts are also seen in animals and protists (Elmendorf et al. 2001; Gunasekera et al. 2004; Katayama et al. 2005). Studies of ncRNA expression profiles have also shown pronounced developmental regulation of ncRNA in *C. elegans* (He et al. 2006) and *Drosophila* (Inagaki et al. 2005). Analysis of *C. elegans* ncRNAs showed that expression of many intronic RNAs was much higher than their host mRNAs, which indicated separate regulatory mechanisms for expression of these RNAs (He et al. 2006).

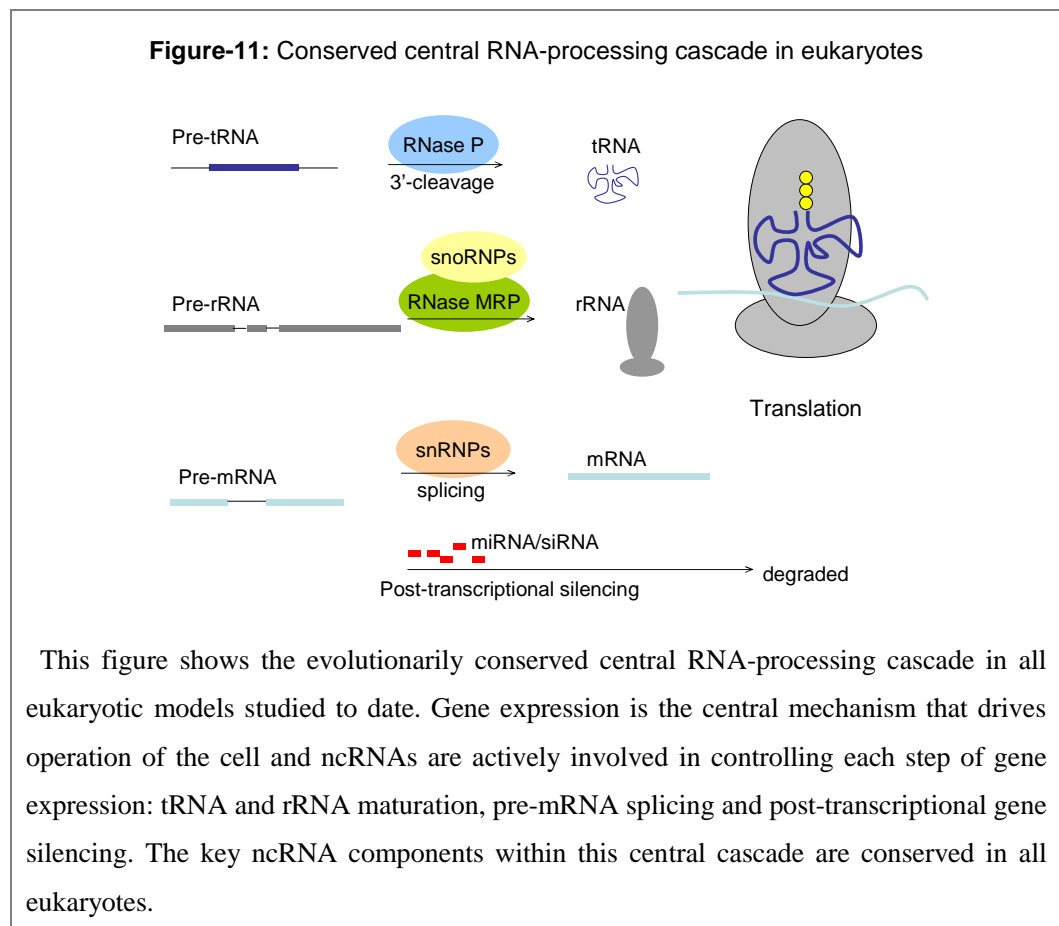
In contrast to small ncRNAs, large ncRNAs show weak evolutionary conservation. Several important large ncRNAs such as mammalian Xist, Tsix (Bernstein and Allis 2005), and *Drosophila* rox RNAs (Oh et al. 2003) have similar functions in epigenetic silencing and dosage compensation, but homologous RNAs have not been found in other eukaryotes. However large ncRNAs with different functions are widely seen, such as, the transcriptional interference seen for the poly-adenylated ncRNA SRG1 in regulation of yeast *SER3* gene (Martens et al. 2004), promoter inactivation of an ncRNA that binds to the DNA sequence of *DHFR* gene promoter, observed in human (Martianov et al. 2007) and epigenetic regulation in mammals through maintenance of chromosomal methylation patterns (Braidotti et al. 2004; Deng and Meller 2006).

1.4 The present and future of ncRNA evolution

The present knowledge of ncRNAs suggests a complex network of RNA-processing pathways, where the transcription and translation of genes are regulated by various ncRNAs, which themselves may also be regulated by other ncRNAs. The concept of RNA-processing cascade has been proposed (Woodhams et al. 2007). Figure-11 shows several major eukaryotic RNA-

* Sterile transcripts: RNA transcripts that are unable to code for proteins and do not have known functions either.

processing pathways that are evolutionarily conserved. The integrated network of ncRNA regulating gene expression at various levels indicates the importance of ncRNA as a fundamental component of life. The functions of ncRNA have expanded during the evolution of eukaryotes, but their basic features, such as the strong structure-function relationship and metal-ion facilitated catalysis, can still be traced back to their ancestors from the RNA World.



ncRNA evolution exhibits complex features of both hierarchical inheritance and parallel expansion. Hence the outcome is seen as the wide distribution of some ncRNAs across broad range of eukaryotic species, and narrow distribution of others. Many newly discovered novel ncRNAs, mostly large ones, have only been observed in higher eukaryotes. Their origin is uncertain and they appear to evolve *de novo*. However, the dynamic pattern of ncRNA expression, especially intronic and antisense expression, suggests that modern ncRNAs have gone through multiple rounds of natural selection associated with the evolution of new genes. Therefore, lack of functional homology

among these RNAs does not necessarily imply *de novo* evolution, as there can be many intermediate evolutionary stages before the current landscape of ncRNAs are formed.

The evolution of eukaryotes has continuously used the resource of ancient RNA motifs and reconstructed them into powerful regulatory tools. Studies of ncRNA evolution will help to understand some fundamental principles of molecular evolution behind the general evolution of cellular lives. With the basic RNA-processing infrastructure being thoroughly studied, the future of ncRNA study will provide more information of specific regulatory RNAs, which are not constitutively transcribed in the cell. Expression of these ncRNAs are widely associated with cell/tissue type, developmental stages and also controlled by epigenetic factors and environmental conditions. Functions of specific regulatory ncRNAs will provide new insights into the evolution of eukaryotes.

Finally, a number of ncRNA databases which annotate homologous ncRNAs from complete genomes are publicly available. The Rfam database (Griffiths-Jones et al. 2003; Griffiths-Jones et al. 2005) is the major resource of ncRNAs found in both prokaryotes and eukaryotes, with coverage of putative ncRNAs from over 200 complete genomes. The RNAdb database (Pang et al. 2005) is a mammalian ncRNA database, which contains almost 20,000 putative ncRNAs and 800 unique experimentally studied ncRNAs. The fRNAdb database (Kin et al. 2007) is a collection of annotated and un-annotated ncRNA sequences from H-inv database (Imanishi et al. 2004), NONCODE (Liu et al. 2005) and RNAdb, and provides an interface for sorting out functional ncRNA sequences. With the large amount of sequence and structural information of ncRNAs covered within these RNA databases and aided by advanced computational tools, searching for new ncRNAs has become more efficient. It is expected that large-scale computational prediction followed by experimental verification will be the major way for discovering new ncRNAs from sequenced genomes. This thesis contributes to our understanding of ncRNA evolution in eukaryotes by commentating ncRNAs in a deeply diverged protist, namely, *Giardia intestinalis*.

Chapter Two – Identification of novel non-protein-coding RNAs from *Giardia intestinalis*

Abstract

This chapter describes my study of the non-protein-coding cDNA library of *Giardia*. The library was constructed from RNA (sized 70 to 600nt) purified from total *Giardia* RNA. Sequencing and structural analysis have identified a number of typical eukaryotic small ncRNAs (including 3 C/D-box snoRNAs, 1 H/ACA-box snoRNAs and an unusual transcript of the RNase P RNA). However most of the ncRNAs identified from this library do not exhibit any conservation with known ncRNAs from other model organisms studied to date. Following computational predictions using a modified Snoscan programme we have identified 60 putative candidates of C/D-box snoRNAs from the *Giardia* genome. In addition, unusual self-cleaving dsRNAs are also found in *Giardia*. Results from this project suggest that the genetic information encoded in ncRNAs of *Giardia* may differ considerably from the standard context of ncRNAs in higher eukaryotes, though the key characteristic ncRNAs of eukaryotes such as snoRNAs and RNase P are present.

The studies included in this chapter have been published as one paper “Combined experimental and computational approach to identify non-protein-coding RNAs in the deep-branching eukaryote *Giardia intestinalis*”. Identification of novel ncRNAs from *Giardia* is based on a size-fractionated cDNA library. In the first stage (as detailed in this paper, 616 clones were sequenced and analysed. The features of *Giardia* snoRNAs are studied here and the expression of some ncRNAs in *Giardia* are discussed. After submitting the paper, another 576 clones were sequenced and analysed. Limited by the length of the paper, additional information including the molecular biology of *Giardia* and techniques used in identification and analysis of ncRNAs are given in Chapter-3, which presents the updated results of *Giardia* ncRNAs identified from the cDNA library (with a total of 1192 clones) and gives detailed discussion of the structures and expression of various types of ncRNAs from *Giardia*. The supplementary data are in Appendix-1.

With respect to my contribution, the *Giardia* was grown at Massey and I extracted and fractionated the RNA. It was then taken to Münster, where I

made, with assistance of locals, the cDNA library. This was brought back to Massey, and the sequencing and computational analysis done here.

Published online 22 June 2007

Nucleic Acids Research, 2007, Vol. 35, No. 14 4619–4628
doi:10.1093/nar/gkm474

Combined experimental and computational approach to identify non-protein-coding RNAs in the deep-branching eukaryote *Giardia intestinalis*

Xiaowei (Sylvia) Chen¹, Timofey S. Rozhdestvensky², Lesley J. Collins^{1,*}, Jürgen Schmitz² and David Penny¹

¹Allan Wilson Centre, IMBS, Massey University, Palmerston North, New Zealand and ²Institute of Experimental Pathology (ZMBE), University of Münster, Münster, Germany

Received February 8, 2007; Revised and Accepted May 29, 2007

ABSTRACT

Non-protein-coding RNAs represent a large proportion of transcribed sequences in eukaryotes. These RNAs often function in large RNA-protein complexes, which are catalysts in various RNA-processing pathways. As RNA processing has become an increasingly important area of research, numerous non-messenger RNAs have been uncovered in all the model eukaryotic organisms. However, knowledge on RNA processing in deep-branching eukaryotes is still limited. This study focuses on the identification of non-protein-coding RNAs from the diplomonad parasite *Giardia intestinalis*, showing that a combined experimental and computational search strategy is a fast method of screening reduced or compact genomes. The analysis of our *Giardia* cDNA library has uncovered 31 novel candidates, including C/D-box and H/ACA box snoRNAs, as well as an unusual transcript of RNase P, and double-stranded RNAs. Subsequent computational analysis has revealed additional putative C/D-box snoRNAs. Our results will lead towards a future understanding of RNA metabolism in the deep-branching eukaryote *Giardia*, as more ncRNAs are characterized.

INTRODUCTION

In the recent past, experimental and computational approaches have identified a vast variety of non-protein-coding RNAs (1), generally abbreviated as non-coding RNAs (ncRNAs), from both unicellular and multicellular eukaryotes. Many ncRNAs in modern eukaryotes function in RNA-protein complexes within which the RNAs may have direct regulatory roles at the reaction centres (1). The size of many ncRNAs is small compared with

protein-coding RNAs, and lack of sequence homology often results in difficulties of identifying ncRNAs in distant eukaryotes through purely biological or computational approaches. In this study, our combined experimental and computational approach has been successful in finding novel ncRNAs in the distant eukaryote *Giardia intestinalis*.

Eukaryotic genomes are rich in non-protein-coding sequences. Large-scale cDNA cloning studies have shown that a large proportion of mammalian RNA transcripts do not appear to encode proteins (2), and an increasing number of ncRNAs have been shown to be functional (1). The origin of ncRNA is likely to date back to the earliest events when life emerged on earth. The theory of the 'RNA-World' (3,4) suggests that self-replicating RNAs are older than protein or DNA. The versatile features of RNA molecules support this hypothesis: first, RNA stores information in the same way as DNA; second, single-stranded RNA molecules are highly flexible to form secondary or tertiary structures, like peptides, they can form enclosed reactive centres and catalyze biological reactions in liquid environment. However, modern natural ribozymes have limited catalytic abilities, as natural ribozymes only perform ligation and/or nucleic acid cleavage reactions. These reactions are normally not limited by the rate of the catalytic reaction (5). Therefore, it is assumed that most ancient ribozymes have gradually been replaced by protein enzymes (5).

On the other hand, the evolution of ncRNAs has been continuous, and functions of ncRNAs have been diversifying throughout the evolution of eukaryotes. Based on structural and functional definition, eukaryotes have several distinct classes of ncRNAs, which form complex RNA-processing networks. Table 1 shows that each type of RNA often participates in the modification of another type of RNA, and the whole network fits into the general RNA-processing cascade (6). It is necessary to provide some brief background on the types of ncRNAs here,

*To whom Correspondence should be addressed. Tel: +64 6 350 9099-7345; Fax: +64 6 350 5626; Email: l.j.collins@massey.ac.nz

© 2007 The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Table 1. A brief summary of ncRNAs in the RNA processing network of eukaryotes

Role	Type of ncRNA	Function
Transcriptional initiation	7SK snRNA (in mammals)	Inhibits transcription by binding to CDK/cyclin kinase complex
Intron splicing	U snRNAs	Function in the catalytic cores of major and minor spliceosomes involving in excision of introns
mRNA degradation	Micro RNAs	Guide the RNAi machinery to homologous mRNAs and trigger mRNA degradation
tRNA processing	RNase P	Involves in 5' end nuclease activity in pre-tRNA processing
rRNA processing	MRP RNA	Involves in the endonuclease activity in pre-rRNA processing
	C/D box snoRNAs	2'-O-methylation guide
	H/ACA box snoRNAs	Pseudouridylation guide

because in this study, we have characterized a number of different types of ncRNAs from *Giardia*.

Probably the best studied ncRNAs are uridine-rich spliceosomal snRNAs (U-snRNAs). They function in the catalytic centre of major and minor spliceosomes. The major spliceosome that splices the majority of eukaryotic introns, consists of 5 U-snRNAs (U1, U2, U4, U5 and U6) and over 200 proteins (7). The minor spliceosome is low-abundant machinery containing U11 and U12 snRNAs instead of U1 and U2, and splices a 'minor' (less frequent) class of introns (8). Both major and minor spliceosomes may be ancestral to eukaryotes because they have now been identified in animals, plants, fungi and recently some distantly related protists (9,10).

The small nucleolar RNAs (snoRNAs) are involved in rRNA biogenesis. An increasing number of novel snoRNAs have been widely identified and have been reviewed in detail (11–15). Based on their structural motifs, snoRNAs are divided into two classes: C/D-box 2'-O-methylation snoRNAs and H/ACA-box pseudouridylation snoRNAs. The snoRNAs bind near the sites of modification through antisense recognition, and guide protein enzymes to the sites of editing. In addition, the functions of snoRNAs can be extended to acting as general chaperones targeting other nuclear or cellular RNAs (16–18).

There are a number of larger ncRNAs (>300 nt) such as the RNase P and RNase MRP RNAs. To date, besides the ribosome, RNase P is the only ribozyme required in both eukaryotes and prokaryotes (19). Eukaryotes have another related ribonuclease, RNase MRP, which processes a specific site in the pre-rRNA which is not found in prokaryotes, however, it seems likely that it is present in all eukaryotic lineages (6). Structural analysis of RNase P RNAs from phylogenetically diverse eukaryotes reveal a very similar minimum core (20). The overall structure of the RNA subunit in RNase MRP is similar to that of RNase P (21), and also the MRP enzyme shares a number of proteins with RNase P (22).

The smallest ncRNAs are micro RNAs (miRNAs) with length ranging from 21–25 nt and function in a variety of gene silencing pathways (23). About 800 miRNAs from different animals and plants have been reported (24). miRNAs from animals are usually transcribed as long and often polycistronic precursors, and then processed into small hairpin intermediates, which are then cleaved by

a conserved protein Dicer (25) into mature miRNAs. The Dicer protein has been well studied for *Giardia* (26).

Recently, new experimental and bioinformatic approaches have identified a great number of novel ncRNA candidates from many organisms including: bacteria (27), animals (28–31) plants (32) and protists (33). The most widely used experimental method for identifying novel RNA candidates is based on size-selected cDNA libraries. Since most mRNAs have lengths greater than 500 nt, it is possible to isolate the majority of ncRNAs by size fractionation on a denaturing PAGE gel. Several methods are available to generate cDNAs from purified RNAs including the addition of poly(C)/poly(A) tail, and adaptor ligation at 5'-end and/or 3'-end, followed by reverse transcription, cloning and cycle sequencing (34). Here, we have constructed a cDNA library for ncRNAs from the deep-branching eukaryote *G. intestinalis*. *Giardia*, a parasitic diplomonad, is phylogenetically distant to all model eukaryotes (35,36). This unicellular organism has reduced mitochondria (mitosomes) and lacks hydrogenosomes (37). Two spliceosomal introns have been found (38,39), as well as several spliceosomal proteins (9) which strongly suggests that *Giardia* has a functional spliceosome. To date, several studies have identified 24 sno-like RNAs and the RNase P of *Giardia* (40–42). However, there is little systematic research reported for the RNomics of *Giardia*.

We have screened our *Giardia* cDNA library, resulting in 31 novel candidates, within which, three are possibly C/D-box snoRNAs, one is possibly an H/ACA box snoRNA, and one is a fragment of the RNase P RNA. A computational study using known *Giardia*'s C/D box snoRNAs has resulted in new putative snoRNAs. In addition, an extended transcript has been found for the RNase P RNA, and two unusual self-cleaving dsRNA candidates have been studied. Given its proposed basal position on the eukaryotic tree (36), *Giardia* is evolutionarily distant to all the eukaryotic species, and probably highly reduced. It is not surprising to see that there may be some different RNA processing components in this organism. Future comparison of RNA-processing between *Giardia* and other eukaryotes is very necessary in understanding the evolution of RNA metabolism in reduced organisms (43). RNA processing in *Giardia* is expected to have changed in both the RNA and protein components as a result of genome reduction (43) due to

the parasitic nature of this organism. Our study moves towards understanding differences in *Giardia* RNA-processing machinery from that of other eukaryotes which to date is largely confined to model, well-studied eukaryotes.

MATERIALS AND METHODS

Preparation of total RNA from *G. intestinalis* WB strain Trophozoites

Cells were collected from TY1-S-33 growth media at a concentration of 1.4×10^7 cells/ml by centrifugation (10 min, 2500 r.p.m., 4°C). Total RNA was prepared using Trizol reagent (Invitrogen) according to the protocol provided by the manufacturer.

cDNA library construction

Total RNA (10 µg) was run on an 8% denaturing PAGE gel (7M urea, 1× TBE buffer). RNA in the range of 70–600 nt was excised and eluted in 0.3M NaOAc overnight at 4°C. Subsequently, 10 µg RNA was treated with tobacco acid pyrophosphatase (Epicenter) for 1 h at 37°C, then C-tailed by poly-A polymerase (Invitrogen) for 2 h at 37°C.

A 5' DNA Sal-I adaptor (5'-CAACGCGTCGACTACGTGAGATTTGAGGTTTC-3') was then ligated to the RNA using T4 RNA ligase at 4°C overnight. First-strand cDNA synthesis was performed using the ThermoScript cDNA synthesis kit (Invitrogen) with Not-I primer (5'-GACTAGTTCTAGATCGCGAGCGGCCCGCCCGGGGGGGGGGGGG-3').

The RNA-cDNA mix was treated with RNase A and PCR amplified using Sal-I and Not-I primers using a Biometra thermocycler. The PCR product was then double digested by Sal-I and Not-I restriction enzymes and ligated into the pSPORT1 vector (Invitrogen), followed by transformation into *Escherichia coli* Top10 cells (Invitrogen).

Sequencing

E. coli cells were grown on LB agar plates (100 µg/ml Ampicilin) at 37°C overnight. Colonies were PCR amplified using the M13for and M13rev primers (Roche Taq polymerase):

M13for: 5'-CGCCAGGGTTTTCCAGTCACGAC-3'
M13rev: 5'-AGCGGATAACAATTTACACAGG-3'

PCR products were cleaned by SAP/EXO-I (GE Healthcare) treatment and cycle sequenced using BigDye Terminator version 3.1 and M13rev primer. The sequencing products were cleaned using CleanSeq (Agencourt) magnetic beads, and capillary sequenced on a capillary ABI3730 Genetic Analyzer (Applied Biosystems Inc.).

Computational analysis

The sequences were assembled using DNAMAN 5.2 and DNASTAR 5.0 packages, and were then blasted against the *Giardia* genomic database (<http://www.mbl.edu/Giardia>) as well as the NCBI databases (<http://www.ncbi.nlm.nih.gov>). Putative snoRNA prediction

used the modified Snoscan program (Snoscan-G) in C for Windows (the original source code is available at <http://lowelab.ucsc.edu/snoscan/>). However, the C-box scoring function was modified so that it read user-specified input of the C-box scoring matrix.

RNA structures were generated using the RNAfold program from the Vienna-RNA-1.4 Package (<http://www.tbi.univie.ac.at/~ivo/RNA/windoze/>), structural alignment was done using RSmatch1.0 converted for Windows (original program is available at <http://exon.umdj.edu/software/RSmatch/>) and FoldalignM (<http://foldalign.ku.dk/software/index.html>). rRNA sequence alignments for preliminary methylation site analysis were generated using ClustalW (44).

RT-PCR and PCR. RT-PCR reactions used Invitrogen ThermoScript first strand cDNA synthesis kit and subsequent PCR reactions used Roche Taq polymerase. Primers:

U5For: 5'-CATTTCATCTCTGCGGTGGATG-3'
U5Rev: 5'-ACCCCAAAAATGCAACTGTCTGCC-3'
U6For: 5'-CAAATTGAAACGATACAGAG-3'
U6Rev: 5'-TCATCCTTGTGCAGGGGCCA-3'
testP/GlsR15_For: 5'-GGGGAAGGTCGTGAGGTCATT-3'
testP/GlsR15_Rev: 5'-AGCTCATAGTCGTGCTTGCTC-3'

In vitro transcription and RNA self-cleavage assays. *In vitro* transcription reactions used the Invitrogen T7 RNA polymerase kit to add T7 promoter sequences to the 5' and 3' ends of the PCR products. The RNA products from *in vitro* transcription were heated to 80°C for 5 min and gradually cooled down to anneal. The dsRNAs were then purified using Roche PCR product purification kit. All the self-cleavage reactions were carried at 37°C for 2 h.

Primers used for generating templates for *in vitro* transcription:

Geniel_T7_For: 5'-TAATACGACTCACTATAGGGAGACGACCTCTCTCCAGCA-3'
Geniel_T7_Rev: 5'-TAATACGACTCACTATAGGGAGAGGAGCGCAAAGAGGATGA-3'
Girep1_T7_For: 5'-TAATACGACTCACTATAGGGAGATGCAGCCCTTCTGTCC-3'
Girep1_T7_Rev: 5'-TAATACGACTCACTATAGGGAGAGATACCCGGCTGTGC-3'

RESULTS

Assembly of cDNA sequences from the RNA library

Assembly of the cDNA sequences resulted in 31 novel ncRNAs, 15 previously known snoRNAs (40–42) and 10 out of 48 characterized tRNAs (<http://www.mbl.edu/Giardia>). Candidates were obtained in the following manner. A total of 616 initial sequences were assembled into 166 contigs and each contig was blasted against the *Giardia* genome database and NCBI databases to screen for easily characterized RNAs. After discarding empty vector contaminants, sequences below the length of 20 nt

and *E. coli* contaminant sequences, the remaining 152 contigs (including repeats or duplicates) contained 33 mRNA fragments, 28 known tRNA sequences, 10 5.8S rRNA sequences, 7 LSU rRNA and SSU rRNA fragments, 29 known ncRNAs sequences and 45 unknown sequences. All the unknown sequences were further analysed so that any broken fragments of a single RNA could be reassembled into a complete sequence, leaving 31 novel RNA candidates. Details of candidate sequences and features are listed in Supplementary Data. In order to carry out further computational analysis, 5'- and 3'-extensions (200 nt from each end) were extracted from the genome database for each candidate.

New C/D box snoRNA candidates and putative snoRNAs from computational studies

Eukaryotic 2'-O-methylation C/D box snoRNAs are characterized by two short sequence motifs near their 5'- and 3'-termini: C-box ('5'-AUGAUGA-3') and D-box ('5'-CUGA-3'), which are brought together by a short (4-8) terminal stem (45). There are one or two 10-20 nt antisense guide elements immediately upstream of the D-box or D'-box, and these elements bind to complementary sequences on rRNAs spanning the methylation sites (46). The position of the nucleotide which is methylated is usually the fifth position upstream of the D-box or D'-box (47).

Since the *Giardia* genome is fully sequenced (NCBI accession number: AACB00000000), it is possible to check our experimentally found RNAs for snoRNA features using potential interactions to rRNA sequences. Once we identify the conserved features of a *Giardia* snoRNA, we can identify more snoRNAs using a computational search. However, to date there are no full-length rRNA large subunit and small subunit rRNA sequences available for *Giardia*. Raw sequence reads from the GiardiaDB (<http://www.mbl.edu/Giardia>) were pulled out individually and assembled using SeqMan. Three contigs were generated, and correspond to the large subunit (LSU), small subunit (SSU) and 5.8S rRNAs, with lengths of 2908, 1449 and 138 nt respectively, and they arrange in the typical eukaryotic rRNA-gene order of SSU-5.8S-LSU, which reveals a site of cleavage by RNase MRP (6). The sequences are listed in Supplementary Data. Shortened lengths of the *Giardia* rRNAs are consistent with an earlier study (48) that *Giardia*'s rRNAs are much shorter than usual eukaryotic rRNAs, and unlike other eukaryotes, *Giardia* does not appear to have the 5S rRNA (48), which was also not found during our searches. The snoRNA search was done using modified source code of the Snoscan program, which was originally used to identify a large number of C/D-box snoRNAs from *Saccharomyces cerevisiae* (49).

We have predicted 3 C/D box snoRNA candidates from the 31 novel candidate sequences. Of the 15 known snoRNAs (40-42) that were found in our cDNA library, 14 are C/D box snoRNAs and 1 is an H/ACA box snoRNA. Comparing all the available C/D box, snoRNA sequences revealed that snoRNAs from *Giardia* share common sequence features within the

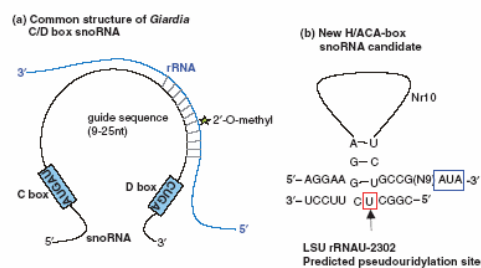


Figure 1. (a) Conserved structure of C/D box methylation snoRNAs in *Giardia*. (b) Structural prediction of the new H/ACA-box snoRNA candidate.

C boxes and D boxes. All but one of the confirmed C/D box snoRNAs has a perfect 'CUGA' D-box near the end of the 3'-end, and most *Giardia* C-boxes have a conserved sequence '5'-AUGAU-3'' allowing one mismatch at either 5'- or 3'-end. The C-box sequences also appear more variable as their lengths range between 5 and 7 nt. The C-box scoring function of Snoscan was adjusted to use the *Giardia* consensus sequence. The C'-box is generally missing or poorly identifiable, and the existence of D'-box is not essential. The length between the C- and D-boxes is varying from 28 to 124 nt. In addition, few of the known *Giardia* C/D box snoRNAs have a terminal stem.

The general structure of *Giardia* C/D box snoRNAs during rRNA modification is shown in Figure 1a. Structural alignment was done on all the experimentally found *Giardia* C/D box snoRNAs using RNA structures generated from Vienna-RNAfold program, but the result did not indicate any additional consensus motifs. Therefore, no further structural features were incorporated into modifying the Snoscan program. Our modified Snoscan program, Snoscan-G, identified 13 out of 18 confirmed C/D-box snoRNAs with the following parameters: cutoff total score (10), C-box score (2.0) and the maximum distant between C and D boxes (150 bp). The others were not recovered due to poorly defined C-boxes or imperfect D-boxes. This testing indicated that it was possible to identify additional C/D-box snoRNAs from the *Giardia* genome with this computational method. Table 2 shows the range of scores obtained from experimentally identified snoRNAs. These are considered as standard scores for *Giardia*, thus used to compare with the scores generated for computationally predicted snoRNAs further on. We refer to these computationally predicted snoRNAs as 'putative' snoRNAs in order to distinguish them from the 'candidate' snoRNAs found experimentally.

Due to the short (5 nt) and less conserved *Giardia* C-box, large volume of output was expected. A whole genome search for C/D box snoRNAs using the same parameter settings yielded many (6280) non-repetitive putative candidates, which were subsequently analysed through a strict three-step post-scan filtering.

Table 2. Snoscan scores obtained for experimentally identified *Giardia* C/D box snoRNAs

Feature	Consensus	Best score	Average score	Worst score
C box	AUGAU(GA)	8.76	7.9	3.55
D box	CUGA	8.05	7.9	3.77
D' box	CUGA	7.34	4.8	0.59
rRNA	9–25 nt with	33.93	22.7	15.92
complement	1 or 2 mismatches			
Total score		21.05	12.4	10

Three features of the putative snoRNAs were looked for during the post-scan filtering:

- (i) The sequences should locate in the non-coding regions.
- (ii) The sequences should locate close to reading frames since *Giardia* appears not to have separate transcription start sites for snoRNAs.
- (iii) The C-boxes of putative snoRNAs are more similar to the experimentally confirmed snoRNAs in *Giardia* (41).

All the output sequences from Snoscan-G were compared against the database of *Giardia* open-reading frames (ORFs) downloaded from GiardiaDB (<http://www.mbl.edu/Giardia>) to exclude possible mRNA sequences. These ORF datasets have been expertly compiled using software such as GLIMMER and CRITICA with parameters adjusted for this unique eukaryote. Our search of this database implicitly filtered out putative candidates with obvious coding potential. The status of the *Giardia* genome is such that a large number of ORFs remain hypothetical. Any explicit assessment for coding potential could be on only a subset of highly conserved proteins and would not be representative of the entire *Giardia* proteome. Hence, the use of this database maximizes our exclusion of contaminant mRNAs.

Unlike other eukaryotes, *Giardia* has only two confirmed introns (38,39), and most ncRNAs characterized to date are located between protein-coding genes, with a small number (less than 10) of them located on the minus strand of protein-coding genes. To exclude any ambiguities, only sequences located between protein-coding genes were considered. Sequence searches showed that most of the Snoscan-G outputs (5857) had full-length 100% match to ORFs, leaving 423 potential putative snoRNAs. After excluding shorter partial sequences and repetitive sequences with different names, 357 sequences remained. To date, all 13 experimentally confirmed C/D-box snoRNAs that had been detected in the small-scale Snoscan-G testing were also found in this large-scale genome search.

It was noticeable that all the experimentally characterized snoRNAs were located in ORF-rich regions of the genome, which could be due to the fact that these snoRNAs do not seem to possess their own promoters. Therefore, further screening was done based on genomic location.

Only putative sequences that are located near ORFs were selected with those appearing in heterochromatic regions excluded because they are less likely to be transcribed. This screening left 101 putative snoRNA sequences. Strict post-scan filtering based on C-box and D-box sequences was then done so that only sequences with 'AUGAU' or 'GUGAU' in C-box and 'CUGA' in D-box were considered as highly likely putative snoRNAs. In the end there were 60 strong putative snoRNAs. All sequences had distinct C-box and D-box motifs and fulfill the criteria for *Giardia* snoRNAs (41,45). In addition, they had average Snoscan-G total score of 12.5, which was slightly above the average total score of experimentally identified snoRNAs. The details of candidates are shown in the Supplementary Data.

As a control, we generated a random database with its size equivalent to *Giardia* genome using a third-order Markov chain based on 4-mer frequencies (49) within the *Giardia* genome. A search of this random sequence database yielded 6721 false positives with an average score of 11.8 and a best score of 25.26. As downstream filtering based on genomic location was impossible to carry out on randomized data, only the last step of the three-stage filtering could be performed on this output. Therefore, a parallel comparison between the *Giardia* Snoscan-G outputs and the randomized data outputs was not entirely applicable since the first two steps of the post-scan filtering were the most important and based on *Giardia* genomic information. However, a strict scan was still performed on this output with more stringent parameter settings based on C-box and D-box motifs, as was done in the final stage of post-scan filtering described above, reduced the positives down to 89 non-overlapping ones. Although these outputs contain C-box and D-box motifs, they do not represent comprehensive data for comparisons. In all, the purpose of generating a randomized dataset was to show that post-scan using genomic information was necessary to improve the selection of putative snoRNAs in a distant organism such as *Giardia*.

To test if the large number of initial output from the random database was due to special features within the *Giardia* genome, another Snoscan-G was run on a partial yeast genome (with a size similar to *Giardia* genome) using the same parameter settings. There were 1756 non-repetitive outputs. This test showed that the *Giardia* genome has less regional variation in its sequence, and this may result in the observation of more false positives. This testing showed that it was necessary to carry out stringent downstream filtering as was done in our Snoscan-G of the *Giardia* genome to obtain acceptable putative snoRNAs.

As an additional analysis, human and yeast C/D box snoRNAs have been mapped onto *Giardia* rRNAs (alignments included in Supplementary Data). Since human and yeast are extremely evolutionarily distant from *Giardia*, most known methylation sites do not have homologues in *Giardia*, apart from two. ncRNA candidate-1 from our cDNA library is predicted to guide methylation of G₁₁₃₁ on SSU-rRNA, which corresponds to the site of modification by human U25 snoRNA. Snoscan-G predicted putative snoRNA U0025 is likely to

guide methylation of C₁₁₉₁ on LSU-rRNA, which corresponds to the site of modification by an undetected human snoRNA. However, as these alignments are between such diverse organisms, no extensive conclusions can be drawn at this time.

In all, our Snoscan-G in combination of the post-scan filtering has identified 60 C/D-box snoRNA putative snoRNAs based on information from previously experimentally characterized snoRNAs. This approach was tested against two negative controls and showed that the use of *Giardia*-specific information made it possible to screen for functional ncRNAs in this reduced genome.

A new H/ACA box snoRNA candidate

The pseudouridylation guide H/ACA box snoRNAs have a common secondary structure consisting of two parallel hairpins linked by a hinge. Two conserved motifs box H (ANANNA) and box ACA are located at the hinge and the 3' tail, respectively, together with the flanking helix, they play important roles in box H/ACA snoRNA accumulation (50). However, compared to the single continuous antisense elements in box C/D snoRNAs, the antisense elements of H/ACA box snoRNAs are very short and bipartite (51). Almost all the H/ACA box snoRNAs adopt the two hairpin model, except one small H/ACA box snoRNA containing only one hairpin described in *Trypanosoma* (52). Based on hallmark sequences and structural features, one of the identified potential novel ncRNA (candidate 16, Supplementary Data), is likely to represent a novel H/ACA box snoRNA. It features a single, long stem positioned upstream from the ACA box motif as shown in Figure 1b. As such, it is strongly reminiscent of archaeal and Trypanosomal H/ACA box snoRNAs, that also feature a single hairpin (52–54). In agreement with the rules applying to eukaryotic H/ACA snoRNAs, the targeted uridine is separated from the H/ACA box by 9–16 nt. Therefore, according to structural modelling, we predict that candidate_16 may guide a pseudouridylation in LSU rRNA.

RNase P

The ribozyme RNase P cleaves the 5'-end of pre-tRNAs. The *Giardia* RNase P RNA was recently identified by sequence similarity search and the RNase P holoenzyme was purified (20), and showed that *Giardia* RNase P RNA has the conserved eukaryotic RNase P core structure, and shared extensive similarity with the RNase P RNA of the microsporidian *Encephalitozoon cuniculi*. Both RNAs lack the conserved P3 helix bulge loop, which has been found in all the other eukaryotes studied so far. The RNase P RNA has been found in our library (candidate 9), but surprisingly, the sequence was not terminated at the previously predicted 3' end, and extended further into the GlsR15 snoRNA (41). These two known RNAs have a 24 nt overlap, which is shown in Figure 2. It is likely that candidate 9 is part of a full-length RNA transcript. To verify this idea, RT-PCR was done using an upstream primer (testP/GlsR15_For) that binds within the RNase P sequence (position 34–53 on the possible full length

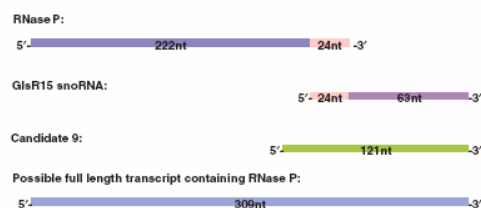


Figure 2. Comparison of RNase P, GlsR15 snoRNA and the new ncRNA candidate 9.

transcript) and a downstream primer (testP/GlsR15_Rev) which binds within the GlsR15 snoRNA sequence (position 269–289 on the possible full length transcript).

RT-PCR results (data not shown) indicate that the RNase P and GlsR15 are indeed transcribed as a single transcript. This rises to a question that whether this transcript is a single functional RNA molecule, or a precursor to give two different RNAs. Structural studies (20) indicate that the shorter transcript could fold with conserved eukaryotic RNase P motifs. Therefore, the second assumption is preferred. It is possible that an as yet unknown ribonuclease is involved in producing two different RNAs from one precursor. However, this leads to a result that only one of the two RNAs can be generated as a full-length molecule and the other one will be non-functional.

Transcribed intergenic repeats

A fragment of the variant surface protein (VSP) mRNA was found in the cDNA library. It has been suggested (55) that antisense regulation controls the expression of VSP genes, and the function of RNA-dependent RNA polymerase (RdRp) is involved to restrict the VSP gene repertoire to a single gene at any one time. Careful sequence mining within the *Giardia* genome observed that there were many tandem repeats sharing short sequence fragments, and these fragments are often complementary to repeated sequences in VSP genes and cysteine-rich protein genes. Blasting a VSP-fragment sequence found in our cDNA library against the *Giardia* genome yielded a potentially functional antisense element. This sequence is a long tandem repeat consisting of nine units, each containing one fragment complementary to the VSP ORF (Figure 3). RT-PCR was carried out targeting both the '+' and '-' strand of this sequence, and the results showed that both strands were transcribed, to give a double-stranded RNA product.

Unlike other tandem repeats of retrotransposons such as LINES or SINES, this tandem repeat shows no feature of any known retrotransposon. In comparison, there have been a few studies on unusual repeated sequences in *Giardia*: one study (56) showed a non-LTR element with site-specific tandem insertions in a chromosomal DNA repeat, and suggested that this element was unlikely to have evolved site specificity unless it did have a function. Another more recent study showed this element was transcribed into a dsRNA (57). In addition, there are

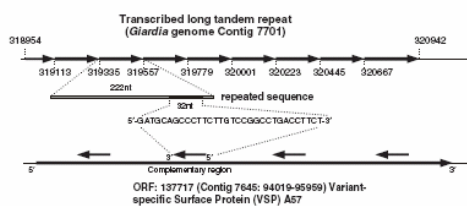


Figure 3. Tandem repeats of the Girep-1 RNA. Each fragment coloured in grey represents a repeating unit (222 nt in length, with the first unit lacking the 5' 63 nt and the last unit extending 54 nt at 3' end) on Girep-1 RNA. Each 32 nt fragment coloured in black represents the repeating Girep-1 sequence that is complementary to the various-surface-protein (VSP) gene.

22 antisense transcripts identified in the *Giardia* genome (www.mbl.edu/Giardia); however, there are no known functions of these transcripts.

Our study has revealed a surprising feature shared by two tandem repeats in *Giardia*: one repeat is the experimentally verified dsRNA with fragments complementary to the VSP (Rep-1); and the other is the non-LTR element Genie-1 (56). A partial sequence from each element was amplified by PCR with T7-promoter attached primers. The PCR products were transcribed by T7 RNA polymerase to produce dsRNAs. As a control, a single stranded Rep-1 RNA was also produced by elimination of T7 promoter sequence from the reverse primers. Both dsRNAs underwent one self-cleavage at roughly the middle of the sequence (under a basic assay condition with Mg^{2+} added to water or buffer) (Figure 4a). The single stranded Rep-1 control did not cleave (Figure 4b). Timing Mg^{2+} titration (Figure 4c) assay and divalent ion assays (Figure 4d) were performed with the Genie-1 dsRNA. Results showed that the self-cleavage did not happen when Mg^{2+} concentration was below 1 mM; and self-cleavage only happened at the present of Mg^{2+} or Co^{2+} , while Mn^{2+} and Ca^{2+} did not have any effect. In addition, addition of EDTA prevented Mg^{2+} induced cleavage. Further investigation will be necessary to analyse this unusual phenomenon.

DISCUSSION

Combined experimental and computational approach

The aim of this study was to explore the variety of ncRNAs in *Giardia* and obtain a view of ncRNA expression in this genomically reduced deep-branching eukaryote. The scale of this cDNA library is small compared with equivalent studies of ncRNAs in other organisms (28–32). However, studying on a relatively small scale can help getting a comprehensive view of the special features and conserved patterns within this organism, before any large scale studies are attempted. There were previously no systemic studies on the ncRNA composition of *Giardia*. As an extant group of eukaryotes, Diplomonads share very low sequence homology with other eukaryotes, which makes characterization of RNAs extremely difficult. From the 31 novel ncRNA candidates,

only 3 can be identified by homology searching as C/D box snoRNAs, the rest have little similarities to known types of ncRNAs.

However, comparing the 18 characterized C/D box snoRNAs from *Giardia* has shown that these snoRNAs still share the basic conserved features seen with snoRNAs from other eukaryotes. This makes a computational screen possible. Within the computationally identified putative snoRNAs, we recovered 13 out of our control set of the 18 experimentally characterized snoRNAs. Snoscan-G used looser parameters than the original Snoscan program in order that the experimentally identified snoRNAs (13 in this study) were included in the results. This ensured the sensitivity of the algorithm which was then used for a whole-genome search. However, the large number of false positive hits obtained from the negative control search on a random database, indicated the requirement for other post-scan filtering of putative snoRNA sequences using data unable to be included in the Snoscan-G software. Also, a fairly large result obtained from scan of the yeast genome confirms that the parameter settings for Snoscan-G are less stringent than the original Snoscan program. Comparing putative snoRNA sequences against the ORF database excluded most of the first-round positive hits, and information from genomic locations of the sequences extended the reliability of the putative snoRNAs.

Possibly due to its reduced genome, *Giardia*'s snoRNAs are less conserved than those of other eukaryotic organisms; therefore it was necessary to apply less stringent searching criteria. This is because there are as yet no additional *Giardia*-specific sequence features, which can be incorporated into the algorithm. This explains the increase in false positives when large databases are screened. However, combining several filtering steps dramatically reduced the number of positive hits, and at the same time did not result in the loss of any true positives. The remaining putative snoRNAs showed greater similarities to the experimentally identified snoRNAs than the first-round Snoscan-G results before post-scan filtering. Therefore, our computational approach is reliable when used in parallel with an experimental approach speeding up the discovery of novel putative ncRNAs.

Encoding patterns of ncRNAs in *Giardia*

Blasting the novel RNA candidates against the *Giardia* genome revealed three types of encoding patterns.

- (i) Most ncRNAs in *Giardia* are encoded as single copies between protein-coding genes. According to current knowledge of *Giardia*, almost all the protein-coding genes are intronless (38,39) and it becomes natural that ncRNAs find their places in intergenic regions. The genome of *Giardia* is compact; and the genes generally have very short gaps (often <200 nt) between one another. Almost all the ncRNAs observed so far are located in ORF-rich regions, but do not appear to possess their own promoters, although this may be due to the fact that *Giardia* does not appear to have well characterized

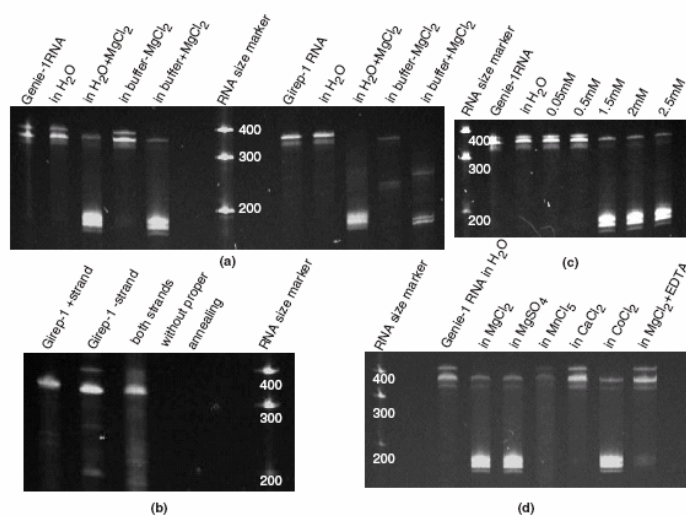


Figure 4. Self-cleavage reactions of the Genie-1 dsRNA and Girep-1 dsRNA. All the reactions were incubated at 37°C for 2 h, and run on 8% denaturing polyacrylamide gel containing 7M urea at 350 V. (a) Self-cleavage reactions of dsGenie-1 RNA (left to the size marker) and dsGirep-1 RNA (right to the size marker); buffer: 20 mM HEPES, 150 mM NaCl, with or without 2.5 mM MgCl₂; (b) The test of ssRNA of Girep-1 in water with 2.5 mM MgCl₂; (c) Mg²⁺ titration assay of Genie-1 dsRNA and (d) the test of different ions with dsGenie-1 RNA in water with 2.5 mM of each divalent ion, and EDTA was added to 50 mM on the last lane.

promoter sequences as there is a lack of conserved sequence in the promoter region. One possibility is that these ncRNAs may co-transcribe with their adjacent ORFs, and the pre-transcripts are later processed to give mRNAs and ncRNAs. If this is the case, there must be specific RNA-processing machinery to carry out the task. One possible candidate is the spliceosome, as it is highly unlikely for a whole spliceosome to remain just for processing two introns (38,39).

- (ii) Three novel candidates from the cDNA library show polymorphic variations in having several nearly identical copies in the genome with most of the polymorphic copies not located near predicted ORFs. It is not known if all the polymorphic copies of these RNAs are transcribed, because for each of the three candidates only one form has been seen in our cDNA library. Some of these polymorphic copies are encoded in tandem repeats, but the rest are located in a distant part of the genome. It has been known that some ncRNAs such as U2 snRNAs in *Xenopus* do have this feature in developmental regulation (58); however, the polymorphic forms of our candidates do not have any sequence similarity to known spliceosomal snRNAs.
- (iii) Long retrotransposon-like tandem repeats of ncRNAs are described in the Results section. The experimentally confirmed tandem repeat is located in an ORF-rich area of the genome with both

'+' and '-' strands adjacent to neighbouring protein-coding genes. We suggest that it is likely that they are co-transcribed with mRNAs and are subsequently cleaved by a specific but yet unknown mechanism. The novel self-cleaving feature of the dsRNAs derived from the two retrotransposon-like elements will require further investigation.

The puzzle of spliceosomal snRNAs in *Giardia*

There is very little known about splicing in *Giardia*. Sequence mining from the genome shows that most of the eukaryotic specific spliceosomal proteins (9) are present in *Giardia*, as well as the important U5 snRNA (59), which functions at the centre of both major and minor spliceosomes. It is common in eukaryotes that the spliceosomal snRNAs are expressed at a high level (60), since intron splicing generally occurs at a high rate. However, it seems not the case in *Giardia*. We did not find any sequence in our cDNA library with similarity to any known spliceosomal snRNA. To determine the possible presence of any spliceosomal snRNAs in the library, PCR reaction using the U5 primers (Materials and Methods section) was done on the cDNAs. Results show that U5 snRNA is expressed and present, but in very low quantities. Another puzzling question concerns the U6 snRNA. U6 snRNA is the most conserved spliceosomal snRNA across all the eukaryotes studied to date. U6 snRNAs take part in the actual catalysis during

splicing (61), and share extensive sequence similarities across eukaryotes. In an early study (62), it has been shown that a single pair of PCR primers could detect U6 snRNAs from 17 different species of eukaryotes. As a trial, the same pair of primers was tried on *Giardia* in both genomic PCR and RT-PCR reactions. Despite extensive effort, there is as yet no detectable candidates for a *Giardia* U6 snRNA. It is therefore concluded that our current approach is not powerful enough to solve the puzzle of *Giardia*'s spliceosomal snRNAs.

Novel ncRNA candidates

Total 26 out of our 31 novel RNA candidates cannot yet be extensively characterized as belonging to any known class of ncRNA; a feature seen in other species-specific studies (29). Structural studies and motif analysis of these RNAs did not show distinct features found in known ncRNAs. A number of these RNAs are GC rich, providing a basis for strong helical structures. Lack of characterization is possibly due to the highly divergent sequences of *Giardia* compared to those of the major eukaryotic groups, and because most computer programs developed for identifying ncRNAs are based on human and yeast. One way to further approach the identification of ncRNA is through more computational studies by incorporating more *Giardia*-specific information into the existing programs, followed by experimental verification of our proposed candidates. Another way is through biochemical studies of central protein components of various RNA processing pathways. These are to be investigated in the future.

In conclusion, our cDNA library successfully uncovered 31 novel ncRNAs from *Giardia*, and our computational approach was shown to be a useful method that worked well in parallel with an experimental approach to aid discovery of 60 potential putative snoRNAs in a deep-branching eukaryote. Although it is hard to characterize each candidate ncRNAs found from the cDNA library due to sequence divergence, as far as we can tell, *Giardia* has quite typical eukaryotic RNA processing despite being reduced and with many introns lost. The transcriptional patterns seen in these ncRNAs may help in understanding the mechanism of RNA processing. Future work will continue to be done in investigating the unusual properties of ncRNAs by combined biochemical and computational methods.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank George Ionas and Errol Kuan from Microaquatech for kind supply of *G. intestinalis* culture, Anja Zemmann and Claudia Marker (ZMBE) for great help with the cDNA library and Timothy White (Allan Wilson Centre) for time and effort on computer programming. This work is supported by the New Zealand Marsden

Fund; the Allan Wilson Centre for Molecular Ecology and Evolution; European Union (EU; LSHG-CT-2003-503022) and the Nationales Genomforschungsnetz (NGFN; 0313358A). Funding to pay the Open Access publication charges for this article was provided by the New Zealand Marsden Fund.

Conflict of interest statement. None declared.

REFERENCES

- Mattick,J.S. and Makunin,I.V. (2006) Non-coding RNA. *Hum. Mol. Genet.*, **15** Spec No 1, R17–R29.
- Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
- Gilbert,W. (1986) The RNA world. *Nature*, **319**, 618.
- Brosius,J. (2005) Echoes from the past – are we still in an RNP world? *Cytogenet. – Genome Res.*, **110**, 8–24.
- Jeffares,D.C., Poole,A.M. and Penny,D. (1998) Relics from the RNA world. *J. Mol. Evol.*, **46**, 18–36.
- Woodhams,M.D., Stadler,P.F., Penny,D. and Collins,L.J. (2007) RNase MRP and the RNA processing cascade in the eukaryotic ancestor. *BMC. Evol. Biol.*, **7**(Suppl. 1), S13.
- Kramer,A. (1996) The structure and function of proteins involved in mammalian pre-mRNA splicing. *Annu. Rev. Biochem.*, **65**, 367–409.
- Patel,A.A. and Steitz,J.A. (2003) Splicing double: insights from the second spliceosome. *Nat. Rev. Mol. Cell Biol.*, **4**, 960–970.
- Collins,L. and Penny,D. (2005) Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.*, **22**, 1053–1066.
- Russell,A.G., Charette,J.M., Spencer,D.F. and Gray,M.W. (2006) An early evolutionary origin for the minor spliceosome. *Nature*, **443**, 863–866.
- Bachelier,J.P., Cavaille,J. and Huttenhofer,A. (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.
- Brown,J.W., Echeverria,M. and Qu,L.H. (2003) Plant snoRNAs: functional evolution and new modes of gene expression. *Trends Plant Sci.*, **8**, 42–49.
- Uliel,S., Liang,X.H., Unger,R. and Michaeli,S. (2003) Small nucleolar RNAs that guide modification in trypanosomatids: repertoire, targets, genome organisation, and unique functions. *Int. J. Parasitol.*, **33**, 235–255.
- Dennis,P.P. and Omer,A. (2005) Small non-coding RNAs in Archaea. *Curr. Opin. Microbiol.*, **8**, 685–694.
- Mehler,M.F. and Mattick,J.S. (2006) Non-coding RNAs in the nervous system. *J. Physiol.*, **575**, 333–341.
- Newby,M.I. and Greenbaum,N.L. (2001) A conserved pseudouridine modification in eukaryotic U2 snRNA induces a change in branch-site architecture. *RNA*, **7**, 833–845.
- Watkins,N.J., Segault,V., Charpentier,B., Nottrott,S., Fabrizio,P., Bachi,A., Wilm,M., Rosbash,M., Branlant,C. *et al.* (2000) A common core RNP structure shared between the small nucleolar box C/D RNPs and the spliceosomal U4 snRNP. *Cell*, **103**, 457–466.
- Staley,J.P. and Guthrie,C. (1998) Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell*, **92**, 315–326.
- Frank,D.N. and Pace,N.R. (1998) Ribonuclease P: unity and diversity in a tRNA processing ribozyme. *Annu. Rev. Biochem.*, **67**, 153–180.
- Marquez,S.M., Harris,J.K., Kelley,S.T., Brown,J.W., Dawson,S.C., Roberts,E.C. and Pace,N.R. (2005) Structural implications of novel diversity in eucaryal RNase P RNA. *RNA*, **11**, 739–751.
- Collins,L.J., Moulton,V. and Penny,D. (2000) Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. *J. Mol. Evol.*, **51**, 194–204.
- Chamberlain,J.R., Lee,Y., Lane,W.S. and Engelke,D.R. (1998) Purification and characterization of the nuclear RNase P holoenzyme complex reveals extensive subunit overlap with RNase MRP. *Genes Dev.*, **12**, 1678–1690.

23. Hammond, S.M., Caudy, A.A. and Hannon, G.J. (2001) Post-transcriptional gene silencing by double-stranded RNA. *Nat. Rev. Genet.*, 2, 110–119.
24. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, 31, 439–441.
25. Bernstein, E., Caudy, A.A., Hammond, S.M. and Hannon, G.J. (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409, 363–366.
26. Macrae, I.J., Zhou, K., Li, F., Repic, A., Brooks, A.N., Cande, W.Z., Adams, P.D. and Doudna, J.A. (2006) Structural basis for double-stranded RNA processing by Dicer. *Science*, 311, 195–198.
27. Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E.G., Margalit, H. and Altuvia, S. (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.*, 11, 941–950.
28. Yuan, G., Klamt, C., Bachelier, J.P., Brosius, J. and Huttenhofer, A. (2003) RNomics in *Drosophila melanogaster*: identification of 66 candidates for novel non-messenger RNAs. *Nucleic Acids Res.*, 31, 2495–2507.
29. Zemann, A., op de Bekke, A., Kiefmann, M., Brosius, J. and Schmitz, J. (2006) Evolution of small nucleolar RNAs in nematodes. *Nucleic Acids Res.*, 34, 2676–2685.
30. Huttenhofer, A., Kiefmann, M., Meier-Ewert, S., O'Brien, J., Leirach, H., Bachelier, J.P. and Brosius, J. (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.*, 20, 2943–2953.
31. Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H. and Bartel, D.P. (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell*, 127, 1193–1207.
32. Marker, C., Zemann, A., Terhorst, T., Kiefmann, M., Kastenmayer, J.P., Green, P., Bachelier, J.P., Brosius, J. and Huttenhofer, A. (2002) Experimental RNomics: identification of 140 candidates for small non-messenger RNAs in the plant *Arabidopsis thaliana*. *Curr. Biol.*, 12, 2002–2013.
33. Aspegren, A., Hinas, A., Larsson, P., Larsson, A. and Soderbom, F. (2004) Novel non-coding RNAs in *Dictyostelium discoideum* and their expression during development. *Nucleic Acids Res.*, 32, 4646–4656.
34. Huttenhofer, A. and Vogel, J. (2006) Experimental approaches to identify non-coding RNAs. *Nucleic Acids Res.*, 34, 635–646.
35. Vanacova, S., Liston, D.R., Tachezy, J. and Johnson, P.J. (2003) Molecular biology of the amitochondriate parasites, *Giardia intestinalis*, *Entamoeba histolytica* and *Trichomonas vaginalis*. *Int. J. Parasitol.*, 33, 235–255.
36. Keeling, P.J., Burger, G., Durnford, D.G., Lang, B.F., Lee, R.W., Pearlman, R.E., Roger, A.J. and Gray, M.W. (2005) The tree of eukaryotes. *Trends Ecol. Evol.*, 20, 670–676.
37. Nixon, J.E., Wang, A., Field, J., Morrison, H.G., McArthur, A.G., Sogin, M.L., Loftus, B.J. and Samuelson, J. (2002) Evidence for lateral transfer of genes encoding ferredoxins, nitroreductases, NADH oxidase, and alcohol dehydrogenase 3 from anaerobic prokaryotes to *Giardia lamblia* and *Entamoeba histolytica*. *Eukaryot. Cell*, 1, 181–190.
38. 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. USA*, 99, 3701–3705.
39. 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.
40. Niu, X.H., Hartshorne, T., He, X.Y. and Agabian, N. (1994) Characterization of putative small nuclear RNAs from *Giardia lamblia*. *Mol. Biochem. Parasitol.*, 66, 49–57.
41. Yang, C.Y., Zhou, H., Luo, J. and Qu, L.H. (2005) Identification of 20 snoRNA-like RNAs from the primitive eukaryote, *Giardia lamblia*. *Biochem. Biophys. Res. Commun.*, 328, 1224–1231.
42. Luo, J., Zhou, H., Chen, C.H., Li, Y., Chen, Y. and Qu, L.H. (2006) Identification and evolutionary implication of four novel box H/ACA snoRNAs from *Giardia lamblia*. *Chin. Sci. Bull.*, 51, 2451–2456.
43. Kurland, C.G., Collins, L.J. and Penny, D. (2006) Genomics and the irreducible nature of eukaryote cells. *Science*, 312, 1011–1014.
44. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. and Thompson, J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, 31, 3497–3500.
45. Samarsky, D.A., Fournier, M.J., Singer, R.H. and Bertrand, E. (1998) The snoRNA box C/D motif directs nucleolar targeting and also couples snoRNA synthesis and localization. *EMBO J.*, 17, 3747–3757.
46. Cavaillat, J., Nicoloso, M. and Bachelier, J.P. (1996) Targeted ribose methylation of RNA in vivo directed by tailored antisense RNA guides. *Nature*, 383, 732–735.
47. Kiss-Laszlo, Z., Henry, Y. and Kiss, T. (1998) Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. *EMBO J.*, 17, 797–807.
48. Edlind, T.D. and Chakraborty, P.R. (1987) Unusual ribosomal RNA of the intestinal parasite *Giardia lamblia*. *Nucleic Acids Res.*, 15, 7889–7901.
49. Lowe, T.M. and Eddy, S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, 283, 1168–1171.
50. Ganot, P., Bortolin, M.L. and Kiss, T. (1997) Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell*, 89, 799–809.
51. Ganot, P., Caizergues-Ferrer, M. and Kiss, T. (1997) The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes Dev.*, 11, 941–956.
52. Liang, X.H., Liu, L. and Michaeli, S. (2001) Identification of the first trypanosome H/ACA RNA that guides pseudouridine formation on rRNA. *J. Biol. Chem.*, 276, 40313–40318.
53. Tang, T.H., Polacek, N., Zywicki, M., Huber, H., Brugger, K., Garrett, R., Bachelier, J.P. and Huttenhofer, A. (2005) Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. *Mol. Microbiol.*, 55, 469–481.
54. Rozhddestvensky, T.S., Tang, T.H., Tchirkova, I.V., Brosius, J., Bachelier, J.P. and Huttenhofer, A. (2003) Binding of L7Ae protein to the K-turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea. *Nucleic Acids Res.*, 31, 869–877.
55. Ullu, E., Tschudi, C. and Chakraborty, T. (2004) RNA interference in protozoan parasites. *Cell Microbiol.*, 6, 509–519.
56. Burke, W.D., Malik, H.S., Rich, S.M. and Eickbush, T.H. (2002) Ancient lineages of non-LTR retrotransposons in the primitive eukaryote, *Giardia lamblia*. *Mol. Biol. Evol.*, 19, 619–630.
57. Ullu, E., Lujan, H.D. and Tschudi, C. (2005) Small sense and antisense RNAs derived from a telomeric retroposon family in *Giardia intestinalis*. *Eukaryot. Cell*, 4, 1155–1157.
58. Mattaj, I.W. and Zeller, R. (1983) *Xenopus laevis* U2 snRNA genes: tandemly repeated transcription units sharing 5' and 3' flanking homology with other RNA polymerase II transcribed genes. *EMBO J.*, 2, 1883–1891.
59. Collins, L.J., Macke, T.J. and Penny, D. (2004) Searching for ncRNAs in eukaryotic genomes: maximizing biological input with RNAmotif. *J. Integr. Bioinformatics*, 1, 61–77.
60. Riedel, N., Wise, J.A., Swerdlow, H., Mak, A. and Guthrie, C. (1986) Small nuclear RNAs from *Saccharomyces cerevisiae*: unexpected diversity in abundance, size, and molecular complexity. *Proc. Natl Acad. Sci. USA*, 83, 8097–8101.
61. Yean, S.L., Wuenschell, G., Termini, J. and Lin, R.J. (2000) Metal-ion coordination by U6 small nuclear RNA contributes to catalysis in the spliceosome. *Nature*, 408, 881–884.
62. Tani, T. and Ohshima, Y. (1991) mRNA-type introns in U6 small nuclear RNA genes: implications for the catalysis in pre-mRNA splicing. *Genes Dev.*, 5, 1022–1031.

Chapter Three – Further analysis of the ncRNA library of *Giardia intestinalis*

Abstract

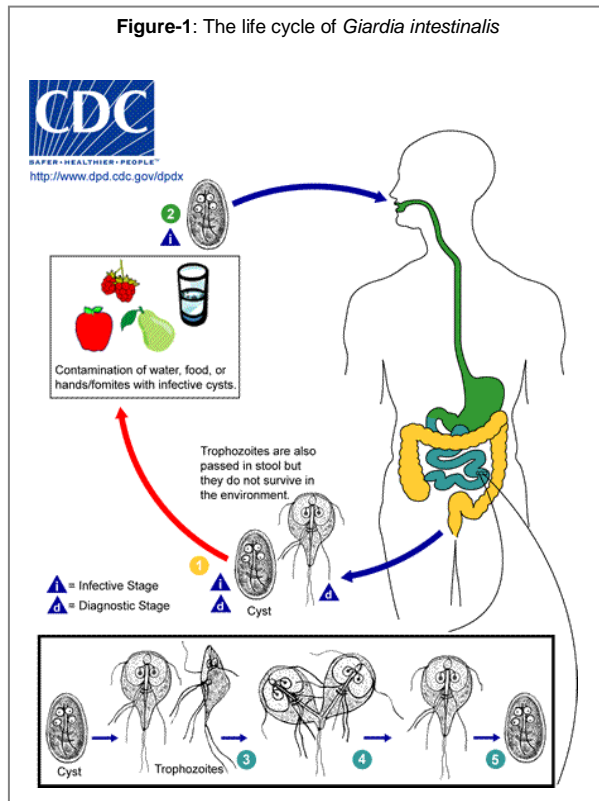
This chapter extends the analysis on our *Giardia* cDNA library. After more clones were sequenced, the collective data agrees with my previous observation that many *Giardia* novel ncRNAs do not show similarity with known types of ncRNAs. The reason behind the phenomenon is not yet clear, but is likely to be resulted from long evolutionary deviation of *Giardia* from the other eukaryotes. Nonetheless, analysis of expressional patterns of various classes of ncRNAs in *Giardia* reveals conservation of certain upstream sequence motifs within proposed promoter regions. The transcription apparatus in *Giardia* is known to be highly reduced. New information obtained from my studies about the potential features of ncRNA transcription in *Giardia* may lead to further investigation of the transcriptional systems in distant eukaryotes. In addition, potential new protein candidates of *Giardia* RNA polymerase system are also presented here. Finally, detailed structural analysis of the novel ncRNA candidates was performed using specialized RNA structural alignment tools. Results indicate a number of conserved structures within these novel ncRNAs of *Giardia*.

3.1 Background: Molecular biology of *Giardia intestinalis* and techniques in the studies of ncRNA

3.1.1 *Giardia* – a deep-branching unicellular eukaryote

Giardia intestinalis (commonly known as *Giardia lamblia*) is an enteric parasite of the small and large intestine, and can cause severe diarrhoea. *Giardia* has a two-stage life cycle consisting of trophozoite and cyst (Figure-1, from <http://www.dpd.cdc.gov>). The life cycle begins with ingested cysts, which release trophozoites. The trophozoites then attach to the surface of the intestinal epithelium, and reproduce by binary fission. However, the trigger for encystment is still unclear. Cysts are released in faeces and can reinfect additional hosts. *Giardia* has a characteristic tear-drop shape and measures 10

–15 μm in length. It has two identical nuclei and an adhesive disk reinforced by microtubules.

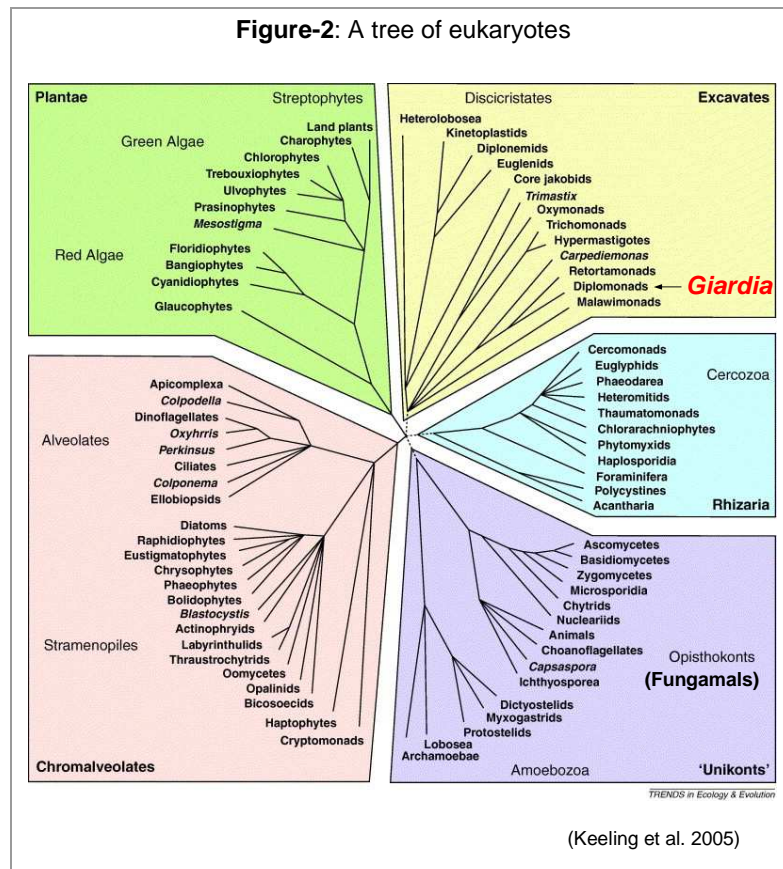


Giardia was traditionally classified with other flagellated protozoans, including kinetoplastids, parabasalids and dientamoeba (Levine et al. 1980). It has now been placed within excavates, as a member of diplomonads along with the mole parasite *Spironucleus muris* (Januschka et al. 1988). The widely accepted classification is based on 18S rRNA sequences, and *Giardia* was proposed to be one of the

most “primitive” eukaryotic organisms, together with other amitochondria eukaryotes such as *Trichomonas vaginalis* and microsporidians (Sogin et al. 1989). However, discovery of nuclear genes with mitochondrial ancestry in *Giardia*, *Entamoeba* and microsporidians (Roger et al. 1998; Tachezy et al. 2001; Arisue et al. 2002), together with the finding of mitochondrial remnant organelles in these amitochondrial protists (Tovar et al. 1999; Williams et al. 2002; Tovar et al. 2003) shows that rather than being primitive, these organisms are more likely to be the result of reductive evolution, in an anaerobic environment.

A recent tree of eukaryotes (Figure-2) divides them into five super-groups (Keeling et al. 2005), with a diverse group of protists named “the excavates”. The order of branching of these five groups is not known, nor is the position of the root. Mitochondrial proteins and remnant organelles have now been found in most of the excavates except for oxymonads and retortamonads (Dyall et al. 2004). It is unclear whether excavates should still be considered early

eukaryotic lineages, although they do show features of ancestral eukaryotic cells such as reduced organelles and transcription apparatus.



Despite the uncertainty of its phylogeny, *Giardia* is a deep-branching eukaryote with reduced cellular architecture, and is very distant from both multicellular plants and animals. As such it is important to study its biochemical properties to be able to help infer properties of ancestral eukaryotes. Its trophozoites are symmetric along the long axis. Lysosomal vacuoles, ribosomal and glycogen granules are found in the cytoplasm (Adam 2001). Golgi complexes are visible in the encysting trophozoites but not in the vegetative trophozoites (Gillin et al. 1996). However, demonstration of stacked membranes suggests the presence of Golgi complexes (Lanfredi-Rangel et al. 1998), and genes of proteins associated with Golgi complexes have been identified (Dacks et al. 2003). *Giardia* does not have recognizable

* Hydrogenosome (Lindmark and Muller, 1973): a membrane-bound organelle of some anaerobic ciliates, trichomonads and fungi. It produces molecular hydrogen and ATP. This organelle is thought to have most likely evolved from mitochondria.

hydrogenosomes* (Lindmark and Muller 1973), but hydrogenase activity has been detected in cell-free extracts under anaerobic conditions (Lloyd et al. 2002). Nucleoli have not been identified in *Giardia*, however important pre-rRNA processing protein homologues normally found in nucleoli of other species are found in *Giardia* (Narcisi et al. 1998; Xin et al. 2005).

The reduced, but fully functional cell of *Giardia* suggests an unusual evolutionary path of this organism. Perhaps not so ancient, *Giardia* (as an excavate) can still represent a basal form of eukaryotes for its relatively simple cellular features, and may still show properties of ancestral eukaryotic cell before the formation of fully specialized multicellular eukaryotes.

3.1.2 The genome of *Giardia*

The *Giardia* genome-sequencing project (McArthur et al. 2000; Morrison et al. 2007) enables extensive experimental and computational studies on genes, transcription and evolution of this organism. The 12 Mb genome of *Giardia* is localized on five chromosomes ranging in size from approximately 1.6 Mb to 3.8 Mb (Adam et al. 1988). The genome of *Giardia* is highly compact in its structure and content with simplified DNA replication, transcription and RNA processing mechanisms (Morrison et al. 2007). The chromosomes have typical eukaryotic features such as the “TAGGG” telomeric repeat (Le Blancq et al. 1991), four core histones (H2a, H2b, H3 and H4), and a linker histone (H1) (Wu et al. 2000). The largest protein family of *Giardia* is the protein kinases, indicating the use of extensive signal transductions and the newly annotated genome (Morrison et al. 2007) has revealed a previously unknown family of cysteine-rich structural proteins.

The transcription apparatus of *Giardia* is basal compared to crown eukaryotes. A survey by Best et al. of the genome revealed homologues to 21 of the 28 proteins comprising the three typical eukaryotic RNA polymerases, and 4 of

the 12 general transcription-initiation factors, also the *Giardia* TATA-box binding protein is highly divergent from the homologous proteins in eukaryotes and archaea (Best et al. 2004). These data suggest two possibilities that either *Giardia* evolved after the origin of the core transcription apparatus but before the complete evolution of all the transcriptional factors; or *Giardia* once possessed complete transcriptional machinery, but subsequently (possibly because of a parasitic lifestyle) reduced to having only a minimum set of transcriptional factors. The deep position of diplomonads, microsporidia and other protists in molecular phylogenies brought about extensive debates. It has been suggested that the deep position is due to artefacts in phylogenetic reconstruction methods such as long-branch attractions (Hirt et al. 1999). However diplomonads have been consistently placed among early-branching eukaryotes in both rRNA and protein phylogenies, and have not associated with any late-emerging phylogenetic groups (Dacks et al. 2002; Inagaki et al. 2003; Moreira et al. 2006). In comparison, the microsporidium *Encephalitozoon cuniculi* also has a highly reduced genome, but its genome contains a full set of RNA polymerase II general transcription factors (Katinka et al. 2001), which is consistent with a late-emerging and less reduced transcriptional system.

Giardia synthesizes a surprisingly abundant and diverse array of sterile transcripts unable to code for proteins. A random sampling of two evolutionarily divergent *Giardia* strains showed that about 20% of the cDNAs analysed were polyadenylated sterile transcripts (Elmendorf et al. 2001a). To date there are only three introns published for *Giardia* (Nixon et al. 2002; Russell et al. 2005) and another three unpublished ones (personal communication with Scott Roy, NIH). The untranslated regions (UTRs) of mRNAs are typically less than 20nt at the 5'-end and less than 50nt at the 3'-end (Adam 2000). It has been shown that only a short region (<50nt) of upstream sequence is required to drive expression of a reporter gene in transfected *Giardia* (Yee et al. 2000; Elmendorf et al. 2001b). Also *Giardia*'s promoters are poorly conserved and are likely degenerate (Holberton and Marshall 1995). Analysis of the antisense transcripts showed that antisense

transcription is not restricted to a few loci (Elmendorf et al. 2001a), hence the authors suggested that they were more likely the results of loose transcriptional regulation than involving in antisense regulation.

In summary, despite the evidence that *Giardia* once had mitochondria, lacking of a number of general transcription factors and having a highly degenerate transcription system suggest a relatively basal phylogenetic position. However, studies on the molecular biology of *Giardia* can still provide glimpses to the cellular state ancestral to the radiation of animals, plants and fungi, because comparing the data from *Giardia* with other eukaryotes will reveal common features that are shared among all lineages, and the common features are likely to represent the ancestral state.

3.1.3 Techniques in the studies of ncRNAs

Several methods for ncRNA identification and characterization have been well established (Huttenhofer and Vogel 2006). The common methods include (1) enzymatical or chemical sequencing of RNA; (2) cloning of ncRNAs by generating cDNA libraries; (3) the use of microarrays to analyse expression of ncRNAs under different conditions; (4) the “genomic SELEX” approach to select ncRNA candidates from the genomes of organisms of interest; (5) bioinformatic tools to screen genomes for ncRNAs.

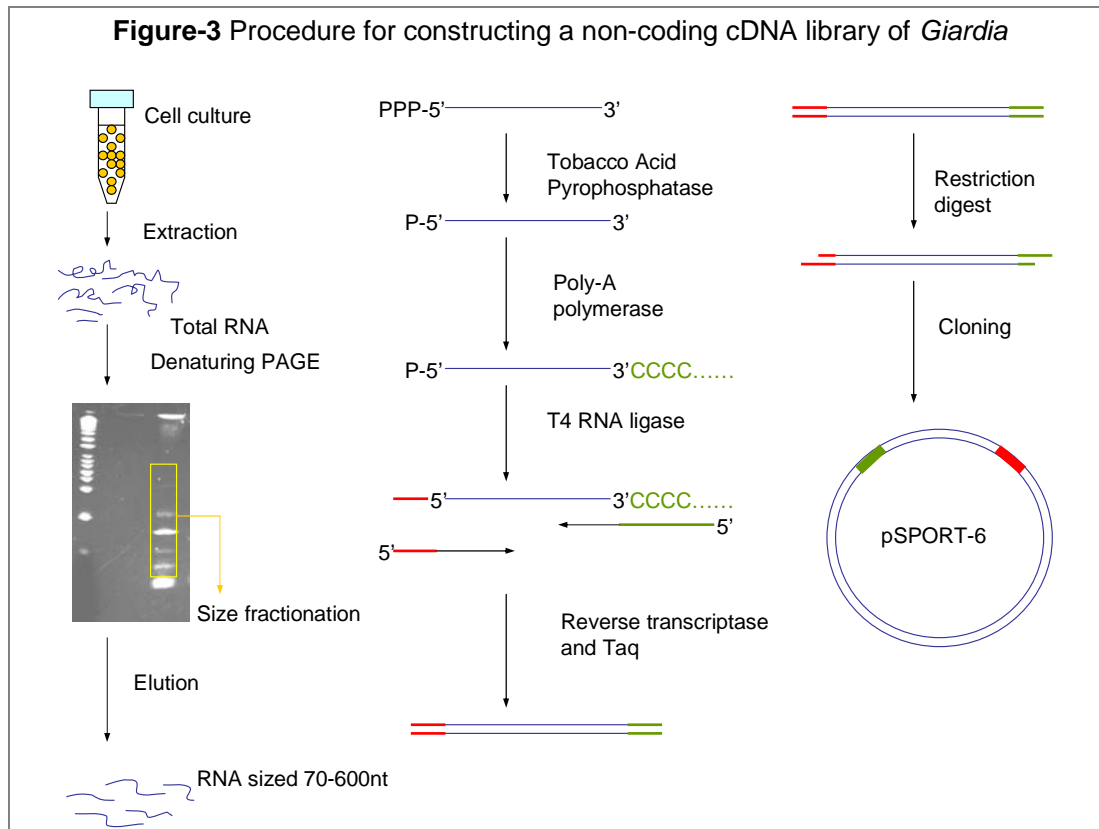
Direct RNA sequencing was used in the very early days of RNA research. By using a mixture of RNases which specifically cleave the radio-labelled RNA substrate at the 3'- end of G, C, U or A (Gupta and Randerath 1977b; Gupta and Randerath 1977a), the sequence of RNA can be determined on a denaturing polyacrylamide gel after autoradiography. Similarly for chemical sequencing, four base-specific chemical reactions generate a means of directly sequencing end-labelled RNA, followed by a partial, specific modification of each kind of RNA base, and an amine-catalyzed strand scission generates labelled fragments whose lengths determine the position of each nucleotide in the sequence (Peattie 1979). Early studies have sequenced tRNAs and rRNAs

using enzymatical or chemical sequencing (Yarus and Barrell 1971; Ehresmann et al. 1977). Direct RNA sequencing has also been used for characterization of small nucleolar RNAs in yeast and vertebrates (Balakin et al. 1996). Direct RNA sequencing requires excision of single bands after separating total RNAs on gels, thus it is very likely to introduce other RNA species of same/similar size and result in ambiguous data. This problem is usually solved by running 2D gels. The applicability of sequencing a particular RNA is also dependent on its size and abundance due to the capacity and resolution of acrylamide gels. However, compared with sequencing cDNA clones, direct sequencing does not require reverse transcription. Therefore it avoids any problems associated with reverse transcription caused by RNA secondary/tertiary structures.

Generation of cDNA libraries is the most widely used method for identifying ncRNAs. Isolation of ncRNAs is based on size. Since most mRNAs have sizes greater than 500nt, RNA sized between 20nt to 500nt is isolated by denaturing polyacrylamide gel electrophoresis for general downstream cDNA synthesis. Alternatively RNAs can be selected based on their ability of binding to proteins through immunoprecipitation (Vitali et al. 2003). Three experimental approaches have been developed for generating cDNA libraries. The first two methods involve addition of oligo(C) or oligo(A) tails to the RNAs by poly(A) polymerase and the third one involves addition of adaptors at both ends of the RNAs (Martin and Keller 1998). In the first method, tailed-RNAs are reverse transcribed using oligo(G) or oligo(T) primer, followed by second-strand synthesis with DNA-polymerase I (Huttenhofer et al. 2004). Subsequently double-stranded DNA linkers are ligated to the cDNAs, which are then cloned into a standard vector system (Huttenhofer et al. 2004). In the second method, RNA samples are treated with tobacco acid pyrophosphatase (TAP) which cleaves the 5'- cap structure of some ncRNAs (Aspegren et al. 2004). Subsequent to 3'- tailing, an oligonucleotide linker carrying a 5'-hydroxyl group is ligated to 5'- end of ncRNAs by T4-RNA ligase (Aspegren et al. 2004). The modified RNAs are then reverse transcribed and cloned. In a third method, RNA oligonucleotide linkers are ligated to both the 5'- end and 3'- end

by T4 RNA ligase; the oligonucleotide at the 5'- end of the RNA lacks a phosphorylated 5'- end, while the oligonucleotide at the 3'- end of the RNA contains a blocked 3'- end (Huttenhofer et al. 2004).

In this study, an ncRNA library of *Giardia* has been made using the second method described above. Figure-3 outlines the procedure of cDNA library construction.



In general, size-selected cDNA libraries enable large-scale identification of ncRNAs by high-throughput sequencing. However the abundance of individual cDNA clones depends on the nature of individual RNAs: less structured/modified RNAs are more easily reverse-transcribed, resulting in more abundant cDNA clones; cDNA clones of smaller size are more abundant than those of larger size, since smaller RNAs are more efficiently reverse-transcribed than larger ones. These obstacles can be overcome by sequencing a large number of cDNA clones (often thousands), or by hybridization using radio-labelled oligonucleotides targeting the most abundant known ncRNAs to exclude them before sequencing.

Microarrays are the favoured method to study levels of expression of many genes in parallel. To date, single-stranded DNA oligonucleotides of 25-70nt are the predominant type of probe, and are generally labelled with fluorescent dyes (Stoughton 2005). Early microarray studies have conducted extensive analysis of *E. coli* transcriptome (Selinger et al. 2000; Tjaden et al. 2002; Zhang et al. 2003), and aided verification of computationally predicted ncRNAs in other bacteria (Pichon and Felden 2005). Global analysis of ncRNA expression has also been applied to eukaryotes in prediction and annotation of ncRNAs (Peng et al. 2003; Inada and Guthrie 2004; Bentwich et al. 2005; Hiley et al. 2005).

The above techniques allow identification of ncRNAs from a pool of transcribed cellular RNAs by direct sequencing, cloning or microarray analysis. The fourth technique can identify ncRNAs through their ability of binding to proteins without isolating RNA transcripts from *in vivo*. This approach is termed genomic SELEX (Singer et al. 1997), and uses *in vitro* transcribed RNAs that are derived from a library of an organism's entire genomic DNA. Genomic SELEX has been used to select mRNAs that bind to certain protein partners (Shtatland et al. 2000; Kim et al. 2003). However this method is not widely used in identification of ncRNAs. The advantage of genomic SELEX is the possibility of identifying low-abundance RNAs that are overlooked by methods that require isolation of RNAs with certain levels of expression. And also, identification of RNAs through this method is not dependent on the developmental stages of an organism.

Identifying ncRNAs through the above methods is the first step towards characterising their functions. The ncRNAs identified at this stage are generally termed candidates before further biochemical analysis is done to confirm their biological functions. Several approaches are generally used in the study of RNA functions.

(1) Analysis of RNA-protein interactions and RNA-RNA interactions

Most ncRNAs are part of ribonucleoprotein particles (RNPs), and bind to specific protein partners. Analysis of RNA-protein complexes can hint towards

the function of ncRNAs based on the known functional domains of proteins. ncRNAs have been used as a “bait” to fish for RNA-binding proteins in cell extracts. For *in vitro* analysis, RNA can be synthesized by T7 RNA polymerase with an “affinity-tag” such as incorporating biotinylated UTP, and the biotinylated RNA is attached to a streptavidin coated solid support and the ncRNP can be isolated by using the ncRNA as a bait (Bardwell and Wickens 1990). For *in vivo* analysis, the yeast three-hybrid system* (Hook et al. 2005) has been developed to study RNA-protein interactions (Zhang et al. 1999; Bernstein et al. 2002). Many ncRNAs target RNAs through antisense binding. Target RNAs include mRNAs, ribosomal RNAs, tRNAs and snRNAs. Finding RNA targets can be done either computationally (Krek et al. 2005; Lewis et al. 2005), or experimentally (Lim et al. 2005).

(2) Expression patterns and structural analysis

Analysis of the expression patterns (cellular/subcellular localizations) of newly identified ncRNAs can hint towards their functions. Fluorescent *in situ* hybridization techniques are used to visualise the location of ncRNAs (Vitali et al. 2005). Subcellular localization of ncRNAs can also be studied by northern blot and RT-PCR. In addition, total RNA extracted from different developmental stages can be extracted and analysed for difference in ncRNA content. Expression patterns can also be analysed through computational method based on available genomic information. Structural analysis is an important computational method for the study of ncRNAs. Currently available structural analysis tools involve single RNA structure prediction (Hofacker 2003), pairwise RNA structural alignment (Havgaard et al. 2005) and multiple RNA structural alignment (Kiryu et al. 2007; Torarinsson et al. 2007).

In this thesis, a total of 38 novel ncRNAs have been found in the *Giardia* non-coding cDNA library and subjected to further analysis. Most of these ncRNA candidates do not show any recognizable features of known types of ncRNAs. The genome of *Giardia* is fully sequenced (McArthur et al. 2000; Morrison et

* Yeast three-hybrid system (Hooks et. Al. 2005): In this system, an RNA sequence is tested in combination of an RNA-binding protein linked to a transcription-activation domain. A productive RNA-protein interaction activates a reporter gene *in vivo*. This system has been used to test candidate RNA-proteins.

al. 2007) and the majority of protein-coding sequences, as well as some ncRNAs (rRNAs, tRNAs) are predicted. However the functions of many proteins are unknown and the major RNA-processing pathways have not been studied, which makes biochemical studies of unknown ncRNAs very difficult. In this chapter, analysis of the novel ncRNA candidates is mainly done by computational methods based on genomic information and published *Giardia* ncRNAs available to date.

3.2 Analysis of the novel ncRNA candidates

3.2.1 Novel ncRNA candidates from *Giardia* cDNA library

Following the first study detailed in Chapter-2, a further 576 clones were sequenced. All together a total of 38 novel ncRNA candidates have been found from *Giardia* cDNA library from a combined pool of 1192 sequences, including the new and updated ncRNA sequences from the earlier study (Chen et al. 2007). Among all 38 candidates, five have been characterized, including three C/D-box snoRNAs, one H/ACA-box snoRNA, and RNase P (Chen et al. 2007). Table-1 summarizes information associated with the ncRNA candidates. All candidates identified in this study are named GncR. The GC content of the whole genome on average is 46.8%.

Table-1: ncRNA candidates from *Giardia* cDNA library

GncR candidate	Length	<i>Giardia</i> genome contig	Start	End	Annotation	Copy No. in genome	GC content
GncR1	163	ctg02_1	834557	834395	none	1	61%
GncR2	150	ctg02_11	93926	93777	none	1	60%
GncR3	92	ctg02_9	223195	223104	none	1	71%
GncR4	95	ctg02_6	75591	75497	none	1	54%
GncR5	64	ctg02_17	151603	151666	C/D-box snoRNA	1	44%
GncR6	106	ctg02_22	85992	86097	none	1	60%
GncR7	121	ctg02_4	139696	139816	5'-P/GlsR15sno-3' fusion	1	52%
GncR8	61	ctg02_57	42770	42710	C/D-box snoRNA	1	61%

GncR9	114	ctg02_3	370386	370499	none	1	59%
GncR11	42	ctg02_26	142067	142026	none	1	50%
GncR12	136	ctg02_2	54704	54569	none	1	52%
GncR13	65	ctg02_4	253747	253811	C/D-box snoRNA	1	38%
GncR14	87	ctg02_82	3835	3921	likely a U	1	57%
GncR15	60	ctg02_1	472843	472902	none	1	45%
GncR16	70	ctg02_2	391655	391724	none	7	56%
	70		391877	391946			
	70		392099	392168			
	70		392321	392390			
	70		392542	392611			
	70		392764	392833			
	70		392986	393055			
GncR17	140	ctg02_188	520	381	none	2	66%
	140		769	630			
GncR18	54	ctg02_11	123685	123738	none	1	57%
GncR19	90	ctg02_34	44	133	none	4	53%
	90	ctg02_3	470373	470284			
	90	ctg02_14	2025	2114			
	90	ctg02_13	254452	254363			
GncR21	42	ctg02_4	98476	98435	none	1	52%
GncR22	66	ctg02_5	59134	59199	none	1	39%
GncR23	62	ctg02_24	26813	26752	none	1	60%
GncR24	41	ctg02_21	66492	66532	none	1	49%
GncR25	33	ctg02_9	223218	223186	none	1	78%
GncR26	66	ctg02_4	314811	314746	none (has a poly-A tail)	1	47%
GncR27	16	ctg02_17	56045	56030	none	1	30%
GncR28	16	ctg02_54	4750	4765	none	1	56%
GncR29	113	ctg02_29	40719	40831	H/ACA-box snoRNA	1	57%
GncR30	80	ctg02_67	17501	17422	none	1	51%
GncR31	60	ctg02_24	90239	90298	none	1	57%
GncR32	71	ctg02_3	378369	378299	none	1	59%
GncR33	136	ctg02_11	95569	95704	none	1	59%
GncR34	72	ctg02_11	140521	140592	none	1	38%
GncR35	86	ctg02_26	5389	5304	none	1	52%
GncR36	43	ctg02_14	199119	199161	none	1	51%
GncR37	77	ctg02_22	154101	154177	none	1	48%

GncR38	96	ctg02_21	174660	174565	none *	1	69%
GncR39	110	ctg02_1	43393	43284	none	1	46%
GncR40	109	ctg02_47	24067	23959	none	1	51%

Within these novel candidates, 13 out of 38 are located on the minus strands of ORFs (GncR28-GncR40). This is consistent with the observation that *Giardia*'s transcribed sequences are rich in antisense transcripts (Elmendorf et al. 2001a). Most of the candidates are transcribed as single copy genes located between ORFs. Analysing the upstream elements and potential promoter elements of known ncRNAs will aid characterising the unknowns. Therefore, the upstream sequences (100nt) of previous identified tRNAs, snoRNAs and RNase P have been pulled out from the genome database, and analyzed as three types of upstream sequences. In addition, potential internal promoter elements of tRNAs and other ncRNAs are also analysed. This is discussed next.

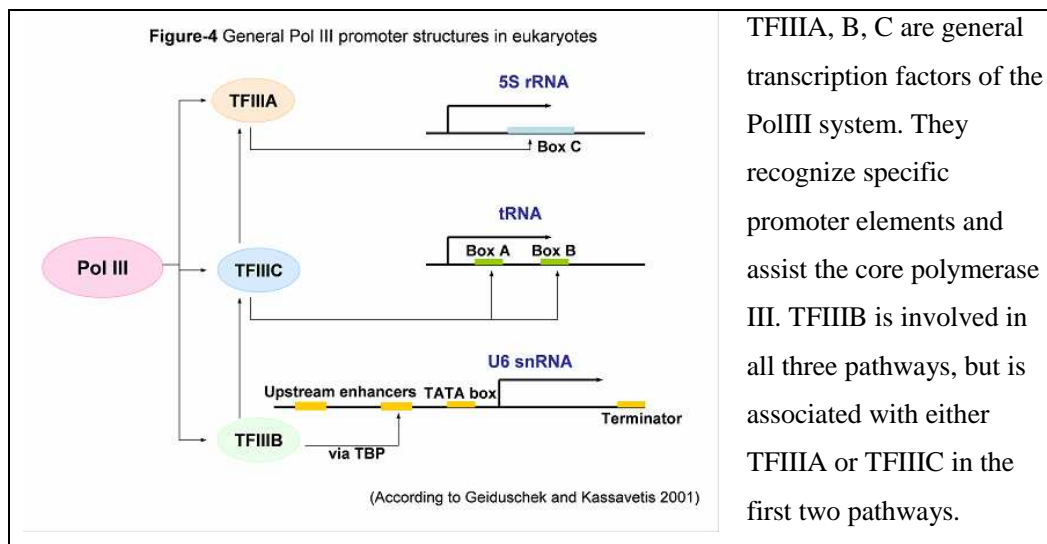
3.2.2 Analysis of potential promoter elements of characterised ncRNAs from *Giardia*

Both RNA polymerase II (Pol II) and polymerase III (Pol III) are involved in transcription of eukaryotic ncRNAs. Pol III transcribes DNA encoding 5S rRNA, tRNAs and other small RNAs of eukaryotes. However spliceosomal snRNAs are transcribed by Pol II except U6 snRNA. Compared with Pol II, Pol III uses fewer regulatory proteins. In most organisms, transcription by Pol III does not require upstream promoter elements; instead, internal promoter elements are recognised. Genes transcribed by Pol III can be divided into three classes based on which transcriptional factor is involved in promoter element recognition (Geiduschek and Kassavetis 2001). Class 1 genes (5S rRNA genes) require direct promoter recognition by TFIIA (Sakonju et al. 1980; Lee et al. 2006). Class 2 genes (tRNAs and other diverse other genes with similar promoter elements) are recognized by TFIIC (Galli et al. 1981); Class 3 genes (U6 snRNAs, RNase P, MRP, 7SK) have a different promoter structure which contains one essential upstream promoter element (PSE) and a dominant upstream enhancer element (Kunkel and Pederson 1988; Jensen et al. 1998).

* In addition to the candidate GncR38, another 4 sequences exist in the genome and their sequences are only differed from GncR38 by a few base-substitutions. It is not known whether these polymorphic forms of GncR38 are transcribed.

There is no 5S rRNA found in *Giardia* (Edlind and Chakraborty 1987), and the presence of U6 snRNA is not certain. To date there are 44 tRNA genes found in the genome of *Giardia* (McArthur et al. 2000; Morrison et al. 2007), and the presence of RNase P has been confirmed (Marquez et al. 2005). Therefore, it is possible to obtain useful information about Pol III transcription and hence help to predict Pol III transcribed ncRNAs by analysing their internal promoter elements and upstream elements.

Internal promoters are the most distinct features of genes that are transcribed by Pol III. Figure-4 shows the general model of the three types of Pol III promoter structures and recognition of essential elements by Pol III transcription factors.



The best studied Pol III polymerase from *S. cerevisiae* contains a core promoter consisting 17 genes, and three transcription factors: TFIIIA, TFIIIB, and TFIIIC. These molecules constitute the essential Pol III transcription apparatus (Geiduschek and Kassavetis 2001). Binding of TFIIIA and TFIIIC implies displacement of the core polymerase at each round of RNA synthesis, but Pol III promoters do not need to be newly marked for every successive round of transcription; to increase the efficiency of transcription TFIIIC places the third initiation factor TFIIIB upstream of the transcription start site, and TFIIIB is able to repeatedly recruit the RNA polymerase to the promoter (Kassavetis and Geiduschek 2006). In addition, TFIIIC also interacts with a

subunit common to Pol I, Pol II and Pol III (Kassavetis and Geiduschek 2006). Table-2 details the features of Pol III subunits. The published *Giardia* Pol III transcription factor subunits are indicated by “√”. The currently unidentified subunits are indicated by “×”, and the possible subunits are also indicated.

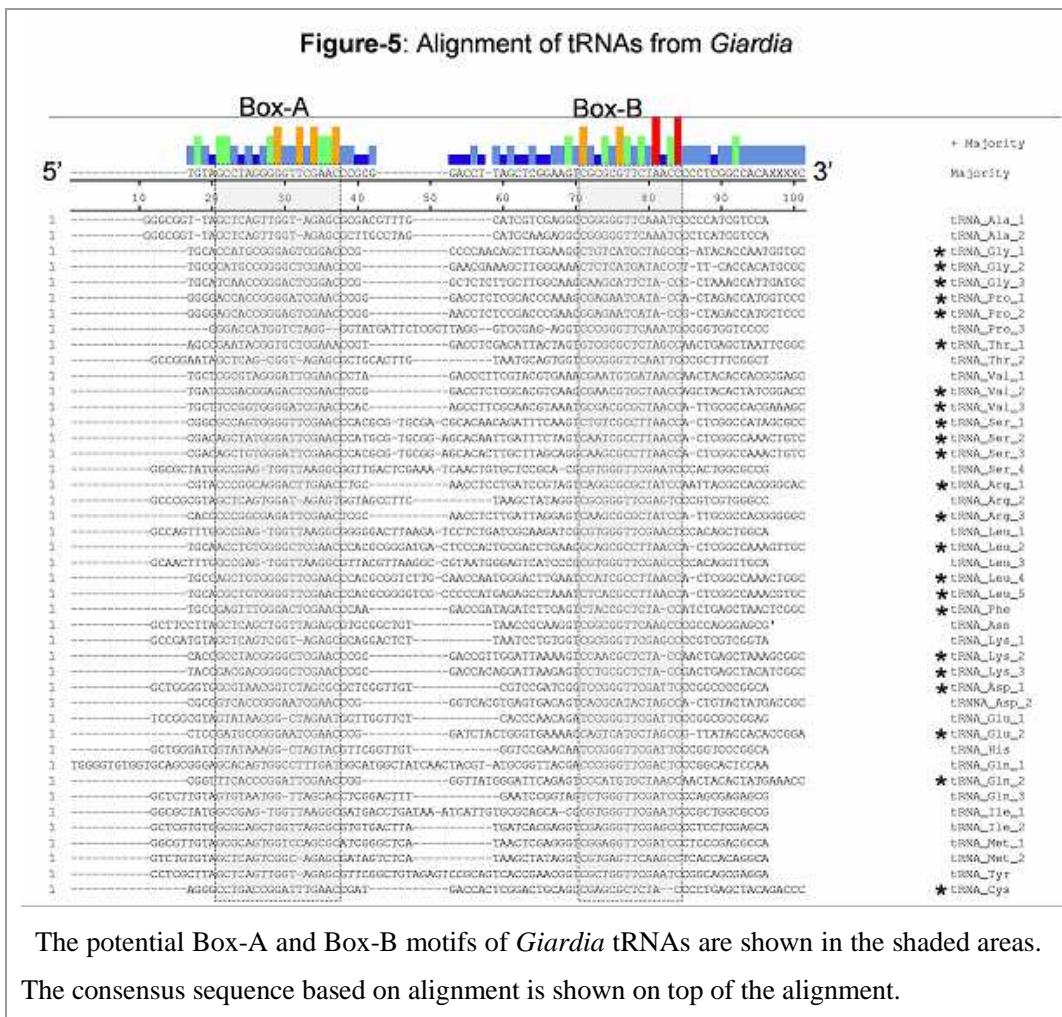
Table-2: Transcription factors of eukaryotic RNA Pol III system and whether they have been identified in *Giardia*

Factors	Components		
	Human	<i>S. cerevisiae</i>	<i>Giardia</i>
TFIIIA	hTFIIIA (9 zinc finger proteins)	Tfc2 (10 zinc finger proteins)	×
TFIIB	TBP	TBP	√
	Brf/TFIIB90	Brf	√
	BrfU/TFIIB50		×
	Brf2		×
		Tfc5	×
TFIIIC		Tfc3 (Box-B-binding subunit)	possible
	TFIIIC-220/C2 α (Box-B-binding subunit)		×
	TFIIIC-102/C2 γ	Tfc4	possible
	TFIIIC-63/C2 ϵ (Box-A-binding motif)	Tfc1 (Box-A-binding motif)	possible
	TFIIIC-110/C2 β	Tfc6	×
	TFIIIC-90/C2 θ	Tfc8	×
		Tfc7	×

Internal promoter elements generally show considerable degree of sequence conservation. Flanking sequences upstream of the transcription start can also affect the activity of many Pol III promoters (Geiduschek and Kassavetis 2001), however, the precise sequences are normally not conserved even between different Pol III regulated genes of the same organism. Nevertheless, certain upstream elements are shown to be generally important. It has been shown that a TATA-box like element located at 25-30nt upstream of tRNA genes in insects is essential for tRNA transcription (Trivedi et al. 1999; Ouyang et al. 2000). TATA-boxes are also present in the promoters of *Schizosaccharomyces pombe* Pol III genes (Geiduschek and Kassavetis 2001). In the case of U6 snRNA transcription, the presence of a TATA-box determines transcription by

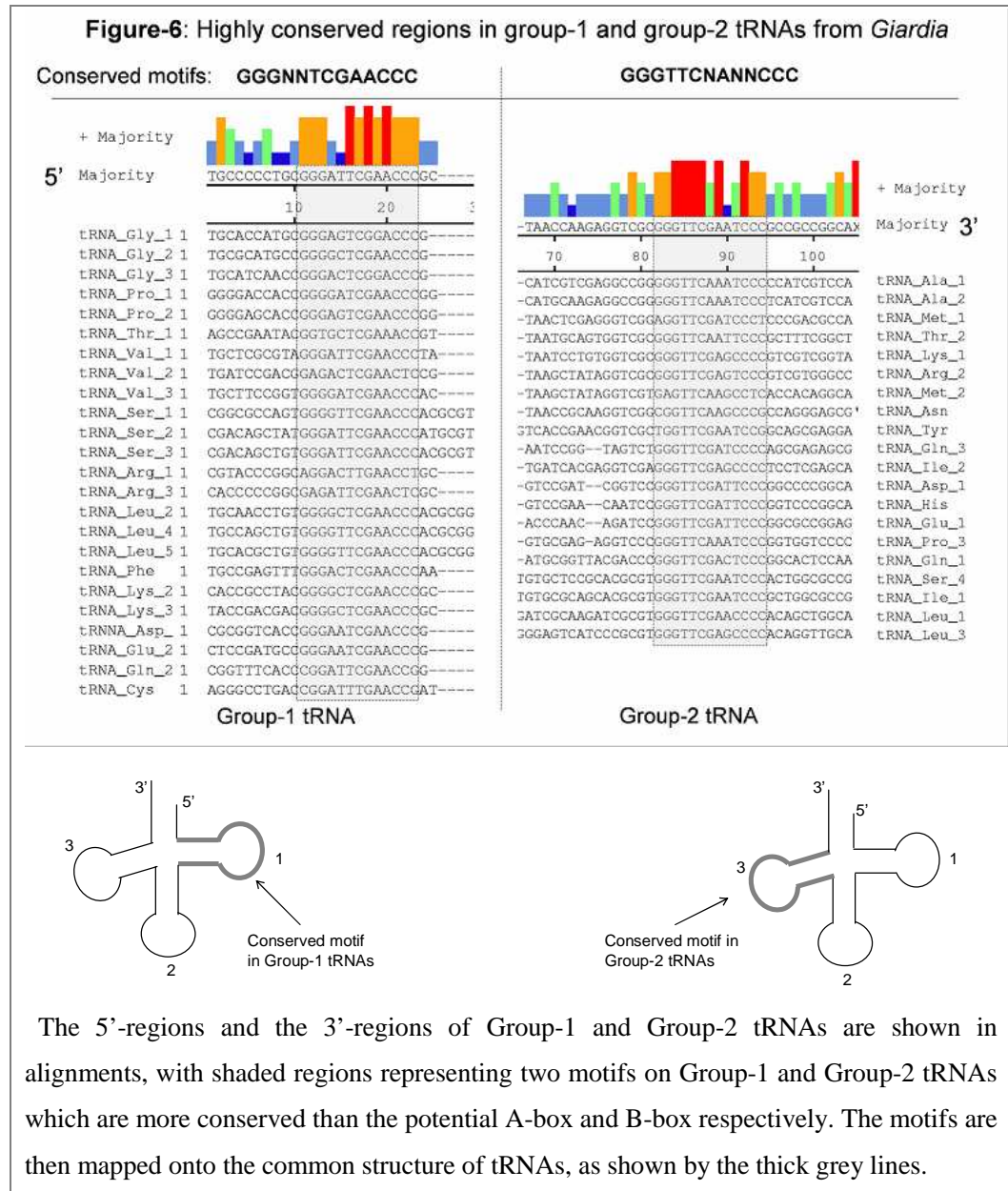
Pol III, whereas U2 which is transcribed by Pol II does not have TATA-box (Geiduschek and Kassavetis 2001).

For analysis of Pol III promoters in *Giardia*, the sequences of 44 tRNA genes have been extracted from the genome database, and aligned by multiple sequence alignment ClustalW. Results show potential Box-A and Box-B motifs, separated by 20-35nt (Figure-5). However, careful examination of the sequence alignment by eye revealed that the 44 tRNA genes can be divided into two separate groups based on their conserved sequence motifs (Figure-6).



Multiple sequence alignments show, that group-1 tRNAs (24 in total, indicated by * in Figure-5) have a compact 12nt conserved motif (GGNNTCGAACCC, with one or two mismatches) near the 5' ends (Figure-6) and a more dispersed motif near the 3' ends (not shown). In contrast, group-2 tRNAs (20 in total) have a more dispersed motif near the 5' ends (not shown) and a compact 13nt conserved motif (GGGTTCNANNCCC, with one or two

mismatches) near the 3' ends. Mapping the conserved motifs onto the tRNA consensus structure reveals that two motifs correspond to the first and third stem-loop respectively, as indicated by thick grey lines (Figure-6).



The conserved motifs and their corresponding folds on tRNA structure suggest an evolutionary divergence of the putative group-1 and group-2 tRNAs of *Giardia*. It is noticed here that some tRNAs (for example tRNA_Gly and tRNA_Ala), all the isoforms are either group-1 or group-2 tRNAs; whereas other tRNAs (for example tRNA_Glu and tRNA_Leu), isoforms are found in both group-1 and group-2 tRNAs. The pattern could mean a different evolutionary history of these tRNAs, with some tRNAs evolved solely within

one group and others diverged into two groups. However highly conserved regions are only found in the first (group-1 tRNAs) or third stem-loop (group-2 tRNAs) regions but not in the second stem-loop where anticodons are located. This pattern is expected because the loop of anticodon characterises each tRNA. In all the highly conserved regions of group-1 and group-2 tRNAs shown here are likely to be associated with tRNA transcription and evolution of tRNAs in *Giardia*.

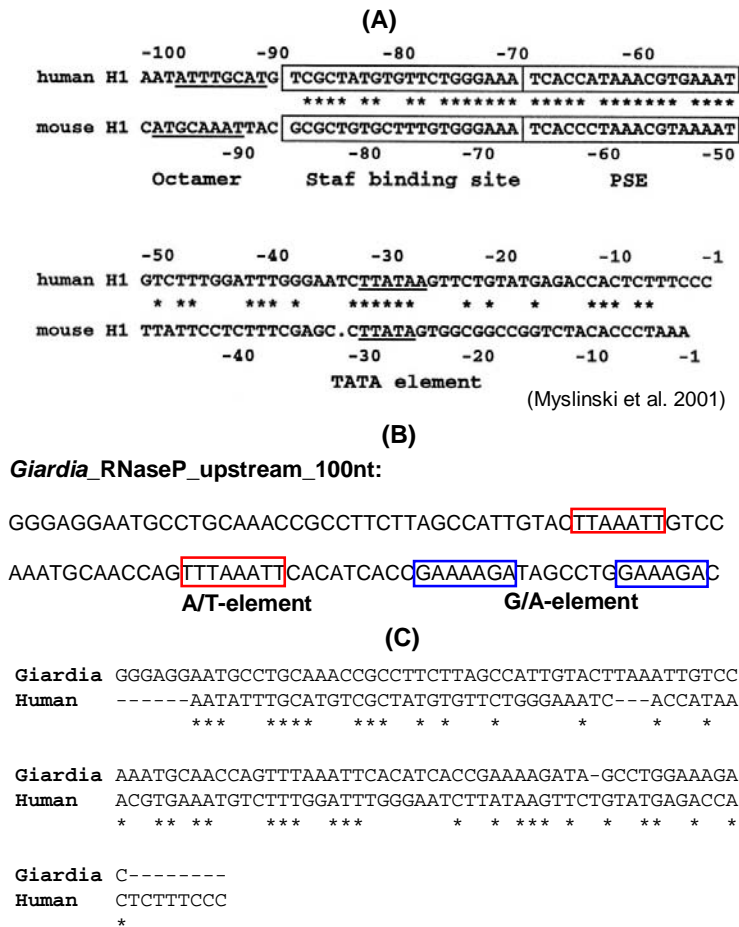
According to a genomic survey of transcriptional proteins in *Giardia* genome (Best et al. 2004), only 8 out of the 23 basal transcriptional factors of eukaryotes are found in *Giardia*, and within the 8 transcription factors, only two of the three subunits of TFIIB: TBP (TATA-box binding protein) and Brf (TBP related factor) are involved in Pol III transcription (see Table-2). Phylogenetic analysis showed that *Giardia* TBP is highly divergent with respect to both archaeal and other eukaryotic TBPs, and contains substitutions of key residues important for TATA-element binding (Best et al. 2004). Therefore, it is likely that transcription initiation of *Giardia*'s Pol III genes is independent of TFIIC and TFIIA, and TATA-box binding of *Giardia* TBP may differ from the general mechanism found in model eukaryotes. Analysing the upstream sequences of tRNAs shows that a number of tRNA upstream sequences have an A/T-rich region located about 10 to 20 nt 5' to the transcription start sites (see Appendix-2), TBP is likely to be involved in tRNA transcription. Lack of TFIIC suggests that the potential internal promoter elements of *Giardia* tRNA are not essential in transcription, or recognition of these elements may be associated with other yet unknown transcription factors. However due to the long phylogenetic distance of *Giardia* to other model eukaryotes, the possibility of TFIIC subunits existing in *Giardia* cannot be excluded.

Following the analysis of potential tRNA promoters, the upstream sequence of RNase P was analysed. In eukaryotes, the RNase P promoter is similar to the snRNA-type promoter, which has class-3 (Figure-4) basal promoter elements (Baer et al. 1990). Efficient basal expression of snRNAs requires a TATA-box

element between -30 to -25, which is the major determinant of Pol III specificity, and a proximal sequence element (PSE) located between -66 to -47, which recruits a five-subunit protein complex known as SNAPc or PTF (Kunkel and Pederson 1988). Activated transcription of snRNA-type genes is also associated with a distal sequence element (DSE), which normally locates between -260 and -190 (Myslinski et al. 1993). However, transcription studies showed that the sequence elements that are required for the transcription of human and mouse RNase P genes lay entirely within 100bp 5' to the transcription start site, with the PSE and TATA-box motifs absolutely required *in vivo* (Myslinski et al. 2001). The upstream sequence of *Giardia* RNase P contains two A/T-elements, which are potential TATA-box-like elements, and there are two G/A-elements located closer to the transcriptional start site. The G/A-element is a short region near the putative transcription start site which usually contains a "G" followed by three to eight "A"s with occasional insertion of a pyrimidine (Figure-7B). The potential G/A-element is also seen in a number of snoRNAs in *Giardia* (see Figure-8).

As shown in Figure-7(C), the upstream sequences of human and *Giardia* RNase P are rather conserved. Although the consensus pattern of this alignment is different from the highly conserved alignment of human and mouse RNase P upstream sequences, the evolutionary divergence between human and *Giardia* makes it difficult to observe strongly conserved motifs. However, the potential upstream elements: the "A/T-element" and "G/A-element" seen in many ncRNAs from *Giardia* (discussed later) suggests their possible functions as being binding sites for transcription factors such as TBP.

Figure-7: Comparison of RNase P upstream sequence elements from *Giardia* and mammals



(A) Alignment of human and mouse RNase P upstream 100nt sequences. The conserved regions contain the important Staf transcription factor binding region, the proximal sequence element (PSE), and the “TATA-box”. (B) Upstream 100nt sequence of *Giardia* RNase P. The potential motifs “A/T-element” and “G/A-element” are indicated in by red boxes and blue boxes respectively. (C) Alignment of human and *Giardia* RNase P upstream 100nt sequences.

The following describes potential upstream promoter elements in snoRNAs. Unlike many eukaryotes, where snoRNAs are either encoded as introns (Selvamurugan et al. 1995) or polycistronic repeats (Dunbar et al. 2000; Brown et al. 2001; Liang et al. 2002; Huang et al. 2004), the snoRNAs found in *Giardia* to date seem to have their own and variable mechanisms of expression. As described earlier (Chen et al. 2007), most of *Giardia*’s snoRNAs are located as single copies between protein-coding genes, and a few of them are located on the minus strands of ORFs. This allows the possibility that either they are

co-transcribed with adjacent protein-coding genes and subsequently cleaved by a yet known mechanism, or they are transcribed separately by either Pol II or Pol III. How the transcription systems in *Giardia* work is far from clear. The promoter sequences of protein-coding genes of *Giardia* are not very well conserved. An early study on seven *Giardia* cytoskeleton genes (Holberton and Marshall 1995) showed that none of the sequences appeared to have a TATA-box. However they contained an A-rich element (CAAAA/TA/CT), which was similar to the hexamer-element (AAAAAT) of a TATA-less promoter in mouse (Hariharan and Perry 1990). According to the information from the first release of *Giardia* genome database (McArthur et al. 2000), the *Giardia* Pol II promoter contains two key upstream elements: the -20 to -35 promoter region (CAAAA[AT][TC]AGA[GT]TC[CT]GAA), and the -40 to -70 promoter region (CAATTT). And also, the Pol II transcription start site is rather strongly conserved, and is marked by a poly-A region (AAT[TC]AAAA). This information on Pol II promoter has remained the same in the new release of *Giardia* genome (Morrison et al. 2007).

Following the above results, I analysed the upstream sequences of *Giardia* snoRNAs to search for potential promoter signals. Among the 25 snoRNAs (Yang et al. 2005; Chen et al. 2007) studied here, three of them have Pol II promoter-like upstream sequence elements, eleven have upstream sequence elements similar with RNase P, the remaining eleven do not show obvious promoter signals. Examples of three categories of upstream sequences of snoRNAs are shown in Figure-8. The first category of snoRNAs has Pol II-like upstream elements as shown by coloured boxes. The potential promoter elements are characterised by sequence similarity with the consensus Pol II promoter elements found in protein-coding genes of *Giardia*. Although the potential elements do not always match perfectly to the published consensus sequences (Hariharan and Perry 1990; McArthur et al. 2000), the likelihood that these snoRNAs are transcribed by Pol II system is strengthened by the presence of an A-rich region directly preceding the putative transcription start sites.

Figure-8: Three categories of *Giardia* snoRNA upstream sequences**Category-1:** Possible Pol II type promoter

♣Upstream_GlsR1_C/D-box_snoRNA (Yang et al. 2005):

CGCCAGATTGCTTAGCAAGCAGCTTTTGAGAAGCACTCAATGTA AATCATATGTTCAAAAAAGCAAATTAATTCCGCTTCTGATTTCATATAAAITTTCAA

♣Upstream_GlsR2_C/D-box_snoRNA (Yang et al. 2005):

GCGAGACATGCTTTTTGTCTGACTCGGATTTGTGCTGAATTACATGTTGCTATTTATGAAACAAAGCTCACGCATACACAGGCTCCGGAAAAATAAA

Category-2: Possible Pol III type promoter

♣Upstream_GlsR8_C/D-box_snoRNA (Yang et al. 2005)

AGCATCTGGCGAGGAAATGTATTACGCCATAATGTGATCAAAAGATAC TTTAAAAATAGATTGATTTTAAATTCACTTTGCAGCTCACAGAAAAGGGTT

♣Upstream_GlsR19_H/ACA-box_snoRNA (Yang et al. 2005)

TTTCGGACATAAAGTAGCAGCAACCAAGGACTAAGTACGCCATGCTTCAGTACTTATAAAAGCTCCTTCTTATAAGTGAACAAAAATTTCTTTTCGTCA

♣Upstream_GncR13_C/D-box_snoRNA (Chen et al. 2007)

TTTACTGCAAGTTACTAGGCAGCAAGTTCAAGTCTGGGAACCGAGATCGTTTCAAAAACGGTTTTAAAAAGCTCCGAAGCAAATGAGAACAAAAGCAGAC

♣Upstream_GncR29_H/ACA-box_snoRNA (Chen et al. 2007)

GGGCGAGCGGACATTTAACCTACGCACGGAATCTATAGATGTCTCCAGAATCAAAATTAATGGATTGCTTCTTAAAAATGATGGCCGGAAGAGAAAAAGA

Category-3: No obvious signal of either Pol II or Pol III type promoter

♣Upstream_GlsR20_H/ACA-box_snoRNA (Yang et al. 2005)

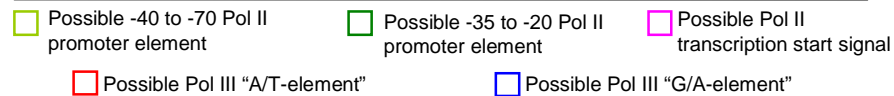
TAACAAAGTCTCCATCTCTACGGCTGGGAACTCATGACTCAGAGTGATGTGTGCGTACGCCATCCAGTTTGATAGGGGGTTCTTTTCTTTTTCGCAAGTT

♣Upstream_GlsR15_C/D-box_snoRNA (Yang et al. 2005)

GAGGAGGGCCTTGCCCGACTGAGAGTGCTCGCTGAAAGAGGCTGCGACGCGGGTTATTCAGTTCGATGCGCCCAGGCTGACGGTAGGACGCCTAACCC

♣Upstream_GncR5_C/D-box_snoRNA (Chen et al. 2007)

GATTCCTCGACCCCTGGTAGGTATACTTTGTGCGGACTAGAAACGAACTAGAAAATCAGTAAAAAGGTCTTGAGCAAACCCAGTAAATTAATAATGATTA



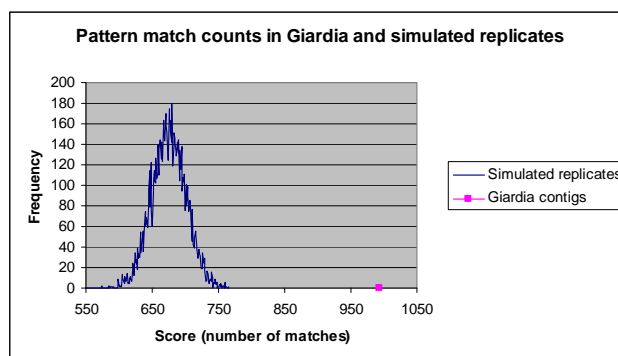
The second category of snoRNAs has upstream elements similar with those of RNase P as discussed before. These upstream sequences generally contain one or two A/T-rich segments (as boxed in red in Figure-8) located before a G/A-element. This category of snoRNAs is classified mainly on the presence of the G/A-element, which appears to be more conserved than the A/T-element. There is a third category of snoRNAs which do not have a consensus pattern of

nucleotide sequence distribution as shown in Figure-8, where three examples are given. The A-rich and T-rich regions highlighted by coloured lines suggest that this category of snoRNAs may not be transcribed as single transcripts as there is no signal of any consensus promoter elements. Therefore they may be co-transcribed with adjacent ORFs.

To further test the likelihood of the combination of A/T-element and G/A-element being specific upstream feature of *Giardia* Pol III promoter, the pattern (as shown below) of this combination was searched in the genome.

Searching model:
 A/T-element: at least six nucleotides of only A and T
 G/A-element: GA-[3 to 6 "A"s with one substitution of "C" allowed]-AG
 Additional restriction: the A/T-element locates at most 20nt upstream of the G/A element.

This search revealed 993 sites matching this pattern in the genome, but this estimation of the combination of A/T-element and G/A-element does not necessarily reflect the number of possible Pol III promoters in the *Giardia* genome. To test the selectivity of the search model, the following permutation test was applied. 10,000 synthetic DNA datasets of the same size as the entire *Giardia* genome (11,192,215 bases) were generated and searched for matching sites. Each dataset was produced by choosing bases randomly with probabilities corresponding to their frequencies in the *Giardia* genome, leading to sequences having nearly identical bases composition with the *Giardia* genome. The minimum, maximum and average match counts of the synthetic datasets were 574, 766 and 675.2 respectively. The standard deviation was 26.26. A histogram comparing the distribution of synthetic dataset match counts to the match count for the actual *Giardia* genome is shown. This figure clearly shows that the



number of matches found in the actual *Giardia* genome is far above the number expected by chance. The fact that the maximum match count across 10,000 synthetic dataset is less than 993 suggests that $p < 0.0001$.

It is very likely that snoRNAs in *Giardia* are expressed through different mechanisms with the possibility of being transcribed by either Pol II or Pol III transcription system, as well as being co-transcribed with adjacent protein-coding genes. The Pol II-like promoter elements seen in a few snoRNAs are less conserved than the elements upstream to protein-coding genes, however they still have recognizable patterns, suggesting that transcription of these RNAs are less tightly regulated compared with protein-coding genes. The putative Pol III promoter elements seen in RNase P and snoRNAs appear to be more conserved than the putative Pol II promoter elements of another small group of snoRNAs. The conserved appearance of the A/T-element and G/A-element are potential binding-sites for Pol III transcription factors. Although *Giardia* does not seem to have two of the three Pol III basal transcription factors (Best et al. 2004), TFIIB alone may be sufficient to initiate Pol III transcription. It has been shown with a minimal RNA Pol III transcription system that Brf and TBP alone assemble Pol III for transcription at a significant level (Kassavetis et al. 1999). It has also been shown using a TFIIC-less transcription system, that TFIIB alone is sufficient to direct efficient Pol III recycling on short (~100bp) Class III genes (Ferrari et al. 2004). Biochemical studies have shown that the TFIIB subunit Brf and Bdp play direct role in DNA melting at the transcriptional start (Kassavetis et al. 2003). Therefore the A/T-element may be responsible for recruiting *Giardia* TFIIB, which acts to open the promoter DNA structure and recruit Pol III.

So far it is fairly likely that many ncRNAs from *Giardia* are transcribed by the Pol III system, with a minority transcribed by Pol II system. Of the three classes of Pol III genes known to date, the third type of Pol III genes are likely to be the dominant type for *Giardia*'s ncRNA genes, due to possibly lack of TFIIA and TFIIC (see Table-2). Most characterised ncRNAs from *Giardia* have A/T-elements upstream to the potential transcriptional start, which indicates the importance of DNA melting, and TFIIB recruiting in

transcription initiation. Analysis of the upstream elements of characterised ncRNAs from *Giardia* show a number of frequently appearing types of sequences which, although they do not always obey a consensus pattern, may be potential binding-sites for transcription factors. It is expected that newly identified ncRNAs from *Giardia* may strengthen the likelihood of these potential upstream elements as being true promoter elements.

3.2.3 Analysis of the upstream sequences elements and internal sequence elements of novel *Giardia* ncRNAs

To further study the 33 uncharacterised novel ncRNA candidates from the cDNA library, a 300nt upstream sequence of each ncRNA candidate was pulled out from the genome and analysed based on the information drawn from the characterised ncRNAs discussed above. 9 of the 38 sequences contain upstream elements similar to that of RNase P, therefore are most likely to be transcribed by Pol III (Appendix-2). The rest do not show distinct upstream sequence elements which may classify them as Pol II genes or Pol III genes as discussed above.

The 9 predicted Pol III genes from the 38 ncRNA candidate genes all contain the A/T-element which is a potential binding site for *Giardia* TBP. Together with the G/A-element found downstream to the A/T-element, the upstream sequences of the 9 ncRNA candidates indicate a strong possibility that they may be transcribed by Pol III. However, in general not all Pol III genes have upstream promoter elements (see Figure-4). It is likely that some of the ncRNA candidates are transcribed by different Pol III mechanisms involving recognition of internal promoter elements by Pol III transcription factors that have not been identified in *Giardia*, and the possibility of being co-transcribed with adjacent protein-coding genes is not excluded.

The sequences of internal promoter elements of Pol III genes are highly variable. As mentioned before, There is little conservation even among different types of Pol III genes of the same organism (Geiduschek and Kassavetis 2001). However, within the same type of Pol III genes, there can be conserved sequence motifs (e. g. the potential internal promoter elements of tRNAs from *Giardia* as discussed above). Unlike tRNAs, multiple sequence

alignment cannot identify potential Pol III internal promoter elements from the novel ncRNA candidates found in the *Giardia* cDNA library. Therefore, more advanced motif finding algorithms were applied to search for potential internal promoter elements within the novel ncRNA candidates. Analysis of all the ncRNA candidates was done using the Gibbs Motif Sampling* Algorithms (Neuwald et al. 1995). Results obtained by using two different searching models based on eukaryotic default parameters from the web-server (which allow five motifs to be searched for one sequence input) showed several potential transcription factor binding sites which share distinct sequence homology. This procedure and results are described in more detail below.

Firstly, all novel ncRNA candidates were analysed using the Gibbs Motif Sampler (Neuwald et al. 1995), which begins with an alignment of motifs randomly spread throughout the sequences. The algorithm starts at the very first position in the long, concatenation of all sequences and checks to see if the position is a possible motif start site. Running the Motif Sampler gave 4 distinct potential motifs; each was shared among a number of sequences. The result is shown in Table-3a. As a comparison and control, 20 *Giardia* snoRNAs (Yang et al. 2005) were analysed using the Gibbs Motif Sampler, and the results showed two distinct motifs shared by a number of snoRNA sequences, as shown in Table-3b. As expected, the alignment of the two motifs contained the conserved C-box and D-box.

Second, all novel ncRNA candidates were analysed by the Gibbs Recursive Sampler (Thompson et al. 2003), which uses recursive sums over all possible alignments from 0 to a maximum in a sequence, to obtain Bayesian inferences on the number of sites for each motif and the total number of sites in each sequence. Results indicate the presence of another potential motif (Table-3c), but here is limited consensus information observed across the majority of the novel ncRNA candidates as shown in Table-3c. However, more conserved sequence motifs are observed among smaller groups of sequences as shown in Table-3a. These sequence motifs might either serve as transcription factor binding sites if these ncRNAs in *Giardia* are transcribed by Pol III or, on the other hand, these sequence motifs may also indicate yet unknown functions such as binding to certain protein factors.

* Gibbs Motif Sampler: <http://bayesweb.wadsworth.org/gibbs/gibbs.html>

Table-3: Sequence motifs in novel ncRNA candidates and 20 snoRNAs from *Giardia*


a) Sequence motifs within the novel ncRNA candidates (GncRs from this study) analysed by Gibbs Motif Sampler

Motif number	Motif information				
	Sequence name	Start position	End position	Motif alignment	Consensus
1	GncR2	98	108	atggt TGCAGGACAAG cttag	
	GncR9	19	29	ggaga TGCTGGACACG gcttt	
	GncR33	14	24	taaca TGCTTGCCACG gcgtc	
	GncR38	17	27	tccag TGCTGGCCAGG ggcaa	
2	GncR1	136	145	gcggg GAAGCCCTGC ggcgc	
	GncR14	17	26	aagag GCAGGCATGC aggat	
	GncR17	30	39	ccggc GAAGGTCTGC aagtg	
	GncR19	81	90	ggcat GCAGCCCTGC	
	GncR32	56	65	gcttc GCAGCTCTAC gggcg	
	GncR35	47	56	aactt GAAGCTCTGA tcggg	
	GncR39	41	50	cttaa GAACCTCTGC ttcta	
3	GncR9	45	54	cccac CGGAGCACAT atgct	
	GncR21	16	25	ctggc CGGAGCACAT ttgtg	
4	GncR2	79	88	agccg CCACACTGAC agtta	
	GncR7	8	17	tgcgc CCAGGCTGAC ggtag	
	GncR17	12	21	ataga CCAGGCTGCC agccc	
	GncR18	3	12	ta CCACTCTGAC cgtga	
	GncR30	53	62	cttgc CCAGTCTGCC tccat	

b) Sequence motifs within 20 snoRNAs (GlsRs: by Yang et al. 2005) from *Giardia* analysed by Gibbs Motif Sampler

Motif number	Motif information				
	Sequence name	Start position	End position	Motif alignment	Consensus
1	GlsR4	26	38	cgccg CCCAGTCTGACC cctga	
	GlsR5	75	87	caagc CAACCGGCTGAGC tc	
	GlsR7	46	58	ctcat AGTACTCTGAGC gg	
	GlsR8	58	70	accgc CTTTCGTCTGACC	
	GlsR13	55	67	acggc CGCCCGTCTTACC ttgtg	
	GlsR13	88	100	tctta CAATGCTCTGACC	
	GlsR16	63	75	cgcac CACCGCTCTGACC tt	
	GlsR17	25	37	taatg CGCTTCTTTGAGC cgcgg	
2	GlsR1	23	35	gaggc AGATGATGACTTT gcgac	
	GlsR4	6	18	tgtct CCATGACGAGAAT tacgc	
	GlsR5	10	22	aaaag CTGTGATGACAGG ttctt	
	GlsR7	4	16	ccg CGATGATTACCGA atcac	
	GlsR9	51	63	ttgca CGCTGATGAGTGA aagca	
	GlsR10	2	14	a GAATGATGAGACG tgttc	
	GlsR11	3	15	gg CGGTGATTAGGCT gcgtg	
	GlsR14	1	13	AAATGATGACAAT gcgca	
	GlsR16	6	18	taaaa CTATGATGAGGTT agcga	
	GlsR20	86	98	gatct GGGTATTAGCAG tcata	

c) Sequence motifs within the novel ncRNA candidates analysed by Gibbs Recursive Sampler

Motif number	Motif information				
	Sequence name	Start position	End position	Motif alignment	Consensus
1	GncR2	17	26	cgggc AGAAAGTGCC ggtcc	
	GncR3	18	27	cggac AGCCGGAGGC cggag	
	GncR4	3	12	ct AGGCTGAAGC tgcca	
	GncR5	38	47	tcttt AGACTGCTGA gacag	
	GncR6	31	40	gttca AGCCAGGTCC aagac	
	GncR7	40	49	gattc AGACTACTCC ttggt	
	GncR12	85	94	ctgtg AGGCAGCTGC cagga	
	GncR12	96	105	ctgcc AGGATGGTCC tgccc	
	GncR13	14	23	aatga AGACAGAACC acaga	
	GncR14	9	18	cctag AGGAAGAGGC aggca	
	GncR14	64	73	gcagc AGAGAGTGGC cacgc	
	GncR16	15	24	agaaa AGACGCGTGC gaggc	
	GncR16	26	35	gtgcg AGGCGGTTC caaca	
	GncR16	62	71	gctgc AGAATGCGGC	
	GncR17	51	60	gacgg AGACAATGGC tacac	
	GncR17	89	98	cacca AGGCGGCTCC tgaca	
	GncR18	17	26	ccgtg AGGCGCATGC ctagg	
	GncR24	1	10	AGACAGAAGT agagc	
	GncR25	6	15	cgcgg AGGCAGGGGC cggcc	
	GncR29	10	19	aagca AGGCTAGAGC catgg	
	GncR29	73	82	aagga AGGATGTGGA tctcc	
	GncR31	6	15	ggcgc AGACAACAGC aagag	
	GncR32	27	36	gcaaa AGCCAGAAGC ccggt	
	GncR33	67	76	tggcg AGGATGAGGA tggga	
	GncR35	27	36	tctca AGGAAGGGGC ccctc	
	GncR36	13	22	acgtc AGGAAGGAGC ctaga	

3.2.4 Polymerase III transcription factors of *Giardia*

The Pol III transcription system of *Giardia* is not well understood. To date TBP and Brf are the only Pol III basal transcription factors that have been identified in *Giardia* through sequence homology search (Best et al. 2004). The results from the analysis of upstream and internal sequences of *Giardia* ncRNAs suggest the possibility of more Pol III transcription factors which are involved in recognition of potential upstream and internal sequence elements.

There is no published evidence for any *Giardia* homologues of TFIIA or TFIIC (Best et al. 2004), and due to lacking the 5S rRNA, it is likely that TFIIA is not present in *Giardia*. However, the potential internal promoter elements (Box-A and Box-B) of tRNAs suggest the presence of TFIIC-like transcription factors. In model organisms such as human and yeast, TFIIC binds DNA, and more importantly recruits TFIIB through its interactions with Brf (Chaussivert et al. 1995). In *Saccharomyces cerevisiae*, the largest subunit of TFIIC complex scTfc4 is the key component recruiting TFIIB (Rameau et al. 1994; Chaussivert et al. 1995). scTfc4 has eleven tetratricopeptide repeat (TPR)* motifs, which mediate protein-protein interactions (Marck et al. 1993). The human homologue of scTfc4 hTFIIC-102 also has eleven TPR domains in its primary sequence, and binds to hBrf through the TPR domains, and both human and yeast proteins contain a helix-loop-helix domain cooperating DNA binding (Geiduschek and Kassavetis 2001). Searching the sequences of predicted ORFs of *Giardia* (downloaded from <http://gmod.mbl.edu>) using Hidden-Markov-Model-based software (HMMer-2.3.2) has revealed a number of TPR-rich proteins. Searching for potential basic helix-loop-helix (HLH) DNA-binding domain in these proteins did not find strong hits ($E < 0.001$), however, several proteins aligned weakly (indicated by E values) to the HMMer alignment generated from 175 seed sequences (Pfam 21.0: HLH) As seen from the results shown in Table-4, it is unlikely that *Giardia* has a distinct homologue to scTfc4 and hTFIIC-102, although a number of TPR-containing proteins (Orf-27310, Orf-16287, Orf-15549, Orf 16226) can be potential

* tetratricopeptide repeat (TPR): The TPR motif consists of 3 to 16 tandem repeats of 34 amino acids residues, and mediates protein-protein interactions and the assembly of multiprotein complexes (InterPro entry: IPR001440).

candidates of TFIIC components. Given the distant relation of *Giardia* with the model organisms, it is perhaps more likely that *Giardia*'s protein components of the TFIIC complex are highly diverged, and therefore hard to identify through current bioinformatic approaches.

Table-4: HMM search for potential TFIIC protein components in *Giardia*

ORF No.	Annotation from <i>Giardia</i> genome	No. of TPR motif	HMMer HLH	E value
16934	Tetratricopeptide repeat family protein	15	X	
27310	similar to transformation-sensitive protein homolog [<i>Acanthamoeba castellanii</i>]	9		0.42
12081	UDP-N-acetylglucosamine--peptide N-acetylglucosaminyltransferase 110 kDa subunit	8	X	
16660	similar to Tg737 protein, isoform 1 [<i>Homo sapiens</i>]	9	X	
2198	serine/threonine protein phosphatase	3	X	
15148	DJC7_HUMANDnaJ homolog subfamily C member 7 (Tetratricopeptide repeat protein 2)	4	X	
21498	TPR repeat	3	X	
21971	similar to ENSANGP00000002840 [<i>Apis mellifera</i>]	4	X	
10529	similar to unnamed protein product [<i>Tetraodon nigroviridis</i>]	6	X	
7287	similar to LOC394994 protein [<i>Xenopus tropicalis</i>]	3	X	
16287	similar to outer arm dynein binding protein [<i>Anthocidaris crassispina</i>]	5		0.22
87202	similar to RIKEN cDNA 4930506L13 [<i>Rattus norvegicus</i>]	4	X	
8508	similar to TTC8_HUMAN Tetratricopeptide repeat protein 8	4	X	
11177	similar to kinesin light chain [<i>Methanosarcina acetivorans</i> C2A]	4	X	
15549	hypothetical protein	4		0.29
7639	TPR domain protein	3	X	
5949	putative tetratricopeptide repeat protein	2	X	
28657	TPR repeat	3	X	
113023	Hypothetical Protein	5	X	
9594	similar to Suppression of tumorigenicity 13 [<i>Gallus gallus</i>]	2	X	
16375	hypothetical protein	4	X	
17624	similar to ENSANGP00000027263 [<i>Anopheles gambiae</i> str. PEST]	3	X	
16226	Hypothetical Protein	3		0.3

(Potential absence of HLH domain is determined by an E-value less than 0.5, indicated by "X")

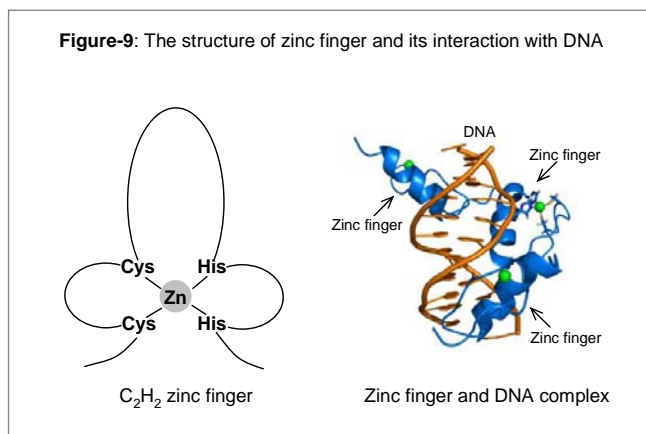
In addition to the eukaryotic general Pol III transcription factors TFIIA, TFIIC and TFIIB, the special promoter structure of Class-3 genes with their

upstream promoter elements (PSE and DSE) require additional transcription factors. The basal transcription factor SNAPc (snRNA activating protein complex) binds specifically to PSE (Murphy et al. 1992) of snRNA gene promoters. In Pol III snRNA promoters, PSE functions in concert with the TATA-box to direct a low level of transcription, thus SNAPc together with TBP nucleates an assembly of an Pol III initiation complex (Henry et al. 1998). SNAPc constitutes a direct target for the transcription factor Oct-1 (containing octamer-binding domain), which is a POU-domain* protein (Herr et al. 1988), binds cooperatively with SNAPc to the DNA through interaction with DSE . In addition, another transcription factor Staf, originally identified in *Xenopus* and a later identified highly conserved human homologue, ZNF143/SBF are seven zinc-finger (C₂H₂ – two cysteines and two histidines) proteins, which bind to varied upstream sequences of U6 snRNA and some tRNAs (Schaub et al. 1997; Myslinski et al. 1998; Schaub et al. 1999).

POU proteins are eukaryotic transcription factors containing a bipartite DNA-binding domain known as the POU domain (Sturm and Herr 1988). POU-domain proteins have been identified in animals, but not yet in plants and fungi (Herr et al. 1988; Petryniak et al. 1990; Verrijzer and Van der Vliet 1993). The POU domain is composed of two subunits, a POU-specific N-terminal subunit and the C-terminal homeobox subunit, separated by a non-conserved region of 15-55 aa (Verrijzer and Van der Vliet 1993). Both subdomains contain the “helix-turn-helix” structural motif, and are required for high-affinity DNA binding and protein-protein interaction (Klemm et al. 1994). Zinc-finger proteins are a major type of DNA-binding proteins, first identified in *Xenopus* transcription factor TFIIIA (Miller et al. 1985). A C₂H₂ zinc-finger domain contains 2 conserved Cys and 2 conserved His residues, and the 12 residues separating the second Cys and the first His are mainly polar and basic, implicating this region in nucleic acid binding (Rosenfeld and Margalit 1993). The zinc-finger motif is a small self-folding domain in which Zn is a crucial component of its tertiary structure; the zinc-finger motif interacts with DNA in the major groove, and the Zn binds to the conserved Cys and His residues

* POU domain: a bipartite DNA-binding domain (InterPro entry: IPR013847)

(Berg 1988; Lu et al. 2003; Simpson et al. 2003). The structure of zinc finger and its interaction with DNA are shown in Figure-9. It has been suggested that zinc-fingers may represent the original nucleic acid binding domain as they have the ability to bind to both RNA and DNA (Rosenfeld and Margalit 1993; Lu et al. 2003).



Since *Giardia* is highly reduced and diverged from vertebrates and other general model organisms, lower similarity between *Giardia*'s proteins and those of other organism is expected. Searching for

the POU domain using Pfam POU-domain alignment in *Giardia*'s ORFs did not identify significant hits, but by using Pfam zinc-finger-domain alignment I have found a number of C₂H₂-zinc-finger-domain containing protein sequences. The output of zinc-finger-domain search applied to *Giardia* contains two subunits of the RNA polymerase system, two splicing related factors and a putative reverse transcriptase. The rest of the output mainly contains uncharacterized zinc-finger-domain proteins, among which there may be potential transcription factors (as shown in Table-5).

Table-5: Putative C₂H₂ Zinc finger-containing proteins from *Giardia*:

ORF number	Annotation	No. of putative zinc finger domains	E value
17003	Zinc finger domain	5	3.4e-19
13007	Zinc finger domain	4	1.1e-8
14069	Hypothetical protein	2	5.4e-5
8920	Zinc finger domain	1	0.0015
4343	Zinc finger domain	3	0.012
27035	Hypothetical protein	1	0.039
19815	Hypothetical protein	1	0.061
8405	Hypothetical protein	3	0.063
3763	Protein 21.4	1	0.13
5822	Hypothetical protein	1	0.22
16877	Zinc finger domain	1	0.3
14119	Zinc finger domain	1	0.93

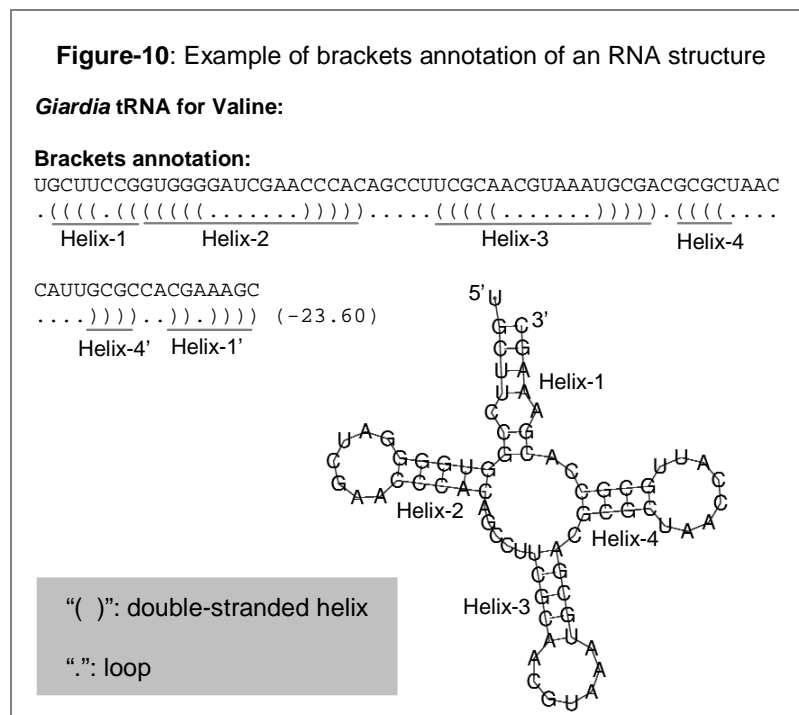
3.2.5 Structural analysis of uncharacterized novel ncRNA candidates

Characterization of ncRNAs based on sequence similarity often encounters major difficulties, as functions of most ncRNAs are determined primarily by their structures. Like proteins, ncRNAs with similar functions may not share extensive sequence similarities; however they can generally fold into similar structures. Therefore, it is very important to study the folding of ncRNAs in order to determine their functions and classify different types of ncRNAs. Structural prediction is a major method of characterising newly found ncRNAs. Until recently a number of reliable computational methods (Lowe and Eddy 1997; Lowe and Eddy 1999; Eddy 2002; Klein and Eddy 2003) for identification of ncRNAs are still based on sequence motif similarity search, which does not work on previously unknown types of ncRNAs.

A number of computational methods have been developed to fold a single RNA sequence (Hofacker 2003; Mathews et al. 2004); however, computationally predicted structures are often different from the true structures *in vivo*, because the folding of RNAs in the cell is usually associated with protein-cofactor binding and different metal ion associations. These variable conditions are hard to simulate. The structures of ncRNAs can more reliably be determined from a set of phylogenetically or functionally related ncRNA sequences through structural alignment. The evolution of structured RNAs has a unique property that substitutions of distant bases are correlated to retain the conserved stem structures. The Sankoff algorithm (Sankoff 1985) is a multiple alignment algorithm that includes the effect of base-pair correlation. Practical variations of the Sankoff algorithm have been intensively investigated in recent years and resulted in a number of algorithms. The first group of algorithms score the structures using the free energy parameters, which give accurate structural predictions but with very high computational cost. These include programmes such as Dynalign (Mathews and Turner 2002) and Foldalign (Havgaard et al. 2005; Kiryu et al. 2007; Torarinsson et al. 2007). The second group of algorithms use a probabilistic model called the pair stochastic context-free grammar (PSCFG), which has the advantage of relatively low computational cost; however the accuracies of structural predictions are only

moderate. This group includes programmes such as Consan (Dowell and Eddy 2006) and Stemloc (Holmes 2005).

As a first test, three ncRNA candidates with high GC-content (above 65% as indicated in Table-1) are aligned using the newly improved multiple structural alignment programme FoldalignM (Torarinsson et al. 2007) and a recently developed multiple alignment programme Murlet (Kiryu et al. 2007). The structures here are all shown in brackets annotation, which is a format widely used in computing RNA structures. Figure-10 shows an example of tRNA-folding using brackets annotation.



Results (Figure-11) show considerable degree of structural conservation among the three sequences tested. The individual structures of all uncharacterized novel ncRNA candidates in this study are shown in Appendix-2 with annotation in both brackets and graphical annotations.

Figure-11: Comparison of two recent multiple alignment programmes using *Giardia* ncRNA candidates

a) Test alignment of three ncRNA candidates using FoldalignM

```
GncR17 GAGGTAATAGACCAGGCTGCCAGCCCGGCGAAGGTCTGCAAGTGTGACGGAGACAATGGCTACACGCTCCAGG
GncR17 .....(((.(((.....)).(((...)).).....))))).(((.....
GncR3  CTTCAA-CTCAGCCGGACAG-----CCGGAGGCCG-G-AGACGGAGCACGGTCAGCGGGCGGGTGCAGT
GncR3  .....-(((.(((.....-----(((.....))-).(((.....)).).....))))).(((.....
GncR38 -----CCCACCGCGTTCAGTGCTGGCCAGGGGC-AAGGAGGCCTGC-TCTCCCTGG-CCTCTGCGGAAA
GncR38 -----(.....((.....)).(((.....))-...(((.....-.))).(((.....))....
```

```
GncR17 GCGACGCGTGCACCAAGGCGGCTCTGACAACGCGTGCCAGACCTGGGAACCGCGGGTGTGCCAC
GncR17 ..).)))).).....).....(((.....((.....)).).....)).).....
GncR3  GCCAGCCCC-----AG--CCGCAGAGCG-G-C--T-----TCCTTA
GncR3  ..).)))).)-----((--(((.....))-)--)-----).....
GncR38 CGG-----G--CAGCTG--CGTGATCCAAGTGC--AGCC----ACCAC
GncR38 ))-----)---(((.....)).)---)-----).....
```

b) Test alignment of three ncRNA candidates using Murlet

```
GncR3      CUU...CAACUCAGCCGGACAGC.....CGGAGGC.....C.GGAGACG...GA
GncR17     GAGGUAAUAGACCAGGCGCCAGCC.CGG.CGAAGGUCUGCAAGUGUGACGGAGACAAUGGCU
GncR38     CC.....C.ACCGGCGUUCAGUGCGCCAGGGGC.....AAGGAGGC.....CU
#=GC SS_cons .....<<<<<.....<.....>.....
```

```
GncR3      GCACGGUC.AGGCGGGCGGGUGCA.....GUGCC.....AG.CCC.CAGCCGCAGAG.
GncR17     ACACGCUCCAGGGCGACGC.GUGCACCAAGGCGGCUCCUGACAACGCGUGCCAGACCCUGGGA
GncR38     GCUCUCCC.UGGCCUCUGCGG.AAAC.GGGCAGC.....UG..CGU..GAUCCACUGA..
#=GC SS_cons .....<<<<<.....>>>>>.....>>>>>.....
```

```
GncR3      .CGGC.....UUCUUUA
GncR17     ACCGCCGGGUGUGCCAC
GncR38     .CAGC.....CA.CCAC
#=GC SS_cons .....
```

The test results show that using the same dataset, FoldalignM and Murlet give very different output. FoldalignM performs pair-wise alignment of all the input sequences and combines the results to give the final output, and it gives more detailed structural prediction of each sequence. On the other hand, Murlet does progressive alignment and gives the highest consensus structure of all the input sequences. However, running FoldalignM requires extremely high amount of computer memory. In order to run FoldalignM for more than three sequences, all-against-all pair-wise alignments were done using Foldalign 2.0.3.

First, all candidates were aligned globally pair-wise using Foldalign 2.0.3. Global alignments show that most of the pair-wise alignments have overall identities between 30% and 40%, with the highest of 51% and lowest of 6%. Through careful analysis, it has been noticed that all pairs of sequences except GncR33 and GncR39 are interlinked by similarities above 40%. However, further analysis showed that some ncRNA candidates can be grouped together so that within each group, three or more sequences share overall similarities above 40%. Multiple structural alignments were performed for these groups of sequences and results are shown in (Appendix-2).

In addition to the potential groups of ncRNAs obtained solely by structural alignment, it was expected that the ncRNA candidates that share similar sequence motifs would fold in to similar structures. To test this, sequences that share potential sequence motifs (Table-3a) were aligned by FoldalignM. Results are shown in Figure-12. The overall structural identities are higher than average identities observed for all-against-all pair-wise alignments of ncRNA candidates. These results suggest that sequence and structural similarities can be linked in classification of new ncRNAs, and the ncRNAs sharing similar sequence motifs and structures may be considered as one type.

Figure-12: Multiple structural alignments of ncRNA candidates that share potential sequence motifs

a) Group1 based on motif:

```
GncR2  TCCCTGGGCGTCGGGCAGAAAGTGCCGGTCTCTGGATTCCGGGGAGTGTCTGGTGCCGATCGGACACTCCCTAGC
GncR2  ..((.....))..((.....((.....)))..)...(((((.....))))).).....
GncR33  T-GTAGGTCT-AAC--AT-----GCTTGCCACGGC-GT--CCCC-GGACATGGCACCGTCTATGTCTGCTTGTT
GncR33  .-((.....-))--((-----((.....))-)--...-(((.....))))).).....
GncR38  CCCAC-----CG-----GCGTTCAGTGC-TGGCCAG--GGGCAAGGAG--GCCT-GCTC-----
GncR38  .....-----((-----((.....))-)...--(((.....))-))-----
GncR9   TGGACG--ATGAACT-GGAGATGCTG-GACACGGCTT--TG-CTCTC-CCACCGGA-GCA--CATATGCTGCAGG
GncR9  .....--((.....-(((.....-(((.....--((.....-((.....-((.....))))))..((

GncR2  CGCCCACTGACAGTTATGGTTGACAGGACAAGCTTAGCGA-GTCCGAACTCGACAGGATACTCTACAGCGTTCC
GncR2  (((.....((.....((.....))))).).....(((((.....-((.....))))).).....)).....).....).....
GncR33  GGCGA--GGATGAGGATGGGAACACC-TGAGCTTGGGGCTGTTAGTACGCCCTCAAGAGCCGTCAGCCCTCCT
GncR33  (((.....-((.....((.....)))))-)...(((((.....-((.....))))).).....)).....).....).....
GncR38  ----TCCCTGGCCTCTGCGGAAACGGG--CAGCT-GC-G-TGAT-CCACTGAC-AGC-----CA-----CCAC
GncR38  ----...((.....((.....))))--...(((((.....-((.....))))).).....-))-----)-----
GncR9   A-TGACCG--GCGCCTG-TCTC--CCACCAGTG-C-CA-GCTAA-ACTGCAG-C-----CACATT
GncR9  .-((.....-)))))..-)))))--.....)---((.....-))..-))-----)-----
```


b) Group2 based on motif:

GncR1 TTCGGGATCAGTTTTGGAGTTAATACCACCAAACCCCTGTGCGTACATGTCGCCCCCTAACCTTCTGATGCGGATACCTTG
 GncR1 ...(((.....(.....((.....)).....((.....)).....)).....(((.....
 GncR14 ---GCT-----CCTAGAGGAA-----GAG-----G-CAGGC--ATG---C-----AGGATATTTT---G
 GncR14 ---(((-----((.....))-----))-----((.....))-----))-----((.....
 GncR17 GA--GG---TAATAGACCAGGCTGCCAGCCCGCGAAGGTCTGCAAGTGTGACGGAGACAATGGCTACACGCTCCAG--G
 GncR17 .--((-----((.....((.....((.....)).....)).....)).....)).....(((.....
 GncR19 ----CA---GAAGAATGCCA-GCAAGTCATGC---AATG-CCTG-T-----GG-ATCC-GTCCCT---C
 GncR19 ----((-----((.....((.....((.....)).....))-----))-----))-----((.....
 GncR32 ACACAAA-----AG-GTG---AGC-----GCGTAAGCAAAA
 GncR32-----((-----((.....
 GncR35 AC-GGGAATAAC--G--CC--CACAGGATCTCA--AGG---AAGG-----GG-CCCCTCAAC---T
 GncR35 ..-(((.....((-----((.....((.....))-----))-----))-----))-----
 GncR39 TGTGCCAC-----TGT-G-GCTTC---GAGCTCTAT-AATGCGCGACT-TAAGAACC--TCTGCTCT--A
 GncR39-----((-----((.....))-----))-----((.....((.....
 GncR1 CCGCAGGGCCGTTAAGCGAGGCTTGGCCCGTGCACGATGAGGCTCCCTGCGGGGAAGCCCTGCGGCGCGTCTTAAGGAGGC
 GncR1))))....((.....))))).....(((.....))))).....
 GncR14 GATGGACAG--CCCTCAT-AAGGGCAGC-----AGAGAGTGGCCACGC--AGGCTGC-----AC
 GncR14))))....((.....))))).....(((.....)))))-----
 GncR17 GCG--ACGCG--TGACACAA-GGCGGCTCCTGACAACCGTGCAGACCCTGGGAACC-GCC-GGG-----TGTGCCAC
 GncR17))--...((.....))))).....(((.....)).....(((.....))-----)).....
 GncR19 GAC-----CTTCTCCTGA-CAGACATGTGTCTTT--TGG-CAT-----GCA-----GCCCTG-C
 GncR19))-----((.....)).....(((.....))-----))-----))-----))-----
 GncR32 GC-C-AGAAGCC-CGTTGC-AGCG-CTTGC--TT-----CGCAGCTCTA---CGGGC-G-----C
 GncR32))-)...((.....))-----))-----))-----((.....))-----))-----
 GncR35 TGA---A---GCTCTGATCGGGT-CCC-----AAGCACAAGTAAAT--AATT-G-----CC
 GncR35))---...((.....))-----))-----((.....))-----))-----
 GncR39 CAGA---CT---TACTTCAAGTAAAGATGTCGC-----AGT-TAGTGCCT-CCTCAAC-----ACAGCTTTC
 GncR39))))---((.....))))).....))-----((.....))-----))-----
 GncR1 CCGCAGGGCCGTTAAGCGAGGCTTGGCCCGTGCACGATGAGGCTCCCTGCGGGGAAGCCCTGCGGCGCGTCTTAAGGAGGC
 GncR1))))....((.....))))).....(((.....))))).....
 GncR14 GATGGACAG--CCCTCAT-AAGGGCAGC-----AGAGAGTGGCCACGC--AGGCTGC-----AC
 GncR14))))....((.....))))).....(((.....)))))-----
 GncR17 GCG--ACGCG--TGACACAA-GGCGGCTCCTGACAACCGTGCAGACCCTGGGAACC-GCC-GGG-----TGTGCCAC
 GncR17))--...((.....))))).....(((.....)).....(((.....))-----)).....
 GncR19 GAC-----CTTCTCCTGA-CAGACATGTGTCTTT--TGG-CAT-----GCA-----GCCCTG-C
 GncR19))-----((.....)).....(((.....))-----))-----))-----))-----
 GncR32 GC-C-AGAAGCC-CGTTGC-AGCG-CTTGC--TT-----CGCAGCTCTA---CGGGC-G-----C
 GncR32))-)...((.....))-----))-----))-----((.....))-----))-----
 GncR35 TGA---A---GCTCTGATCGGGT-CCC-----AAGCACAAGTAAAT--AATT-G-----CC
 GncR35))---...((.....))-----))-----((.....))-----))-----
 GncR39 CAGA---CT---TACTTCAAGTAAAGATGTCGC-----AGT-TAGTGCCT-CCTCAAC-----ACAGCTTTC
 GncR39))))---((.....))))).....))-----((.....))-----))-----

c) Group3 based on motif:

GncR17 GAGGTAATAGACCAGGCTGCCAGCCCGG--CGAAGGTCTGCAAGTGTGACGGAGACA-ATGGCT-ACACGCTCCAGGGCGA
 GncR17-((.....((.....))-----))-----))-----))-----
 GncR18 TAC-----CA-----CTCT--GACCGTG--AGG-C-GCATGCCTA-GGGCA-
 GncR18 ..-.....((.....))-----((.....))-----))-----))-----
 GncR2 TCCCTGGGCGTGGGCGAGAAAGTCCCGTCTCTGATTCGGGGAGTGTCTGGTGCCGATCGGACACTCCCTAGCCGCCA
 GncR2((.....((.....)).....)).....(((.....((.....)).....)).....
 GncR30 CAC-----GA-GG-----AAACGAGTG--TTTCGCCGGGCATAACTGGGCA
 GncR30 ..-.....((.....))-----((.....))-----))-----))-----
 GncR17 CGCGTG--CACCAAGCGGCTCCTGACAACGC--GTGCCAGACC--CTGGGAACCGCCGGGTGTGCCAC
 GncR17))-----))-----))-----((.....))-----))-----
 GncR18 ---TG-----GAG--AAGAGCAG--ACTT---G-----AG
 GncR18 ---))-----))-----))-----((.....))-----))-----
 GncR2 CACTGACAGTTATGGTTGACAGACAAGCTTAGCGAGTCCGAACCTCGACAGGGATACTCTACAGCGTTCC
 GncR2)).....)).....)).....)).....)).....)).....)).....)).....)).....
 GncR30 TGCATTTTCC-TTGCC---CA----GTCT--GCCT--CCATACT--AAT-TTC---TC----CTA
 GncR30))-----))-----))-----))-----))-----))-----))-----
 GncR17 GAGGTAATAGACCAGGCTGCCAGCCCGG--CGAAGGTCTGCAAGTGTGACGGAGACA-ATGGCT-ACACGCTCCAGGGCGA
 GncR17-((.....((.....))-----))-----))-----))-----
 GncR18 TAC-----CA-----CTCT--GACCGTG--AGG-C-GCATGCCTA-GGGCA-
 GncR18 ..-.....((.....))-----((.....))-----))-----))-----
 GncR2 TCCCTGGGCGTGGGCGAGAAAGTCCCGTCTCTGATTCGGGGAGTGTCTGGTGCCGATCGGACACTCCCTAGCCGCCA
 GncR2((.....((.....)).....)).....(((.....((.....)).....)).....
 GncR30 CAC-----GA-GG-----AAACGAGTG--TTTCGCCGGGCATAACTGGGCA
 GncR30 ..-.....((.....))-----((.....))-----))-----))-----
 GncR17 CGCGTG--CACCAAGCGGCTCCTGACAACGC--GTGCCAGACC--CTGGGAACCGCCGGGTGTGCCAC
 GncR17))-----))-----))-----((.....))-----))-----
 GncR18 ---TG-----GAG--AAGAGCAG--ACTT---G-----AG
 GncR18 ---))-----))-----))-----((.....))-----))-----
 GncR2 CACTGACAGTTATGGTTGACAGACAAGCTTAGCGAGTCCGAACCTCGACAGGGATACTCTACAGCGTTCC
 GncR2)).....)).....)).....)).....)).....)).....)).....)).....)).....
 GncR30 TGCATTTTCC-TTGCC---CA----GTCT--GCCT--CCATACT--AAT-TTC---TC----CTA
 GncR30))-----))-----))-----))-----))-----))-----))-----

3.3 Conclusion

Various studies have identified ncRNAs in *Giardia*, including 3 rRNAs (Edlind and Chakraborty 1987), 44 tRNAs (McArthur et al. 2000), 28 snoRNAs (Yang et al. 2005; Luo et al. 2006; Chen et al. 2007), RNase P (Marquez et al. 2005), U5 snRNA (Collins et al. 2003), and 33 novel ncRNAs of unknown types (this study). Compared with other eukaryotes, the number of known *Giardia* ncRNAs is yet small, and searching for snRNAs from *Giardia* has encountered obstacles. At present constructing cDNA libraries is one of the most efficient ways to uncover previously unknown ncRNAs on a large scale, and this method has been applied to many organisms including bacteria and eukaryotes. 38 ncRNAs have been identified in our *Giardia* ncRNA library, including four new snoRNAs. However, the long phylogenetic distance between *Giardia* and other eukaryotes leads to difficulties in characterising the novel ncRNA candidates. Structural analysis provides useful information for classifying unknown ncRNAs that are potentially of the same type.

Gene transcription and regulation in *Giardia* is not well understood. The upstream elements of the Pol II transcription system have roughly conserved sequences. However the Pol III transcription system appears to have more flexible upstream sequence elements. By analyzing the upstream sequences of various ncRNAs from *Giardia*, several potential Pol III upstream and internal elements have been observed especially for tRNAs. Results from the present analysis suggest the possibility that many ncRNAs in *Giardia* may be transcribed by RNA Pol III. However it does not exclude the possibility that a number of the ncRNAs are co-transcribed with adjacent protein-coding genes.

In conclusion, our size-fractionated cDNA library of *Giardia* provided an overview of the various types of ncRNAs within this organism although most of the uncovered ncRNAs are not yet functionally characterized. The presence of many unknown ncRNAs suggests that there may be unusual RNA involved mechanisms in *Giardia*. Comparing the known *Giardia* ncRNAs with those of the higher eukaryotes has shown that the central RNA-processing pathway, which involves the well-studied snoRNAs, snRNAs, RNase P and microRNAs,

has evolved during early eukaryotic evolution. Although some key ncRNAs (such as the spliceosomal snRNAs and microRNAs) have not been thoroughly studied in *Giardia*, they are highly likely to be present, because the protein Dicer is present (Macrae et al. 2006). Comparing the ncRNAs from *Giardia* with those from other deep-branching eukaryotes will help understand the overall RNA processing during early eukaryotic evolution. However, information on ncRNAs in these organisms is still even more limited.

Studies of ncRNAs from deep-branching eukaryotes provide insights into the evolution of ncRNAs as important components of the cellular machinery. It is likely that the types of ncRNAs in deep-branching organisms differ from those in higher eukaryotes, although the divergence of eukaryotes brings major difficulties for characterization of novel ncRNAs identified from deep-branching eukaryotes. More studies on different deep-branching eukaryotes (e.g. other protists) will help understanding the conservation and changes of ncRNAs during eukaryotic evolution. At this stage, the analysis of novel ncRNAs from *Giardia* has shown that the major types of ncRNAs present through out eukaryotic species, and the presence of *Giardia*-specific ncRNAs is highly likely.

Chapter-Four: Studies of the major spliceosomal snRNAs in *Giardia*

Abstract

Pre-mRNA splicing is one of the most important RNA-processing mechanisms in eukaryotes. Splicing is mostly catalysed by a macromolecular complex – the major spliceosome which consists of five uridine-rich small nuclear RNAs (U-snRNAs) and over 200 proteins. Three major spliceosomal introns have been found experimentally in *Giardia*. One *Giardia* U-snRNA (U5) and a number of spliceosomal proteins have also been identified. However the other U-snRNAs of *Giardia* have not been found previously due to expected low sequence similarity between the *Giardia* ncRNAs and those of other eukaryotes. This chapter describes my studies on searching for the other four spliceosomal U-snRNAs in *Giardia* plus the analysis of the *Giardia* homologue of a prominent spliceosomal protein Prp8 protein. Using two computational methods, candidates for *Giardia* U1, U2, U4 and U6 snRNAs were identified. Expression of these candidates was confirmed by RT-PCR. Secondary structural modelling of these *Giardia* U-snRNA candidates revealed typical features of eukaryotic U-snRNAs. In addition to the identification of *Giardia* U-snRNA candidates, one central protein component of the spliceosome Prp8 protein was analysed. Computational analysis revealed putative functional domains within the *Giardia* Prp8 protein, and a small scale biochemical study was done to test potential RNA-binding properties of the putative RNA-recognition domain within the *Giardia* Prp8 protein. In all this chapter shows that it has been successful to combine different computational and experimental methods to identify expected ncRNAs in a highly divergent protist genome. Although the experimental studies on the *Giardia* spliceosomal proteins are still at a primary stage, the results obtained in this study provide useful information for future research on spliceosomes and splicing mechanisms in deep-branching eukaryotes.

4.1 Introduction – Does *Giardia* have a functional spliceosome?

The spliceosome is one of the most important RNA processing units in eukaryotes. The presence of spliceosomal introns in deep-branching eukaryotes (Nixon et al. 2002; Russell et al. 2005; Vanacova et al. 2005; Slamovits and Keeling 2006) suggests that the splicing mechanism is likely to have evolved very early during eukaryotic evolution (Collins and Penny 2005), despite the

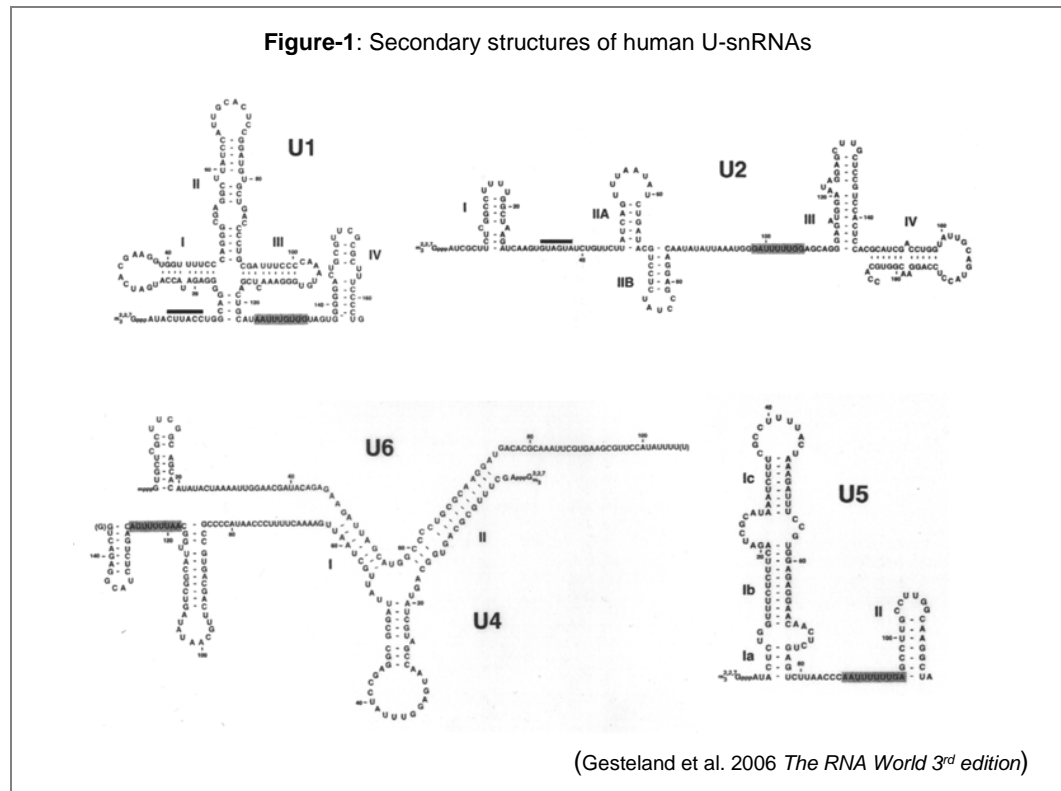
small number of introns found in some deep-branched eukaryotic species such as *Trichomonas vaginalis* (Vanacova et al. 2005) and *Giardia lamblia* (Nixon et al. 2002; Russell et al. 2005). To date only three spliceosomal introns have been experimentally confirmed in *Giardia*. The first one is a short (35nt) non-canonical intron (5'-CT – AG-3') located within the mitochondrial [2Fe-2S] ferredoxin protein (Nixon et al. 2002), the second one is a 109nt canonical intron (5'-GT – AG-3') found in the ribosomal protein Rp17a (Russell et al. 2005) and the third one is a 220nt canonical intron found in an unassigned ORF (Russell et al. 2005).

Genomic surveys (Nixon et al. 2002; Collins and Penny 2005) have revealed a number of spliceosomal proteins from the *Giardia* genome. These include homologues of Prp8, Prp11, Prp28 and Prp31; a number of DExH-box RNA-helicases which have homologues in bacteria but which also have important roles in eukaryotic intron splicing; 11 archaeal-like Sm and Lsm core peptides which coat the spliceosomal snRNAs; and a number of U-snRNA-specific peptides. It is therefore very likely that *Giardia* has a functional spliceosome, although there have been no extensive biochemical studies.

The aim of this part of the study is to look for more evidence which may support the hypothesis that *Giardia* has a functional spliceosome by looking for candidates of the U-snRNAs and studying a major protein component: the *Giardia* homologue of splicing-related protein Prp8.

In humans, the major spliceosome is composed of over 200 proteins and five uridine-rich small nuclear RNAs (U1, U2, U4, U5 and U6) that form dynamic protein-RNA and RNA-RNA interactions (Nilsen 2003). The detailed mechanism of splicing is described in the introductory Chapter-1 (page 36). Like other ribozymes, the RNA components of the spliceosome are the major catalysts of splicing. It has been shown that human protein-free spliceosomes are capable of catalysing reactions that resemble both the first (Valadkhan et al. 2007) and second (Valadkhan 2005) steps of trans-esterification reactions during splicing. The U-snRNAs are found across the eukaryotic kingdom and

have the characteristic Sm-protein binding site, which is a conserved 8-10nt uridine-rich sequence flanked by two stem-loops. The structures of these snRNAs are also highly conserved. Figure-1 shows the secondary structures of human U-snRNAs. To date many studies have shown that the U-snRNAs from a wide range of organisms share the same stem-loop folds (Vankan et al. 1988; Brown and Waugh 1989; Hofmann et al. 1992; Miranda et al. 1996; Valadkhan 2005; Hinas et al. 2006; Ambrosio et al. 2007).



The stem-loops within these snRNAs are important for interactions with snRNA-specific proteins. Each of the five snRNAs has a number of specific interacting proteins ranging from 4 in human to 10 in yeast (Jurica and Moore 2003). However in deep-branching eukaryotes, the protein components are usually reduced. Bioinformatic studies have shown that *Giardia* is likely to have most of the more conserved snRNA associated major spliceosomal proteins although the less conserved ones may be lost (Collins and Penny 2005). The predicted presence of many spliceosomal proteins suggests that *Giardia* is highly likely to possess a functional spliceosome.

The *Giardia* U5-snRNA has been found by computational analysis (Collins et al. 2003), and it folds into a conserved U4 U5 secondary structure although the

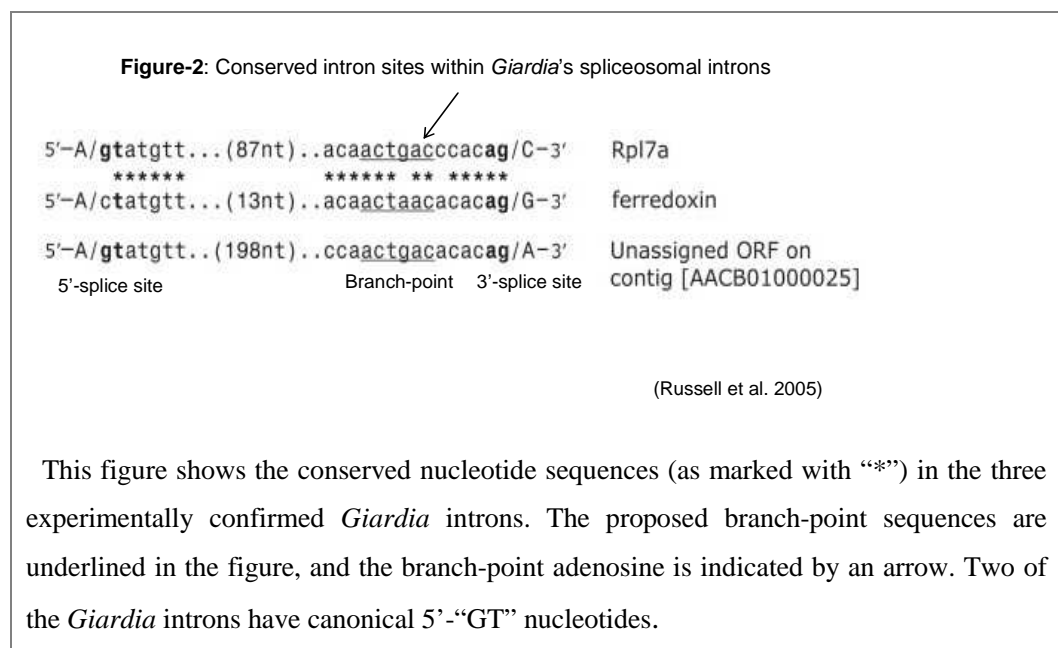
primary sequence itself does not show homology with U5-snRNAs from other species. The U5-snRNP is required for both steps of splicing (Dix et al. 1998) and is the only snRNP found in all three types of splicing: major-, minor- and trans-splicing. The U5snRNP-specific proteins Prp8 and Brr2 are also found in other deep-branching eukaryotes including *Trypanosoma brucei* (Lucke et al. 1997) and *Trichomonas vaginalis* (Fast and Doolittle 1999). The Prp8 protein, a unique and highly conserved protein which has no obvious homology to other proteins, has a central role within the spliceosome and makes extensive protein-protein interactions throughout the various stages of pre-mRNA splicing (Grainger and Beggs 2005). Therefore, given the presence of U5 and Prp8, it is highly likely that *Giardia* has a functional spliceosome. The presence of U5 snRNA and all the protein components from *Giardia* suggest the high possibility that *Giardia* possesses other U-snRNAs too. The aim here is to test these predictions.

4.2 Searching for U-snRNAs in *Giardia*

4.2.1 Prediction of *Giardia* U1-snRNAs candidate

Searching for U-snRNA candidates from *Giardia* based on primary sequence similarity (Blast and profile HMM) failed as expected, due to the observed low sequence homology between *Giardia* and other eukaryotes. However, the generally conserved structures of the U-snRNAs may allow a more advanced computational search for new U-snRNA candidates from the fully sequenced *Giardia* genome (McArthur et al. 2000; Morrison et al. 2007). Due the reduced nature of the *Giardia* genome (Edlind and Chakraborty 1987; Adam 2001; Vanacova et al. 2003; Best et al. 2004; Morrison et al. 2007), it is not unlikely that some of the ncRNAs from *Giardia* also have been reduced in size and structure. For example, it has been shown that the U1 snRNA from *Trypanosoma brucei* (see Figure-4) is unusually reduced that it only contains one stem-loop structure in contrast to the usual four stem-loops seen in other organisms (Palfi et al. 2005).

Besides structural information, certain sequence motifs of the U-snRNAs can also aid computational searches. It is known that U1-snRNA and U2-snRNA have direct interactions with introns through complementary nucleotide sequences; U1 binds to the 5'-intron splice site and U2 binds loosely at the branch site (Das et al. 2000). The three spliceosomal introns in *Giardia* (Nixon et al. 2002; Russell et al. 2005) share sequence similarities which indicate the presence of conserved 5'-, 3'- splice sites and the branch site as shown in Figure-2. Together with the conserved U-rich Sm-binding site, these sequence elements can be incorporated into a computational search for snRNAs from *Giardia*.



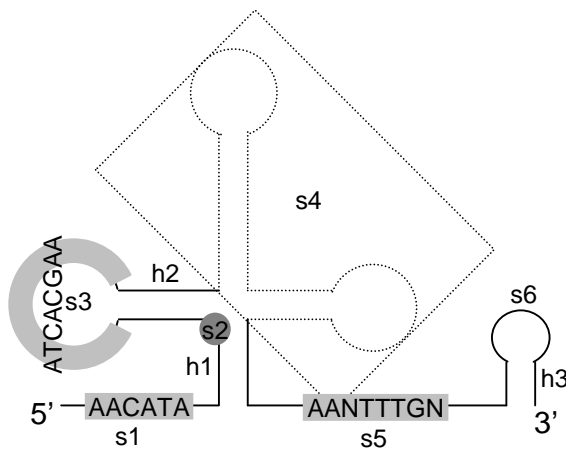
The computational prediction for U1-snRNA candidates was done using RNAbob* programme. This programme uses a descriptor file which specifies the structure and sequence motifs of the RNA to be searched, and looks for matching candidates from a sequence database. The descriptor file for U1-snRNA was constructed using the information available for *Giardia* (e.g. intron-binding sequence, expected Sm-binding site and predicted conserved loop sequence as detailed below). Since it was not known whether the U1-snRNA from *Giardia* was typical with conserved structure similar with human

* The source code of RNAbob was downloaded from:

<http://selab.wustl.edu/cgi-bin/selab.pl?mode=software#rnabob>.

U1 or reduced like U1 from *T. brucei* (Palfi et al. 2005), a relaxed model was set using the structural information from both the human and *T. brucei* U1-snRNA with human U1-snRNA as the upper limit of complexity and *T. brucei* U1-snRNA as the lower limit of complexity (Figure-4). The searching model was set so that the expected output would have the 5'-intron site recognition sequence "AACAU" which complements "UUGUAU" sequence at the 5' of intron. The Sm-binding sequence was set to "AANUUUGN" where N indicates an uncertain nucleotide. The stem-1 and stem-2 which were seen in both human and *T. brucei* are highly conserved at the loop sequence (Figure-4). Therefore this loop sequence (conserved as "AUCACGAA") is also incorporated into the search. Finally, a terminal stem which is present in both human and *T. brucei* was also used as searching criterion. All the "U"s are written as "T"s in the descriptor file for searching in a DNA genome. The descriptor file for U1 was written according to the proposed structure of the U1 candidate as shown in Figure-3. This proposed structure is deduced based on known U1-snRNA structures.

Figure-3: Proposed structure for writing the U1 descriptor file:



The content in the U-1 descriptor cell can be visualized in the drawing shown in this figure. "s" stands for strand and "h" stands for helix. The elements within the proposed U-1 structure are marked in order from the 5'-end to the 3'-end. The two stem-loops drawn as dotted lines are not compulsory in the proposed structure of *Giardia* U-1 candidate; therefore they are marked as a free-folding strand s4.

In the descriptor file below, lines started with “#” are notational. The “strands” and “helices” elements within the proposed structure are listed in order, and each of them is then specified. “N” represents an uncertain nucleotide which is definitely present and “*” represents an optional nucleotide. [] indicates the maximum number of nucleotides present. Since the presence of stem 3 and 4 (as marked on human U1-snRNA structure in Figure-4a) is uncertain, these two optional stems were replaced by a long strand s4. The numbers immediately following element (s1, h1 etc.) described indicate number of mismatches allowed. For example “0:0” shows that no mismatches are allowed in the helix h1.

```
# U1 snRNA descriptor (Giardia)
# 5' intron recognition site: 5' AACATA 3'
# Sm-binding site AANTTTGN
# Giardia's snRNAs may be reduced as seen with the U1 snRNA from T.
brucei.
# the search is done by restricting stem-1, stem-2 and stem-5 only, in
case
# stem-3 and stem-4 are missing.
# conserved loop-2 sequence seen in human and T. brucei: ATCACGAA

s1 h1 s2 h2 s3 h2' s4 h1' s5 h3 s6 h3'

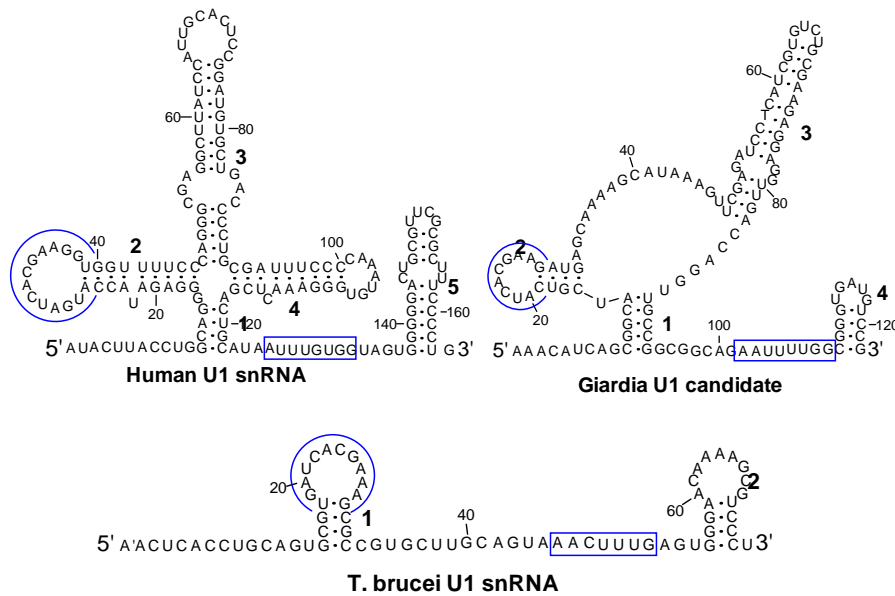
s1 1 NAACATANN
h1 0:0 NNNN:NNNN
s2 0 N
h2 1:0 *****NNNN:NNNN*****
s3 1 *ATCACGAA*
s4 0 NNNNNNNNNNNNNNNNNNNNN[50]
s5 0 NNNNN*****AANTTTGN*****
h3 2:2 *****NNNN:NNNN*****
s6 0 NNNN**
```

This search produced only one output sequence, which has two copies in the *Giardia* genome. The sequence was then folded by RNAstructure (Mathews et al. 2004) and drawn in RnaViz-2.0 (De Rijk et al. 2003). The output structure has one more stem-loop (stem-loop 3 in Figure-4a) compared with *T. brucei*. Thus the *Giardia* candidate is intermediate between the standard eukaryotic pattern as found in human, and the reduced one in *T. brucei*. Expression of this

Giardia U1-snRNA candidate was confirmed by RT-PCR (Figure-4b). Structural modelling of this candidate shows that it is a good candidate for U1-snRNA.

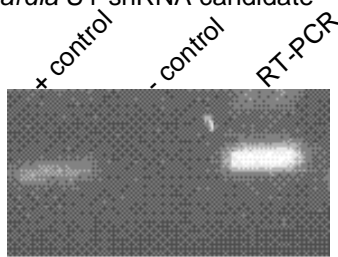
Figure-4: Identification of *Giardia* U1-snRNA candidate

(a) The structures of Human, *T. brucei* and *Giardia*-candidate U1-snRNAs



The conserved loops among the human, *Giardia* and *Trypanosome* U1-snRNAs are indicated by the circles. The Sm-protein-binding sites are boxed.

(b) RT-PCR test for expression of *Giardia* U1-snRNA candidate



RT-PCR results show highly expressed *Giardia* U1-snRNA candidate.

+ control: PCR with genomic DNA

- control: PCR with total RNA without reverse transcription

4.2.2 Prediction of *Giardia* U2-snRNA candidate

The same method was applied to search for U2 snRNAs from *Giardia*. However, this search did not give any results due to the high degree of specificity required for constructing the descriptor file. Subsequently, a more general approach was tried. The new approach used the available sequences of

U-snRNAs from Rfam (Griffiths-Jones et al. 2003) to search for the corresponding ncRNAs from *Giardia* genome using the INFERNAL software package (Eddy 2006). The INFERNAL software uses covariance models (Eddy 2002) which optimizes the aligning of an RNA sequence to a conserved RNA structure. The INFERNAL package is comparable to HMMER package, which builds profile Hidden Markov models in searching for homologous protein sequences from a database. Eukaryotic U-snRNAs from Rfam have been annotated with the INFERNAL package with multiple alignments and conserved secondary structures. Therefore, these alignments were used in searching for potential U-snRNAs from *Giardia* genome. The programmes **cmbuild** and **cmsearch** within the INFERNAL package were used here.

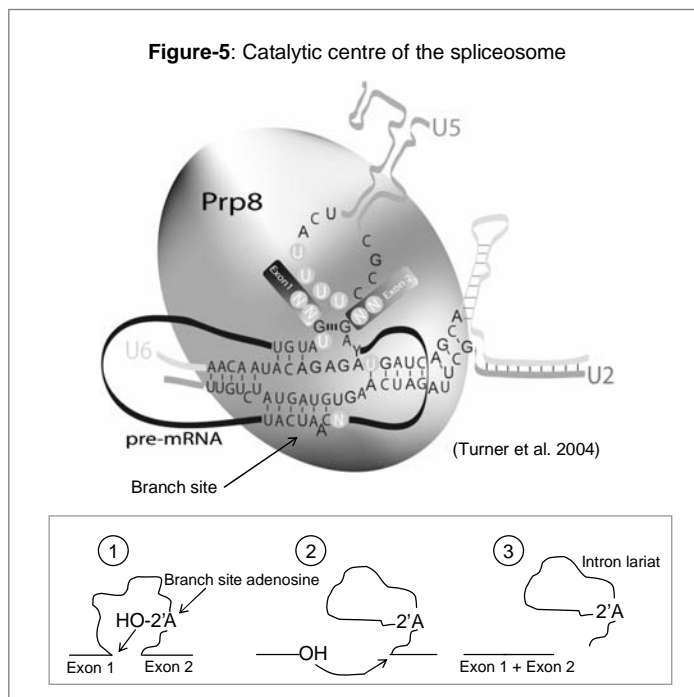
The alignments of U1, U2, U4, U5, and U6 were downloaded from Rfam in Stockholm format and Covariance models for these alignments were built using the **cmbuild** programme. Searching for potential U-snRNAs from *Giardia* genome was done by the **cmsearch** programme. An output hit from **cmsearch** consists of an alignment and score. By default, scores above 0 are considered as hits.

As a control, a **cmsearch** for U5 snRNA was performed first. Using the model built from the alignment of 33 seed-sequences resulted in 394 potential U5 sequences, as well as the experimentally confirmed U5 candidate (Collins et al. 2003). This control strengthened the likelihood of obtaining a true candidate. A second control searching for U1 candidates was also performed. However, the putative U1 candidate described above was not in the output which contains 29 sequences in total. The absence of the predicted U1 sequence in the output from **cmsearch** could be due to the high degree of conservation among the seed sequences used for building the Covariance model, thus the search may have bypassed possible *Giardia* U1 sequence which has one stem-loop less than a typical U1-snRNA. This important structural difference may have resulted in the searching algorithm bypassing the putative U1-candidate obtained above.

Different searching algorithms have varying degrees of sensitivity. The RNAbob programme used here is highly sensitive on searching RNAs with

known structures and conserved sequence motifs but requires enough information to construct a descriptor file. On the other hand, the INFERNAL software applies to more general searches using alignments of both sequences and structures of seeds* RNAs; however successful searches using this method largely depends on the prerequisite that the candidate RNA is highly conserved at both sequence and structural level with the seeds RNAs used for the search. In this study of *Giardia* U-snRNAs, it is not clear as to what degree *Giardia* U-snRNAs may be conserved with other known U-snRNAs, therefore it is necessary to use two searching methods of differing focus and sensitivity to achieve a high efficiency of finding the putative candidates.

Subsequently, U2, U6 and U4 candidates were searched in sequence. In the general model of eukaryotic spliceosome, the catalytic centre contains three U-snRNAs: U2, U6 and U5, positioned by the important scaffold protein Prp8 (Grainger and Beggs 2005; Turner et al. 2006). The centre of an active spliceosome is shown in Figure-5 (Turner et al. 2004).



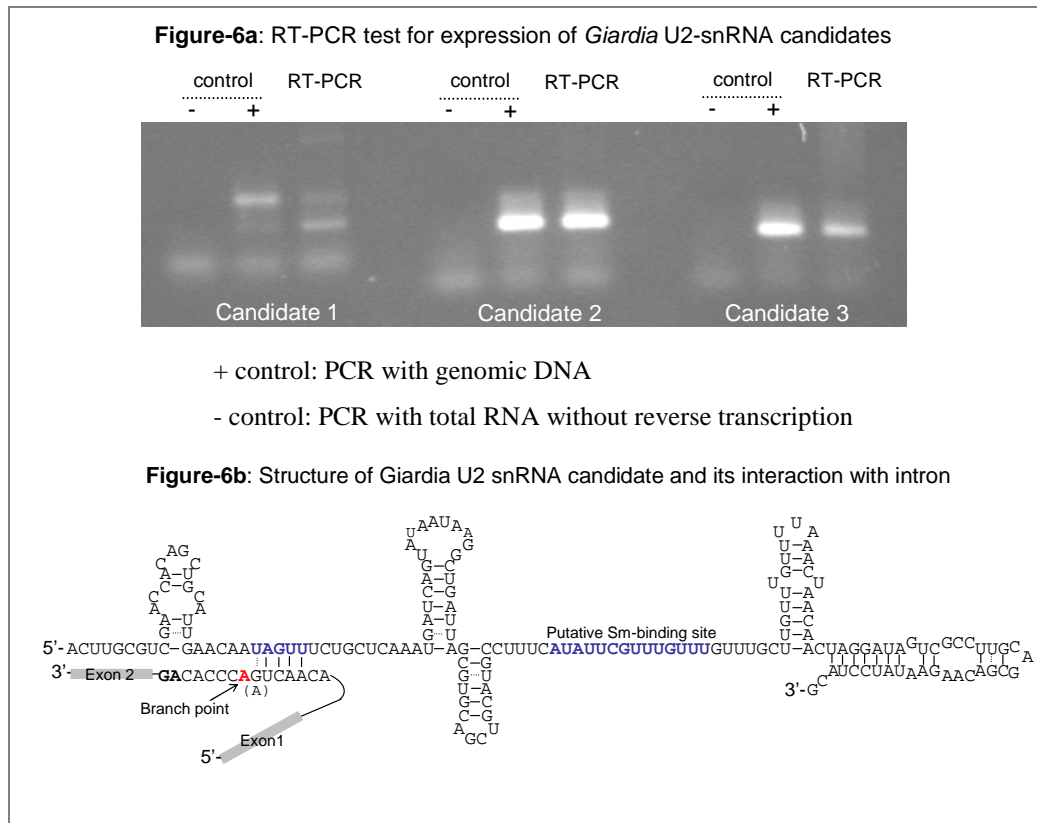
In this RNA-protein complex shown in the above figure, U2-snRNA loosely binds to the branch site of the intron, leaving the unbound branch-site adenosine, which can then interact with the phosphate group on the guanosine at 5' of the intron through its 2'-OH group, and form an

intron lariat. Therefore the bulged branch-site adenosine is crucial for the function of the spliceosome. U2-snRNA also binds to the U6. It is expected that any potential U2-candidate from *Giardia* must have a sequence motif

* Seeds: a representative set of known members of the same family that are used to construct an alignment for searching putative candidates in a sequence database.

complementary to the branch site. And based on the U2-U6 base-pairing, a U6 candidate may be found after U2.

First of all, **cmsearch** was run using models build from U2, U4 and U6 seeds and outputs were obtained. The outputs for U4 and U6 are large (217 and 1052 sequences respectively) whereas the output for U2 has only 5 hits. Blasting the hits for U2, U4 and U6 at the *Giardia* genome database (<http://www.giardadb.org/giardadb/>) showed that 3 of the U2 hits, 114 of the U4 hits and 649 of the U6 hits lie within non-coding regions. Since the number of potential U2 candidates is small, RT-PCR analysis was carried out to test the expression of these hits, though the small number of hits may not cover all possible U2 candidates. Results (Figure-6a) clearly show that two of the three candidates (candidate-2 and candidate-3) are expressed and candidate-2 is highly expressed. Although candidate-3 is also shown to be expressed, it appears much less abundant than candidate-2. Structural modelling (Figure-6b) and sequence analysis show that candidate-2 is more likely to be U2-snRNA. U2-snRNA is part of the catalytic centre of spliceosome. The likely U2-candidate shown in Figure-6b contains a “UAGUU” motif which complements the 5' of intron branch site “AACUG (or AACUA)”, but does not have upstream bases that can bind to 3' of the branch-site adenosine (coloured red), thus instead of leaving the branch-site adenosine bulged this interaction leaves an open-end of the branch site. However this alteration of branch-site recognition may not have any functional difference because the branch-site adenosine is still free to attack the 5'-guanosine phosphate. The overall sequence of this U2-snRNA candidate can fold into a typical U2-snRNA structure (see Figure-1) with the presence of a putative Sm-binding site, suggesting it to be a good candidate for U2-snRNA.



4.2.3 Prediction of *Giardia* U6 and U4 snRNA candidates

It is known that conserved base pairings form between U2 and U6, and between U6 and U4 snRNAs during the dynamic process of splicing. These conserved base-pairings are shown in Figure-7. In the U2-U6 hybrid, the central region of U6-snRNA folds into an intramolecular-stem-loop (ISL) structure, which is highly conserved in the active spliceosome and juxtaposes the regions interacting with U2-snRNA (Fortner et al. 1994). The ISL has been shown to have important roles in the catalytic centre of the spliceosome with the uridine (indicated by * in the *S. cerevisiae* model shown in Figure-7) serving as a binding site for an Mg^{2+} ion during the catalytic step of splicing (Huppler et al. 2002). This uridine is seen in all but two U6-snRNAs from Rfam (Griffiths-Jones et al. 2005), and the metal-binding uridine is usually situated below a “C·A” wobble base pair, which is readily protonated (Huppler et al. 2002). As mentioned in Chapter-1, the structure of U6 ISL is highly similar with the catalytic stem-loop structure of Group-II ribozyme (Sashital et al. 2004; Valadkhan 2005) and it appears that this structure has been maintained through evolution of the splicing mechanism (Lehmann and

Schmidt 2003; Seetharaman et al. 2006). In addition, two sequence motifs on the U6-snRNA are also conserved (coloured red in Figure-7a). The “ACAGAG” is involved in base-pairing with the 5'-intron site and the branch site (Sashital et al. 2004). The invariant “AGC” tri-nucleotide is seen in all identified U6-snRNAs recorded in Rfam (Griffiths-Jones et al. 2005), and has both structural and functional roles during splicing (Sashital et al. 2004). A recent study also showed that the “ACAGAG” loop and “AGC” tri-nucleotide were binding sites of Mg^{2+} (Yuan et al. 2007). U6 and U4 also form extensive base-pairings (Nottrott et al. 2002) as shown in Figure-7b. In this hybrid, the U6-snRNA has formed a 5'-stem-loop structure. Gathering all the sequence and structural features of U-snRNAs, Table-1 lists all the consensus properties used for searching U6 and U4 snRNA candidates. Searching for U6 and U4 snRNAs were based on the previous result that the U2 candidate identified here was highly likely to be the true *Giardia* U2-snRNA.

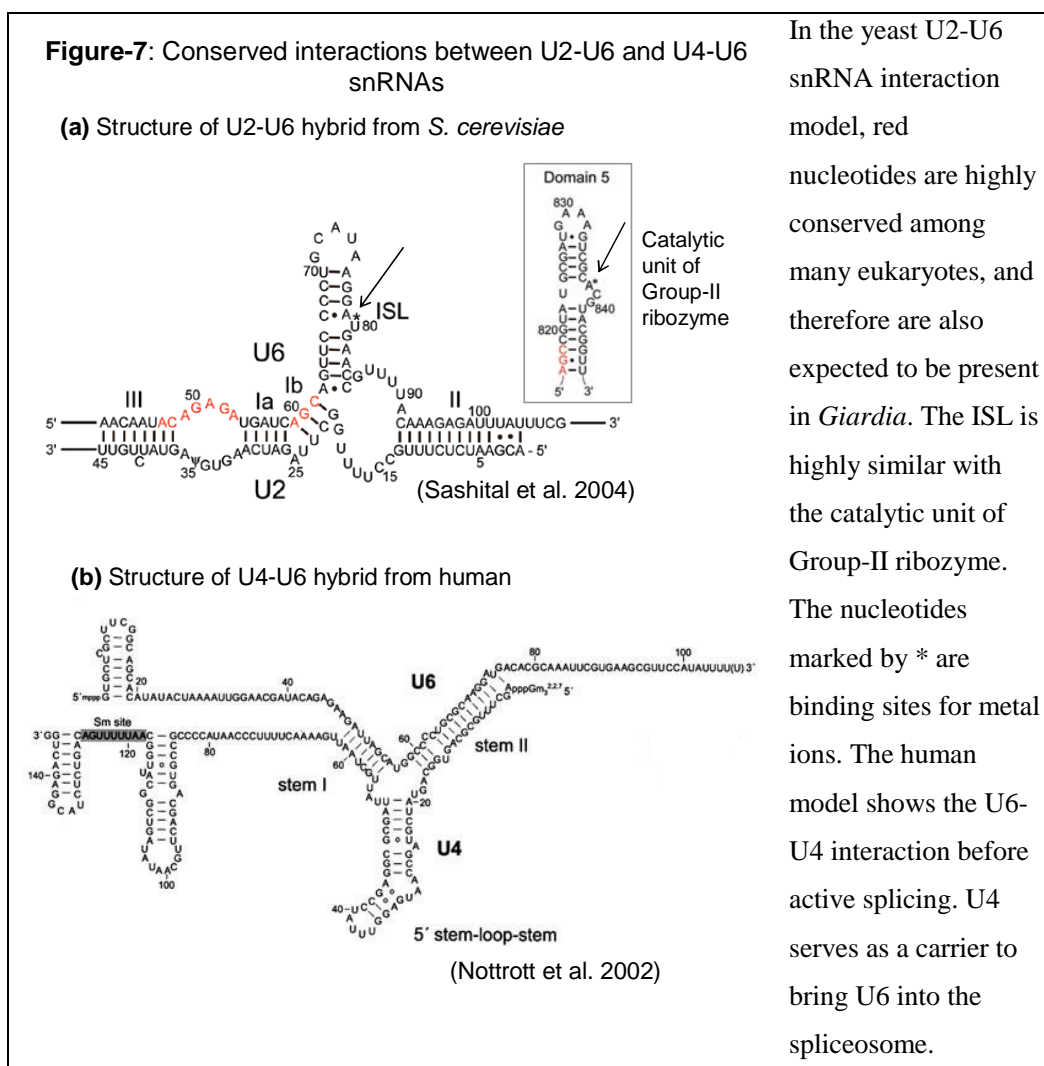


Table-1: Criteria for searching U6 and U4 snRNA candidates in *Giardia*:

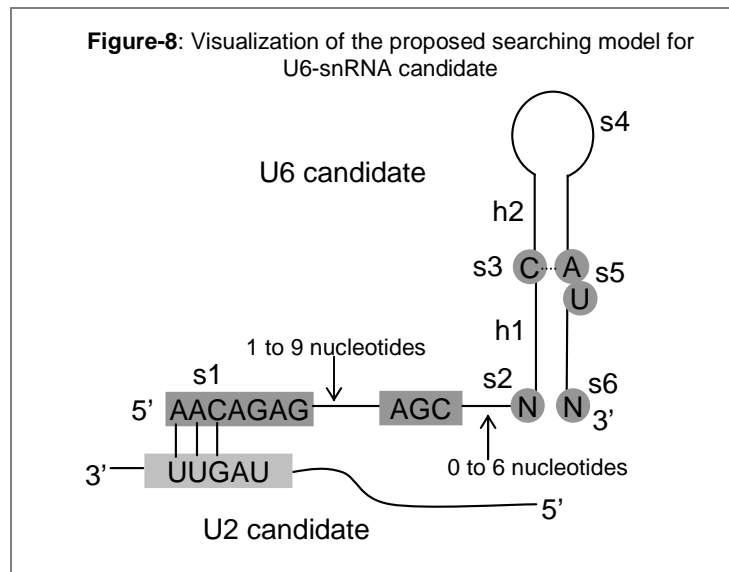
U-snRNA	Features
U6-snRNA	5'-stem-loop
	ISL with a bulged uridine, likely to be located below a "C-A" wobble pair
	ACAGAG motif
	AGC invariant tri-nucleotide
	Base-pairing with U2-snRNA on 5' and 3' of the ISL
U4-snRNA	GCT tri-nucleotide which base pairs with "AGC" tri-nucleotide of U6
	5'-sequence which base-pairs with U6 central region and sequence immediately after "GCT" which base-pairs with U6 near its 5'-stem-loop
	Sm-protein binding site (usually starts with 'A' followed by a number of 'U's and terminates with 'G')

A trial to search for *Giardia* U6-snRNA candidate was carried out first because there are more conserved features known for the U6-snRNA. A descriptor file for RNAbob programme was written based on the consensus features around the ISL, including the "AAC" motif which binds *Giardia* U2 at the 5' of the "ACAGAG" loop, the "ACAGAG" motif and "AGC" invariant tri-nucleotide which are two of the important characteristic features of U6-snRNA. The criteria used for writing the descriptor file can be visualized in Figure-8.

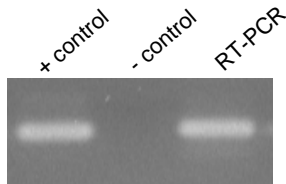
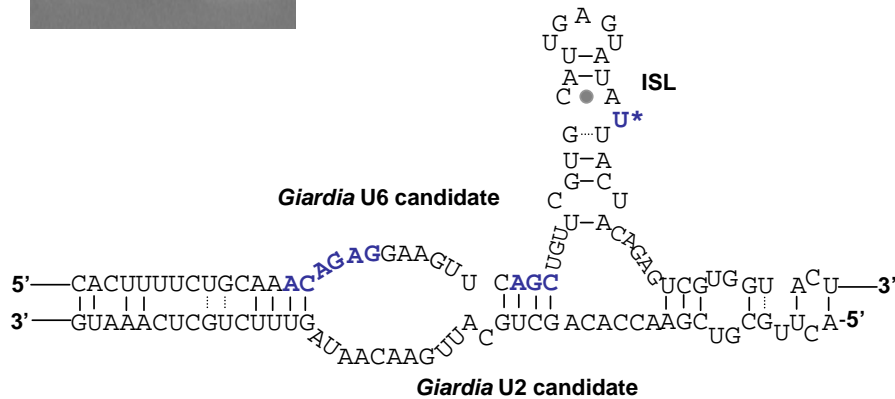
```
# U6_central region descriptor
# features: 1) AAC binding GUU of U2
#           2) ACAGAG and AGC conserved motifs
#           3) ISL with the catalytic 'U' below 'C-A' wobble base-
pair

s1 s2 h1 s3 h2 s4 h2' s5 h1' s6

s1 0 AACAGAGN*****AGC*****
s2 0 N
h1 0:0 *NNN:NNN*
s3 0 C
h2 0:0 *NN:NN*
s4 0 NNNN**
s5 0 AT
s6 0 N
```



The descriptor file shown above was then used to search against the whole genome sequence of *Giardia*, and gave 4 output sequences. By comparing with coding sequences, two of the four output sequences were eliminated. 40nt upstream and downstream sequences of the two output sequences were pulled out from the genome and analysed by eye. One of the remaining two sequences have all the compulsory features of U6-snRNA (see Table-1), therefore was identified as a candidate, even though this candidate is not found from **INFERNAL-cmsearch**. This may be due to the low sequence conservation between *Giardia* U6 and U6 from most other organisms which were used as seeds for constructing the **cmsearch** model. Sequence homology is the major method for searching U6-snRNAs in the majority of eukaryotes because sequences of U6-snRNAs are highly conserved among many eukaryotes. The covariance model built for **cmsearch** therefore very likely bypassed U6-snRNA during the search in *Giardia* genome. Indeed low sequence conservation was the major problem in identifying *Giardia* ncRNAs and earlier trials to look for U6-candidates failed with sequence homology search. RT-PCR test has confirmed that this potential U6-snRNA candidate is highly expressed. Results are shown in Figure-9a. Figure-9b shows the two-RNA-hybrid formed by the U2 and U6 snRNA candidates from *Giardia*. Conserved sequence elements on U6-snRNA candidate are coloured in blue.

Figure-9: Identification of *Giardia* U6-snRNA candidate**(a)** RT-PCR test for expression of the U6 candidate**(b)** Interaction between *Giardia* U6 and U2 snRNA candidates

Controls for the RT-PCR test:

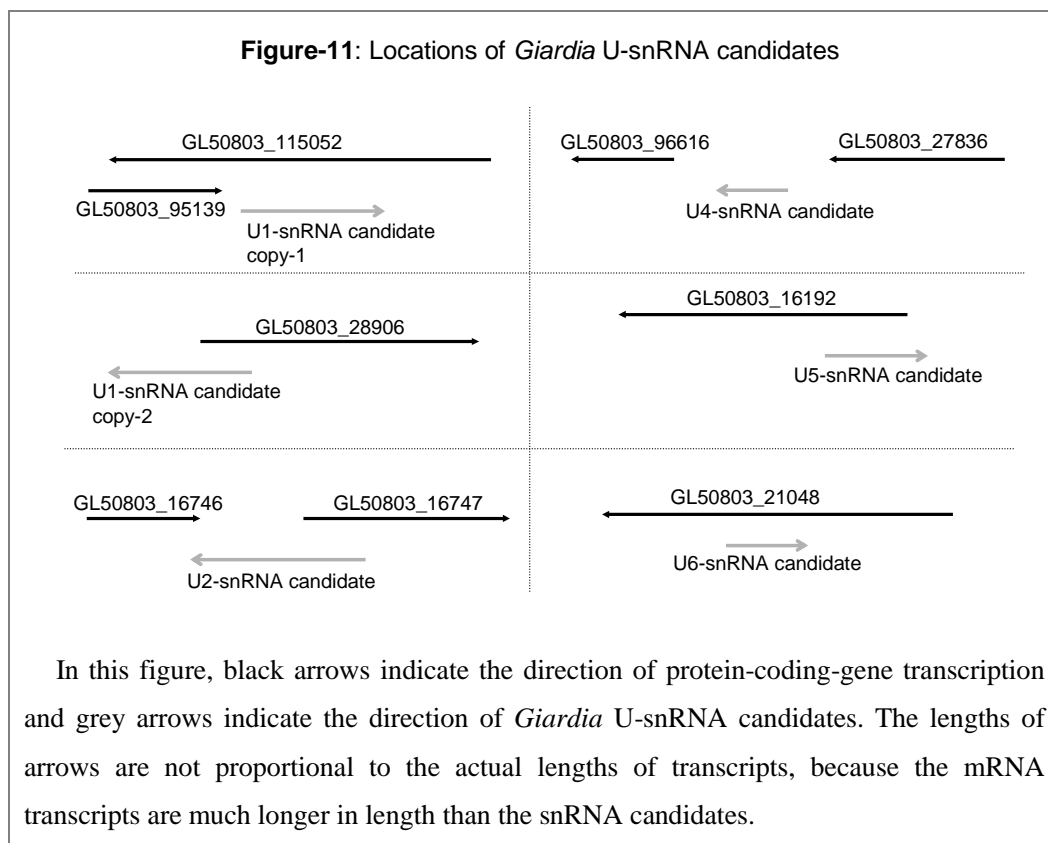
+ control: PCR with genomic DNA

- control: PCR with total RNA without reverse transcription

The U6 candidate was then used to search for a possible U4 candidate based on the conserved U6-U4 base-pairing feature shown in the human model in Figure-7b. First, a potential U4-snRNA candidate was searched for from the 114 output sequences of Infernal-**cmsearch**. A few sequences from **cmsearch** output contain a putative Sm-binding site and one of them shows base-pairing with the U6-snRNA candidate. Expression of this sequence was tested by RT-PCR and result (Figure-10a) shows clear and high expression. The interaction between *Giardia* U6 and U4 snRNA candidates is shown in Figure-10b.

et al. 1989); such example is seen in U6-ISL. The structural features of non-Watson-Crick base pairs are closely related with RNA catalysis, and also protein interactions. It has been shown that base substitutions within the “A·C” wobble base pair (above U80 in Figure-7a) severely impaired yeast growth and it was suggested that the defect might disrupt Prp24 protein binding and reduced stability of the U4/U6 hybrid (McManus et al. 2007). The wobble-base pairs are often evolutionary conserved in large ribozymes such as the ribosome (Mokdad et al. 2006), but less conserved in small ribozymes. In case of the snRNAs, the conserved “A·C” wobble base pair is the best studied example. Mutation of the bulged uridine within U6-ISL (see Figure-7a and Figure-9b) has been shown to be lethal due to its resulted alteration of “A·C” wobble base pair which is important for melting the U6-ISL during structural rearrangement necessary for association with U4-snRNA (Sashital et al. 2003). This important wobble base pair is also seen in the newly identified *Giardia* U6-snRNA candidate in my study. There have not been extensive studies on the roles of other wobble base pairs in snRNAs, and the other wobble base pairs (apart from the highly conserved “A·C” pair) in the structural modelling of *Giardia* U-snRNA candidates do not have conserved counterparts in other eukaryotes. However In the case of *Giardia*, a high degree of sequence divergence from most other eukaryotes causes difficulties in comparing the position of every base pair with other eukaryotes. Detailed biochemical experiments will be needed to fully verify the U-snRNA candidates identified here, but that is beyond the scope of this study. Structural modelling of the *Giardia* U-snRNA candidates shows all the expected features of eukaryotic U-snRNAs, therefore they are most likely to be the true *Giardia* U-snRNAs.

All five *Giardia* U-snRNA candidates are found in transcriptional intense regions of the genome; most of them overlap with protein-coding genes on the antisense strands. Except U1-candidate which has two copies (one copy has a base-substitution to the other one), the other candidates Figure-11 shows the locations of *Giardia* U-snRNA candidates in relation to the positions of nearby protein-coding genes.



The upstream sequences of *Giardia* U-snRNA candidates were also analysed. The upstream 100nt sequence for each U-snRNA candidate was extracted from the genome (see Appendix-3 for sequences). It is known that in most eukaryotes, the U6-snRNA is transcribed by RNA Pol III (Kunkel and Pederson 1988), and the other four snRNAs are transcribed by RNA Pol II. The general eukaryotic U6 promoter contains an upstream “TATA-box” and also upstream enhancer elements (Kunkel and Pederson 1988; Jensen et al. 1998). The potential *Giardia* Pol III promoter elements are discussed in Chapter-3, where the “A/T-element” and “G/A” element have been shown to be possible Pol III upstream elements. The upstream sequence of *Giardia* U6-snRNA candidate does not show strong signals of either “A/T-element” or “G/A-element”, but the absence of strong signals of either Pol II or Pol III promoter elements in the other four U-snRNA candidates shows that these candidates may be another example of ncRNA genes without clearly observable promoters. The same feature is seen in more than half of the uncharacterised novel ncRNAs identified in *Giardia* as discussed in Chapter-3.

In conclusion, this study has found four likely candidates of *Giardia* snRNAs through computational method, and confirmed that they are expressed. The sequences and genomic locations of five *Giardia* U-snRNA candidates are listed in Appendix-3. Combining sequence and structural information which summarises conserved features of characterised ncRNAs appears to be an efficient way of searching the unknown homologues of these ncRNAs in phylogenetically distant lineages. However, apart from the primary tests of expression, the *Giardia* U-snRNA candidates found here have not been extensively verified by biochemical methods such as functional knockout. It still remains uncertain as whether these candidates are truly U-snRNAs, but the characteristic structures, sequence motifs and RNA-RNA interactions indicate that they are very likely to be U-snRNAs. The following section describes a small-scale analysis of a central protein component of the spliceosome: the *Giardia* homologue of Prp8 protein.

4.3 *Giardia* homologue of Prp8 protein – the central protein component of the spliceosome

Formation of the catalytically competent spliceosome involves a series of protein-RNA rearrangements. A number of RNA-dependent helicases are required in these processes including Brr2, Prp5, Prp5, Prp8, Prp16, Prp17, Prp18, Prp22, Prp28 Prp43, Sub2, and Slu7 (de la Cruz et al. 1999; James et al. 2002). Among these proteins, Prp8 is the most highly conserved and involved in both the first and second trans-esterification reactions during splicing catalysis (Grainger and Beggs 2005). The Prp8 protein is a component of the U5snRNP (Lossky et al. 1987) and U5.U4/U6 tri-snRNP (Stevens and Abelson 1999), and can be UV-cross-linked to the 5'- splice site (SS) (Wyatt et al. 1992; Maroney et al. 2000), the branch point (BP) (MacMillan et al. 1994; McPheeters and Muhlenkamp 2003), the 3'- SS (Teigelkamp et al. 1995), and also to the U5 (Dix et al. 1998) and U6 snRNAs (Vidal et al. 1999). The interactions between Prp8 and RNA active sites suggest an essential function of Prp8 at the catalytic centre of the spliceosome (Collins and Guthrie 1999).

The Prp8 protein is evolved in both major and minor splicing (Lucke et al. 1997; Luo et al. 1999) and exhibits high degree of conservation across all eukaryotes from which it has been identified. For example, an overall 60% amino-acid sequence identity has been observed between the Prp8 protein of human and yeast (Hodges et al. 1995). Prp8 belongs to the PRO8 splicing-factor family, which has 72 proteins recorded in InterPro database (<http://www.ebi.ac.uk/interpro/IEntry?ac=IPR012591>). While highly conserved they do not show homology to other protein domains. These proteins are classified by their N-terminal PRO8NT domains located at N-termini. 28 full-length sequences of the Prp8 genes from 26 eukaryotic organisms are available and their large size is conserved, varying between 230 and 280 kDa (Grainger and Beggs 2005).

Most Prp8 proteins have nuclear localization signal peptides at the N-terminus, a 3'- splice site fidelity region at the middle, followed by a conserved RNA recognition motif (RRM), and an MPN domain at the C-terminus (Grainger and Beggs 2005). The *Giardia* homologue of Prp8 has been identified based on sequence homology (Nixon et al. 2002). Aligning the *Giardia* Prp8 protein sequence with those of other deep-branching unicellular eukaryotes showed that Prp8 protein homologues are highly conserved across the entire sequence (Appendix-3).

4.3.1 Bioinformatical analysis of the *Giardia* homologue of Prp8 protein

Being the central protein component of the spliceosome, extensive biochemical studies have been carried out in order to understand the functional domains of the Prp8 protein using yeast as a model (Grainger and Beggs 2005; Turner et al. 2006). The Prp8 protein is known to interact with U5-snRNA (Turner et al. 2006) as a conserved central protein component of the spliceosome. The presence of a highly conserved Prp8 homologue in *Giardia* (*Giardia* genome ID: GL50803_112114) suggests that this protein may have the same functions in *Giardia* as those known for higher eukaryotes. In order to obtain more information about *Giardia* Prp8, biochemical studies are necessary. However, before any experimental studies could be carried out, detailed

analysis of the amino-acid sequence and possible structural properties were needed. A number of bioinformatics tools have been used to analyze possible functional domains and secondary structures of the *Giardia* Prp8 protein.

In addition to being conserved with Prp8 homologues from various unicellular organisms, *Giardia* Prp8 is also highly conserved with the yeast Prp8 protein. Based on the sequence alignment of *Giardia* and yeast Prp8 proteins, the functional domains on the yeast Prp8 protein can be mapped by eye onto the *Giardia* Prp8 protein (Appendix-3). Table-2 lists the functional domains of yeast Prp8 protein, and Figure-12 shows the location of potential functional domains within *Giardia* Prp8 protein mapped in my study.

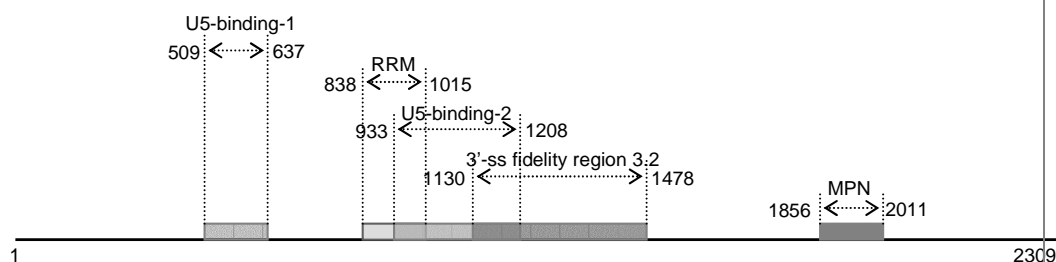
Table-2: Functional domains of Prp8 proteins from *S. cerevisiae*.

Domain	Position (aa)	Function
U5-snRNA-binding site 1	770-871	Interaction with U5-snRNA
U5-snRNA-binding site 2	1281-1413	Interaction with U5-snRNA
RNA recognition motif	1059-1151	Possible interaction with 5'-ss site, 3'-ss site, U5-snRNA and U6-snRNA
3'-splice site filelity region 3.2	1372-1660	Highly conserved and likely to be responsible in promoting RNA-mediated catalysis; overlaps with U6-interaction domain (aa 1503-1673)
MPN domain*	2178-2310	Regulating protein-protein interactions, may be specific to higher eukaryotes only

(Maytal-Kivity et al. 2002; Grainger and Beggs 2005; Turner et al. 2006)

* MPN domain (Maytal-Kivity et al. 2002): highly conserved in a number of MPN-domain proteins such as Rpn11 and Csn5/Jab1. The MPN domain consists of five polar residues that resemble the active site residues of hydrolytic enzyme classes, particularly that of metalloproteases.

Figure-12: Potential functional domains in *Giardia* Prp8 protein from comparison with *S. cerevisiae*



The putative domains on the *Giardia* Prp8 protein were mapped by eye from the sequence alignment of *Giardia* and yeast Prp8 proteins. All the functional domains of yeast Prp8 protein can be aligned with high degree of sequence similarity with parts of the *Giardia* Prp8 protein, and the corresponding positions on the *Giardia* Prp8 protein are shown above. The putative RRM, U5-binding domain-2 and 3'-ss fidelity region 3.2 are overlapping on the *Giardia* Prp8 protein sequence.

The *Giardia* Prp8 protein is a large protein of 2309aa (approximately 260 kDa), and it is not possible to construct a small-scale protein analysis on such a large protein (personal communication with Dr. Gill Norris). According to Figure-12, it appears that the putative RNA-binding domains are clustered at the centre of *Giardia* Prp8 protein. In this study, the potential RNA-recognition motif (RRM) is selected as a candidate domain to analyse. A recombinant peptide (named Gp8d1) containing the potential RRM has been studied here and details are shown below. The peptide Gp8d1 is 249aa in length corresponding to amino-acid position 843 to 1082 on *Giardia* Prp8 protein.

4.3.2 Analysis of the potential RNA-recognition motif of *Giardia* Prp8 protein

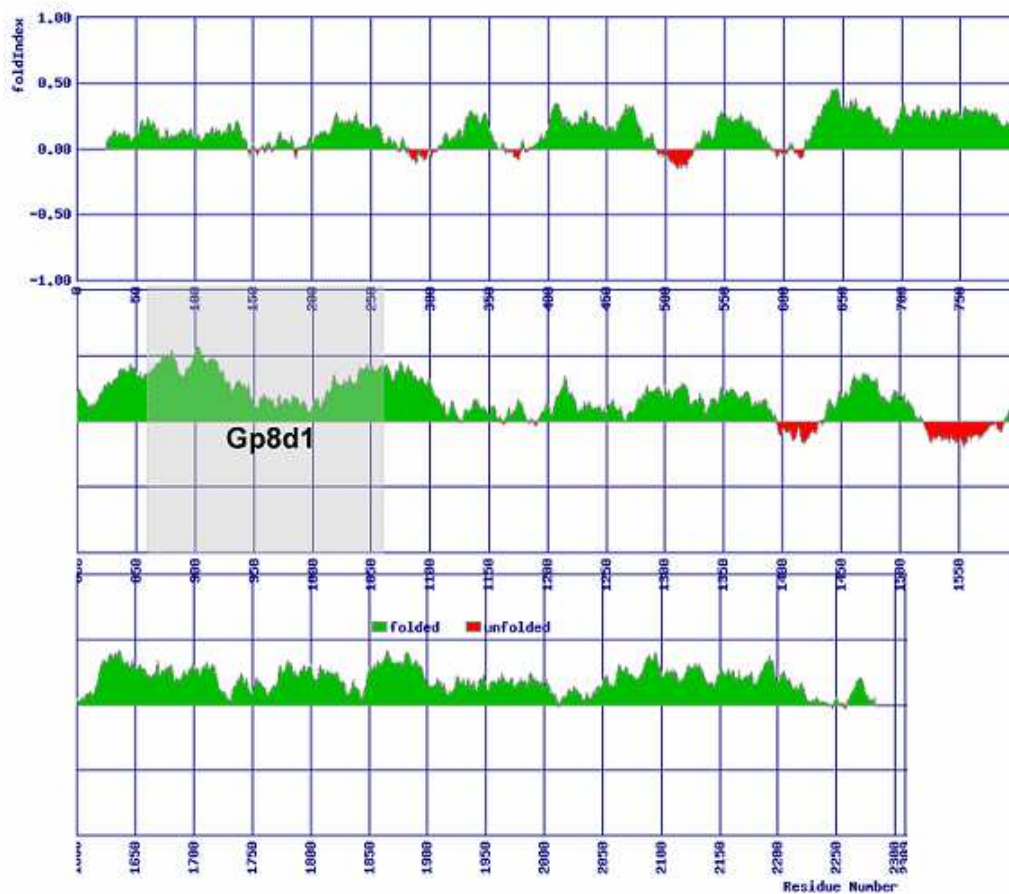
4.3.2.1 Computational analysis of the likelihood of Gp8d1 being a protein domain

Prior to biochemical study of previously uncharacterised protein domains, it is usually necessary to learn the possible folding of the particular domain to increase the possibility of success in experiments. Most of the characterized protein domains resemble globular form but it has been noticed that many functionally important protein segments lie outside globular domains in regions that are intrinsically disordered* (Wright and Dyson 1999), and may only become ordered when bound to another molecule (Dunker et al. 2001; Uversky 2002). There are a number of computational methods for analysis of protein globularity and disorder, such as DisEMBL (Linding et al. 2003a), GlobPlot (Linding et al. 2003b) and FoldIndex (Prilusky et al. 2005) etc.

Using the three computational methods mentioned above, the globularity and disorder of Gp8d1 in comparison to the whole *Giardia* Prp8 protein was analysed. First a rough scan of the full-length *Giardia* Prp8 protein sequence was performed using FoldIndex (<http://bip.weizmann.ac.il/fldbin/findex>). The output (Figure-13) shows that the entire Prp8 protein is likely to be ordered (green) with a few short disordered connection segments (red). The position of Gp8d1 is at the central region of the protein and indicated by grey shade. Prediction shows that Gp8d1 is highly ordered at both ends and less ordered at its centre.

* Intrinsic disorder: This term refers to segments or whole proteins that fail to self-fold into fixed 3-D structure.

Figure-13: General Foldability of *Giardia* Prp8 protein



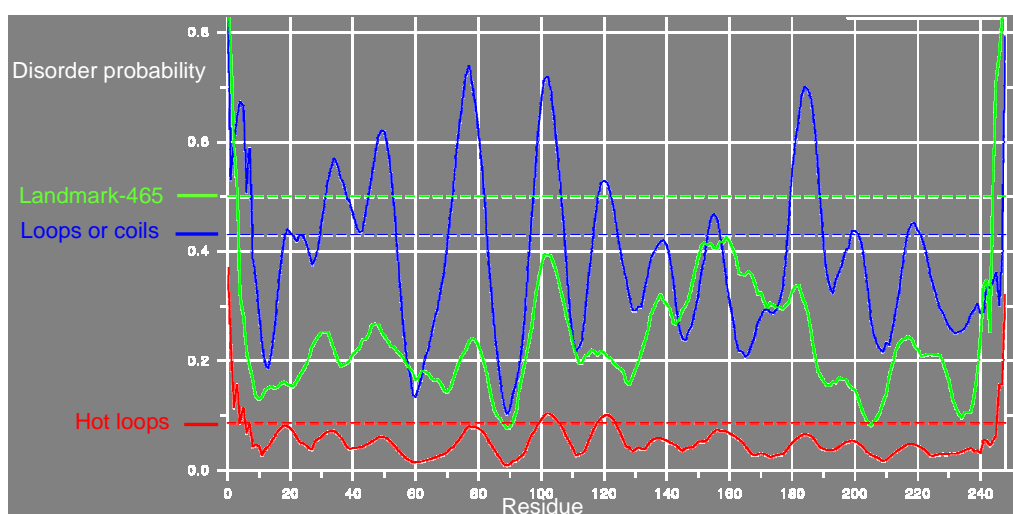
The order/disorder of a protein sequence is determined by the foldability of the sequence. The green regions shown in the figure represent folded regions. The height of the green region indicates the likelihood of the residue being in a folded region. Red colour represents unfolded regions. The *Giardia* Prp8 protein appears highly folded, with its central regions most folded. Gp8d1 (as shaded in grey) is located in the middle part of the Prp8 protein and is highly folded at two ends.

Further tests for protein disorder of Gp8d1 were done using DisEMBL (<http://dis.embl.de>) and GlobPlot (<http://globplot.embl.de>). Both methods predict putative domains within a given amino-acid sequence. Results are shown in Figure-14. In Figure-14a, the output of DisEMBL shows three types of putative structures within the Gp8d1 sequence. The dotted lines indicate thresholds for structure definition. The term “Loops or coils” corresponds to residues that are predicted as within helices or strands which are necessary but

not sufficient determinants for protein disorder. The “Hot loops” defines highly dynamic and mobile loops and is considered protein disorder. “Landmark 465” is a term used in X-ray structure with non-assigned electron densities and often reflects intrinsic disorder. As shown in the results (Figure-13), Gp8d1 sequence appears to be moderately ordered, consisting mainly helices and sheets with highly dynamic residues at its N- and C- terminus. Output of GlobPlot is consistent with DisEMBL. It has been tested in other studies that the downhill region in GlobPlot curve often co-locates with characterised protein domains (Linding et al. 2003b).

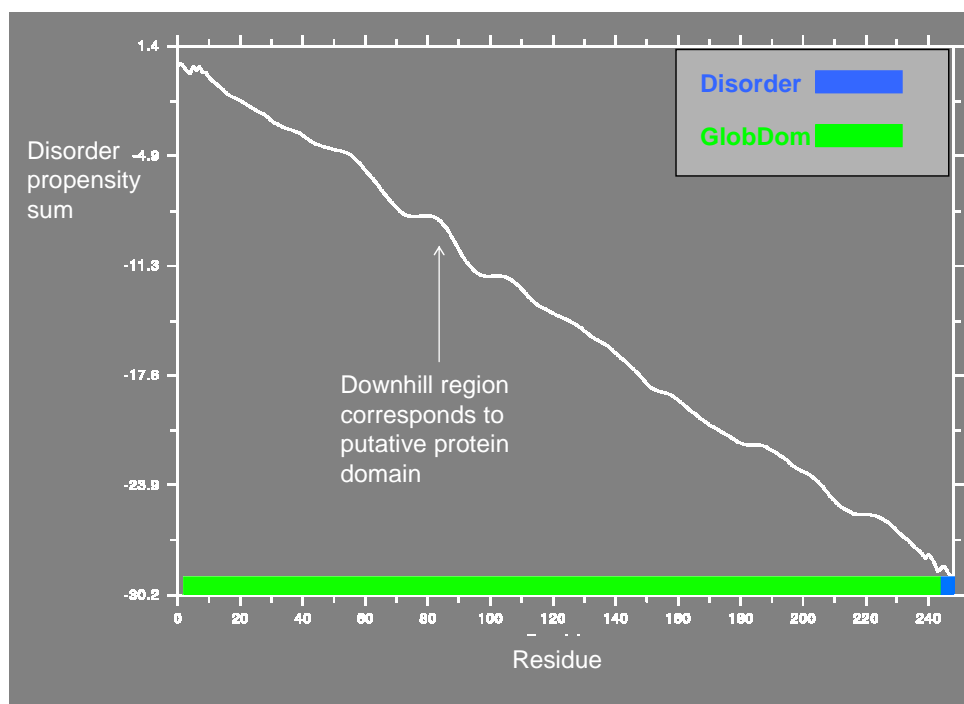
Figure-14: Analysis of intrinsic protein disorder using DisEMBL and GlobPlot

(a) Output of DisEMBL analysis



In the above figure, the dotted lines indicate the threshold for each type of unstable protein structures (disordered): hot loops, loops of coils and landmark-465. Regions that are above the thresholds are most probably to be disordered. The probability of disorder (being any one of the three type of unstable structures) of residues along the Gp8d1 sequence is plotted as curves. It is clear that most regions of the Gp8d1 are below the thresholds for being hot loops or landmark-465, while many residues are likely to be parts of loops or coils. Also it can be seen that the probability of disorder increases dramatically at the two termini of the peptide. The result indicates that the Gp8d1 peptide most likely folds into a moderately ordered structure, thus is likely to contain a protein domain.

(b) Output of GlobPlot analysis



The GlobPlot test is an alternative way to analyse protein disorder by defining regions of globularity and disorder based on a running sum of the propensity (P) of amino acids to be in an ordered or disordered state. P is expressed as $P=RC-SS$, where RC and SS are the propensity of a given amino-acid to be in “random coil” and regular “secondary structure”. The frequencies of RC and SS for each amino-acid has been calculated based on a database containing one representative sequence from each protein family (Linding et al. 2003b). A reducing sum of P (the curve in the figure) indicates that the residues along the protein sequences are more frequently in SS than in RC thus are more likely to be ordered and result in a defined protein domain.

With the above three tests performed, it was rather certain that the Gp8d1 sequence should fold into a single protein domain, which is likely to undergo self-folding. However the analysis does not reveal any highly dynamic region (that is highly disordered) within this peptide. To test the potential RNA-binding property of this peptide, recombinant Gp8d1 was made as described below. Detailed experimental methods are detailed in section 4.5.

A number of *E. coli* expression constructs were made, but all failed due to unknown reasons resulting in no expression under a variety of conditions. Table-3 lists the different vector constructs tried. Induction of protein expression at various temperatures (37°C, 25°C and 16°C) and different

concentrations of induction chemical (IPTG or arabinose) were tried, but the results did not change. Due to time restriction, only the standard BL21 expression cell and KRX expression cell were used. It was suspected that certain feature of this peptide leads to toxicity to *E. coli*; therefore *in vitro* protein expression was tried.

Table-3: Vector constructs used for expression of Gp8d1 peptide in *E. coli*.

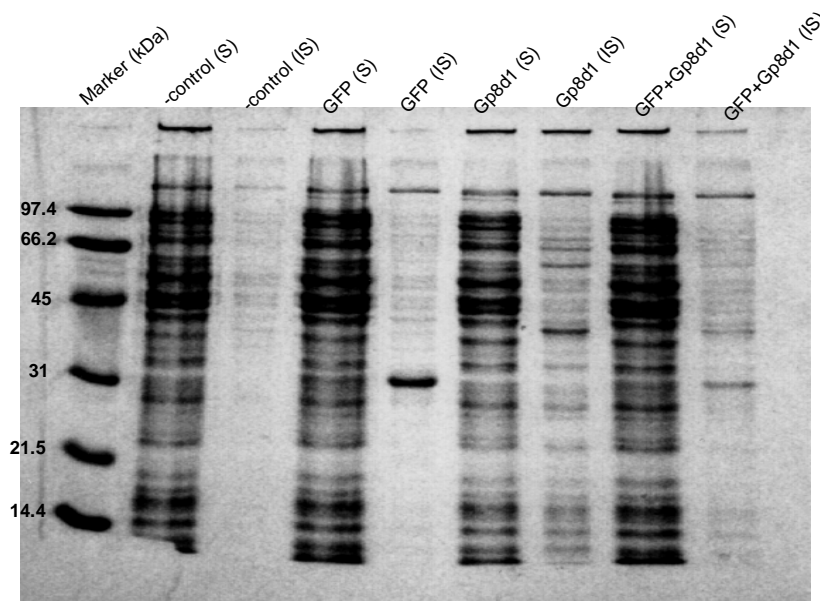
Vector	Feature	Result
pET-24b (BL21 expression cell, Invitrogen)	Basic vector construction for His-tag recombinant protein expression driven by T7 promoter, induced by IPTG	No expression
pETDuet_DsbC (BL21 expression cell, Invitrogen)	Co-expression with <i>E. coli</i> DsbC (disulphide isomerase II) peptide to aid correct folding of the recombinant peptide	No expression
pETDuet_MalE (BL21 expression cell, Invitrogen)	Co-expression with <i>E. coli</i> MalE (maltose-binding protein) peptide to increase solubility of the recombinant peptide	No expression
pIVEX-2.4d (KRX expression cell, Invitro Technologies)	Expression is activated by an inducible (by L-arabinose) production of a genomic copy of T7 polymerase	No expression

4.3.2.2 Cloning and *in vitro* recombinant protein expression

The potential RNA-binding domain (Gp8d1) of Prp8 was amplified by PCR from *Giardia* genomic DNA with restriction sites tagged primers, inserted into an *in vitro* expression vector (pIVEX-2.3d) and cloned into *E. coli* DH5 α cells for purification of plasmid. The purified plasmid was then used in the *E. coli* *in vitro* expression system for recombinant protein expression. The detailed protocol is described in Materials and Methods of this chapter.

The recombinant protein was produced and results are shown in **Figure-15**. Compared with the positive control of recombinant GFP expression, the expression level of Gp8d1 was much lower, and most of the recombinant proteins (both GFP and Gp8d1) were present in the precipitates. The precipitation may be due to the salt concentration in the expression mix, therefore buffer with higher salt concentration (500mM NaHPO₄, 0.02% Triton-X 100) was used in order to dissolve the precipitates. This turned out to be partially effective as part of the precipitated protein could be dissolved in the buffer used and GFP protein showed green fluorescence in UV light.

Figure-15: *in vitro* recombinant protein expression of potential RNA-binding domain in *Giardia* Prp8 protein.



Marker: BioRad low-range protein standard; S: soluble; IS: insoluble;

This figure compares expressions of the – control, the + control (GFP) and Gp8d1 recombinant peptide. There is not much visible difference among the soluble fractions of the – control, the + control and Gp8d1 expression mixture. There are clear bands showing expression of the GFP and Gp8d1 in the insoluble fractions shown on the gel. Co-expression of GFP and Gp8d1 reduced the expression level of both peptides.

Since the expression level of *in vitro* expression system is low, it was not possible to purify the recombinant protein by traditional methods of column chromatography despite that the recombinant protein was His-tagged at the N-terminus. Therefore two alternative methods were used to assay the proposed RNA-binding property of this domain.

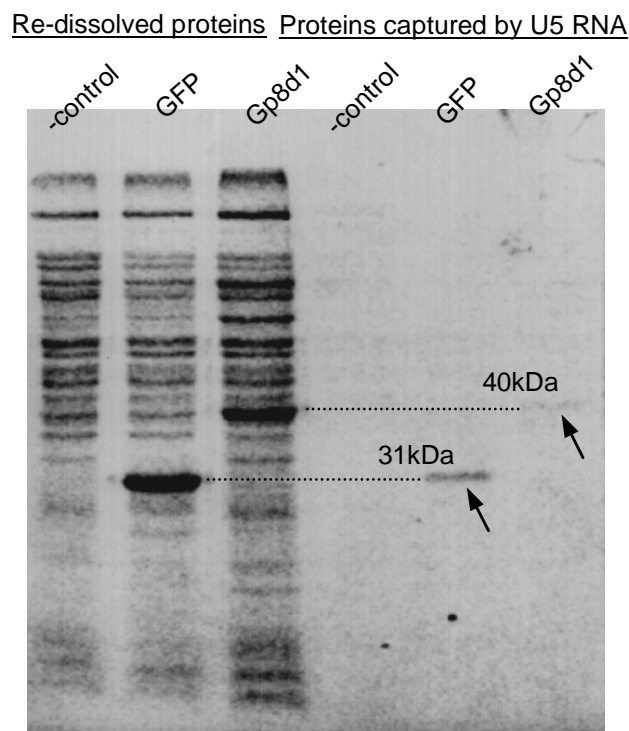
4.3.2.3 RNA-protein binding assays of Gp8d1 versus *Giardia* snRNA candidates.

The actual functions of the RRM-domains of Prp8 proteins in model eukaryotes studied to date remains uncertain (Turner et al. 2006). However current knowledge does not rule out the possibility that this domain can interact with a number of snRNAs. Mutation within the RRM could affect U4/U6 unwinding and it has been proposed that the RRM-domain of Prp8 may interact

with the U6-snRNA to regulate the formation the active spliceosome (Grainger and Beggs 2005). To test the RNA-binding ability of the recombinant Gp8d1, two approaches followed.

The first approach was affinity binding using a poly-A tailed U5-snRNA candidate. The U5-snRNA candidate was PCR-amplified and 3'-extended with a poly-A tail. The purified RNA was than hybridised with biotinylated oligo-dT primers and immobilized on the inside surface of streptavidin-coated tubes. Expression mixtures of Gp8d1, GFP and a negative control mixture containing no recombinant protein was added to the tubes, which were incubated to allow binding. The tubes were then washed three times to remove anything that did not bind to the RNA. Finally SDS-PAGE denaturing buffer was added to the tubes to break any possible interactions between protein and RNA. The final solutions in the tubes were loaded on to an SDS-PAGE gel for analysis. Results are shown in Figure-16.

Figure-16: Capturing proteins capable of binding to the U5-snRNA candidate

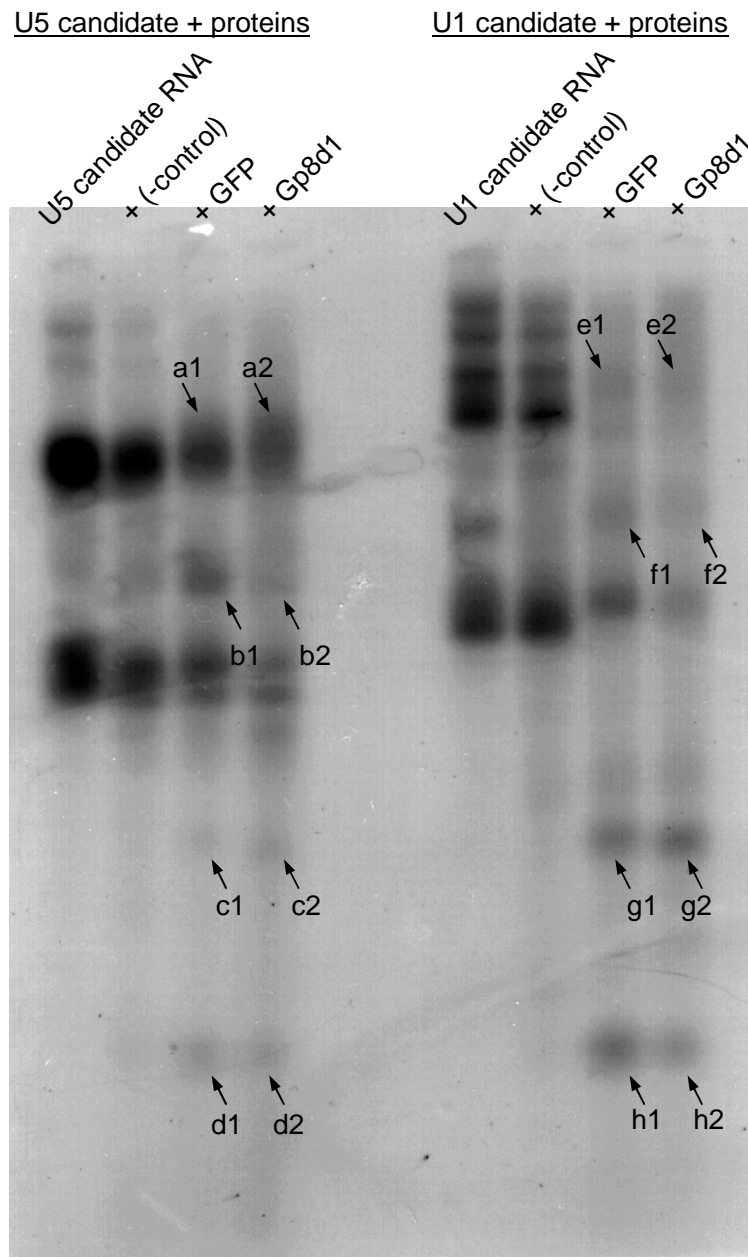


The arrows here indicate the proteins remained in the tubes after three washes. The sizes of two indicated bands correspond to GFP and Gp8d1. There are no visible bands in the lane showing '- control', indicating no protein left after washing.

Unexpectedly the lane showing the capturing result for GFP expression mixture showed a distinct band corresponding to the GFP protein, whereas the lane showing the result for Gp8d1 expression mixture only revealed a faint band corresponding to the Gp8d1 recombinant protein. The negative control lane clearly showed no proteins remaining after washing, therefore the results suggest that GFP protein has higher affinity for the U5-snRNA candidate than the Gp8d1 protein. Given the earlier observation that the GFP protein glowed in UV light, therefore it has folded correctly, and also it is certain that GFP protein does not bind any RNAs; the band seen here may be a result of residual protein not being washed away completely. Hence the presence of the Gp8d1 band may not indicate interaction between Gp8d1 and U5-snRNA candidate. To tackle the problem, a different method was used.

The second method to study RNA-binding property of Gp8d1 was a gel shift assay. Gel shift assays are widely used for analysis of protein-nucleic acid binding. Due to the fact that the proteins used in this study were not purified from reaction mixtures, radio isotope labelled RNAs were used to visualize the gel. The expression mixtures of negative control, GFP and Gp8d1 were incubated with ³²P-labelled U5- and U1-snRNA candidates before loading onto a native PAGE gel. At this stage of study, only the U1-snRNA candidate had been identified, and used as a control RNA because there is no evidence for U1-snRNA interacting with the RRM. Results (Figure-17) are, however again difficult to resolve. It would have been helpful if a positive control with known RNA-binding protein was included, but there was none available at the time when this experiment was done.

Figure-17: Gel shift assays of negative control, GFP and Gp8d1 expression mixture versus U1- and U5-snRNA candidates.



Both lanes of RNAs without proteins show several bands, which likely indicate differently folded RNA molecules. The negative control expression mixture did not show any changes in RNA mobility. Addition of GFP and Gp8d1 expression mixtures into U5-snRNA candidates resulted in slight mobility shift, as indicated by arrows a1, a2 and b1, b2, however consistent with the earlier results obtained from affinity capture, the effect of GFP appears stronger than Gp8d1. In the same assay, addition of GFP and Gp8d1

expression mixture also resulted in the appearance of potential RNA-cleavage products c1, c2 and d1, d2. In the parallel assay with the U1-snRNA candidate, the reducing intensity of original U1-snRNA candidate bands and the appearance of bands g1, g2 and h1, h2 also suggested RNA-cleavage upon addition of GFP and Gp8d1 expression mixtures. The potential mobility shift (e1, e2 and f1, f2) detected in the U1-candidate control assay was less noticeable than the U5-candidate assay.

The mobility shift assay revealed unexpected results. It is clear that addition of recombinant GFP and Gp8d1 resulted in RNA degradation in possibly specific positions of RNA sequences (as indicated by defined bands in Figure-17). It is not known what caused this result. However, both the U5-candidate assay and U1-candidate control assay suggest that there may be unspecific interactions between the RNA molecules and recombinant peptides.

In summary, this primary study of the putative RRM of *Giardia* Prp8 has encountered many problems, and the main reason behind this situation is likely to be in part the uncertainty of both the RNA candidates and selection of protein segments analysed. There is very limited knowledge on the spliceosomes of unicellular protists, and it is always difficult to begin with little information. The results obtained here will hopefully aid future research to carry on with the study of the *Giardia* spliceosome.

4.4 Conclusion and overview of the major spliceosomal components in *Giardia*

This chapter mainly focused on the identification of major spliceosomal snRNAs in *Giardia*. Using computational methods based on known structural and sequence information of eukaryotic snRNAs, four *Giardia* U-snRNA candidates were uncovered from the genome, and expression of these candidates were confirmed by RT-PCR. *Giardia* U-snRNA candidates can fold into characteristic stem-loop structures and conserved interactions are observed

in the models of U2-U6 and U6-U4 hybrids. This study has shown that combined structural and sequence search has the ability of identifying expected ncRNAs in a genome highly diverged from other eukaryotes. Identification of all the *Giardia* U-snRNA candidates suggests that *Giardia* has a full spliceosome similar with other eukaryotes. Although the RNA candidates identified here require further verification, the primary results are promising.

The small-scale analysis of the potential RNA-binding domain in *Giardia* Prp8 protein was not successful mainly due to limited background information and was not further investigated. However, as seen from a number of computational analyses, *Giardia* Prp8 protein is highly conserved with human and yeast Prp8 protein, and hence is likely to be functionally similar as well. It is expected that future studies will be able to reveal the biochemical details of the *Giardia* spliceosome.

4.5 Experimental Materials and Methods

4.5.1 PCR amplification and cloning

DNA encoding Gp8d1, U1- and U5-snRNA candidates were amplified by PCR from genomic DNA using HiFi DNA polymerase (Invitrogen, Gp8d1 reaction) and Taq polymerase (Roche, U1 and U5 reactions). PCR primers used for amplification of Gp8d1 fragment, U5 and U1 snRNA candidate genes are listed as following:

Gp8_d1_F_pIVEX	ATAGGCGGCCGCCTTAATACTGATAGCTACTT
Gp8_d1_R_pIVEX	TCGAGTCGACGCTACGATTAAGCTCATC
GiU5For	CATTCATCTCTGCGGTGGATG
GiU5Rev	ACCCCAAAAATGCAACTGTCTGCC
GU1_cand_1_F	AAACATCAGCGGCATCGTCA
GU1_cand_1_R	CGGACATCACCCGCCAAAA

PCR products of U1- and U5-snRNAs candidates were inserted into pGEM-T-easy T/A cloning vectors (Promega) and re-amplified by PCR using the universal forward primer (GTTGTAAAACGACGGCCAGT) and the U1- or U5-specific reverse primer to obtain the DNA templates for *in vitro* transcription.

The PCR product of Gp8d1 was double digested by Not-1 and Sal-1 restriction enzymes (Fermentas) and ligated into Not-1/Sal-1 double digested pIVEX-2.3d vector (Roche). The ligation reaction was carried out using T4-DNA ligase (Roche) at 4°C overnight. 1:5 dilution of 1µl ligation reaction was then used for heat-shock transformation into *E. coli* DH5α cells.

For transformation, 50 µl of *E. coli* DH5α cells was thawed on ice before addition of the diluted ligation mix. The cells were incubated on ice for 20min and heat shocked for 45 sec at exactly 42°C. Then the cells were immediately transferred on to ice and 300 µl room temperature S.O.C medium was added. The transformed cells were then shaken at 37°C for 1 h at 225 rpm. 50ul cells were plated on LB agar plate containing 100 µg/ml ampicilin. The plate was incubated at 37°C overnight.

4.5.2 Plasmid preparation:

All the plasmids were prepared from overnight cell cultures inoculated by single colonies from LB agar plates. Cells were collected by centrifugation and resuspended in 0.2ml TE buffer A (50mM Tris, 10mM EDTA) with addition of 40 µg/ml RNase A and 0.2ml alkaline lysis solution (8g/l NaOH, 1% SDS). The lysis solution was kept at room temperature for 15min and 0.2ml 3M NaOAc (pH 5) was added. The solution was then incubated on ice for another 15min and centrifuged at 4°C to collect the supernatant. The supernatant was centrifuged again and the final supernatant contained mostly purified plasmid DNA. Plasmid DNA was precipitated using ice cold 95% EtOH and resuspended in TE buffer B (10mM Tris, 1mM EDTA), and purified further using the PCR product purification kit (Roche, Cat# 11 732 668 001).

4.5.3 *In vitro* recombinant protein expression

The *in vitro* recombinant protein was expressed using the Rapid Translation System RTS-100 *E. coli* HY-Kit (Roche, Cat# 3 186 148) according to the standard protocol. The following reagents were added in order into tubes on ice to make a 50 µl reaction:

12 µl *E. coli* lysate

10 µl reaction mix

12 µl amino acids mix (without Methionine)

1 μ l Methionine

5 μ l reaction buffer

10 μ l plasmid DNA (0.5 μ g GFP or Gp8d1 plasmid DNA) or distilled H₂O (-control)

The reaction mixtures were then incubated at 30°C with shaking at 150 rpm for 6 h. After the reactions were finished, the reaction mixtures were centrifuged at 10000 rpm for 5min at 4°C and precipitants were isolated from supernatant. The precipitants were then resuspended in protein-RNA binding buffer (20mM Tris, 10mM MgCl₂, 300mM KCl, pH 7.5) and stored at -20°C.

Aliquots of the resuspended precipitants and supernatants were analyzed by 10% SDS-PAGE. (10% resolving gel, 6% stacking gel, running at 120V in Tris-glycine buffer and stained with Bio-Safe Coomassie stain (BioRad, Cat# 161-0786).

4.5.4 *In vitro* RNA transcription

In vitro RNA transcription was done using T7-RNA polymerase (Invitrogen, Cat# 18033019). The reaction mixtures were assembled in two steps at room temperature. First, 5 μ l 10 \times reaction buffer, 5 μ l 10mM rNTP mix (with 20 μ Ci of ³²P-UTP added for making radio-isotope labelled RNAs), 5 μ l DNA template from PCR were mixed and the volume was adjusted to 48 μ l. The mixture was incubated at 37°C for 1 h and 1 unit of RNaseOUT and 1 unit of T7-RNA polymerase were added. The reaction continued at 37°C for 2 h, and the DNA template was digested by addition of 1 unit of DNaseI. The RNA product was extracted with phenol:chloroform (5:1, pH 5) and then chloroform and precipitated by ice cold 100% EtOH.

4.5.5 Affinity capturing of proteins with ability to bind U5- and U1-snRNA candidates

The *in vitro* transcribed U5-snRNA candidate was first extended at 3'-end using Poly-A polymerase (Invitrogen, Cat# 18032029). The reaction mixture was assembled on ice as following:

20 μ l 5 \times reaction buffer (50mM Tris-HCl pH 7.9, 250mM NaCl, 0.5mg/ml BSA)

40 μ l 25mM MgCl₂

10 μ l 25mM MnCl₂

10 μ l 10mM ATP
18 μ l RNA in H₂O
1 μ l 1 unit/ μ l RNaseOUT
1 μ l 5 unit/ μ l Poly-A polymerase

The mixture was then incubated at 37°C for 30 min and the RNA was extracted by phenol and chloroform and precipitated in EtOH, and finally resuspended in RNA-protein-binding buffer.

The Poly-A tailed RNA was heated to 85°C for 2 min and cooled down gradually to allow folding, and then incubated with 50mM biotinylated oligo-dT[20] primer at 37°C for 10min in streptavidin-coated PCR tubes. The liquid was then aspirated from the tubes, which were washed twice with washing buffer (20mM Tris-HCl pH 7.5, 500mM NaCl, 1mM EDTA, 0.01% Triton X-100).

20 μ l *in vitro* protein expression mixtures (resuspended precipitants of the negative control, GFP and Gp8d1) were added to the streptavidin-coated tubes and incubated for 20 min at 37°C. Then the liquid was taken out and the tubes were washed 3 times with washing buffer.

Finally, 20 μ l 1 \times SDS loading buffer was added to the tubes, which were transferred into 95°C heating block and incubated for 5min. The 20 μ l solutions were loaded onto a 10% SDS-PAGE.

4.5.6 Gel shift assays with radio-isotope labelled RNAs

The *in vitro* transcribed ³²P labelled U1- and U5-snRNA candidates resuspended in protein-RNA binding buffer were heated to 85°C for 2 min, and then cooled down gradually for folding. 10 μ l *in vitro* protein expression mixtures (resuspended precipitants of negative control, GFP and Gp8d1) were added into 5ul U1- and U5-snRNA candidates on ice. The mixtures were incubated at 37°C for 1 h before loading on to 8% native polyacrylamide gel. The gel was then run at 150V for 3 h. After running, the gel was transferred into the dark room, covered by an X-ray film (Kodak) and the gel-cassette was left standing overnight at 4°C. The film was developed the next day.

4.5.7 Reverse transcription (RT) PCR

All the RT-PCR reactions were performed using the Thermoscript cDNA synthesis kit (Invitrogen, Cat# 11146024). Total RNA treated with DNase was

mixed with the corresponding reverse primer and dNTPs. The mixture was heated to 85°C for 2 min and cooled down gradually. Then a mixture of reaction buffer, RNaseOUT and reverse transcription enzyme was added in. All RT reactions were carried out for 1 h at 55°C and heated to 85°C to inactivate the enzyme. 2µl RT reaction was taken out to serve as the template for downstream PCR reaction. Results were analyzed on 2% agarose gels. Primers used for testing expression of the U2, U4 and U6 snRNA candidates are listed below:

U2_cand_1_F	CTATATGATGACTATTAATAGTAAGTTTAAAGA
U2_cand_1_R	GTTGCTTCTAATATATAGTGAGGGA
U2_cand_2_F	ACAGCTGCATTGAACAATAGTTTCT
U2_cand_2_R	CAAGGCGACTATCCTAGTTG
U2_cand_3_F	TCA CCT CAC ATG ATT TGG TGA
U2_cand_3_R	TACATTTCTGCGGGGAGTCT
Likely_U6_F	AGTGTCCGGGAACAAGTGAG
Likely_U6_R	TAGGGTCTGAGTACCACGAC
Likely_U4_F	TATTGCGAGAAAACCCTCTTAG
Likely_U4_R	CCCACAAAATTTCGACACCAC

Chapter Five – Unusual ncRNAs in *Giardia* and the putative RNAi pathway

Abstract:

A number of transcribed dsRNAs in *Giardia* raised my interest to further look into their unusual features. Double-stranded RNAs are known to be involved with gene silencing mechanisms in various eukaryotic organisms. Recent biochemical studies have characterised Dicer: the key protein component of RNAi mechanism from *Giardia*. This finding reinforces the earlier suggestions that *Giardia* uses RNAi to regulate gene expression. However *Giardia* endogenous RNAs which are possibly involved in gene silencing have not yet been identified. In this chapter, several long tandem repeats of dsRNAs have been observed to be highly transcribed, and some of them undergo self-cleavage at the presence of divalent metal ions. The repeating units of these repeats are homologous to part of the large number of VSP (variant surface protein) genes that are expressed on the cell surface. The transcriptional patterns and sequences of these novel dsRNAs are then analysed. They are likely to be candidates of Dicer protein substrates, although further verification is still needed. In addition, my earlier study discovered a truncated transcript of the Dicer mRNA, which led to investigations of the individual RNase III domain of *Giardia* Dicer protein. The overall view of the possible RNA-induced silencing in *Giardia* is also reviewed.

5.1 Introduction: the mechanism of dsRNA-induced gene silencing and RNAi in eukaryotes

Since its discovery in 1998 (Fire et al. 1998), RNA interference (RNAi) has been found in a variety of organisms including animals (Collins and Cheng 2006), plants (Gazzani et al. 2004) and protists (Ullu et al. 2004), and is implicated in a wide range of gene silencing mechanisms including down-regulating mRNA levels (Sen and Roy 2007), heterochromatin assembly and maintenance (Grewal and Elgin 2007), DNA elimination (Collins and Cheng 2006), promoter silencing (Morris et al. 2004), developmental control (Chan et al. 2006), and up-regulation of transcription during the cell cycle (Vasudevan

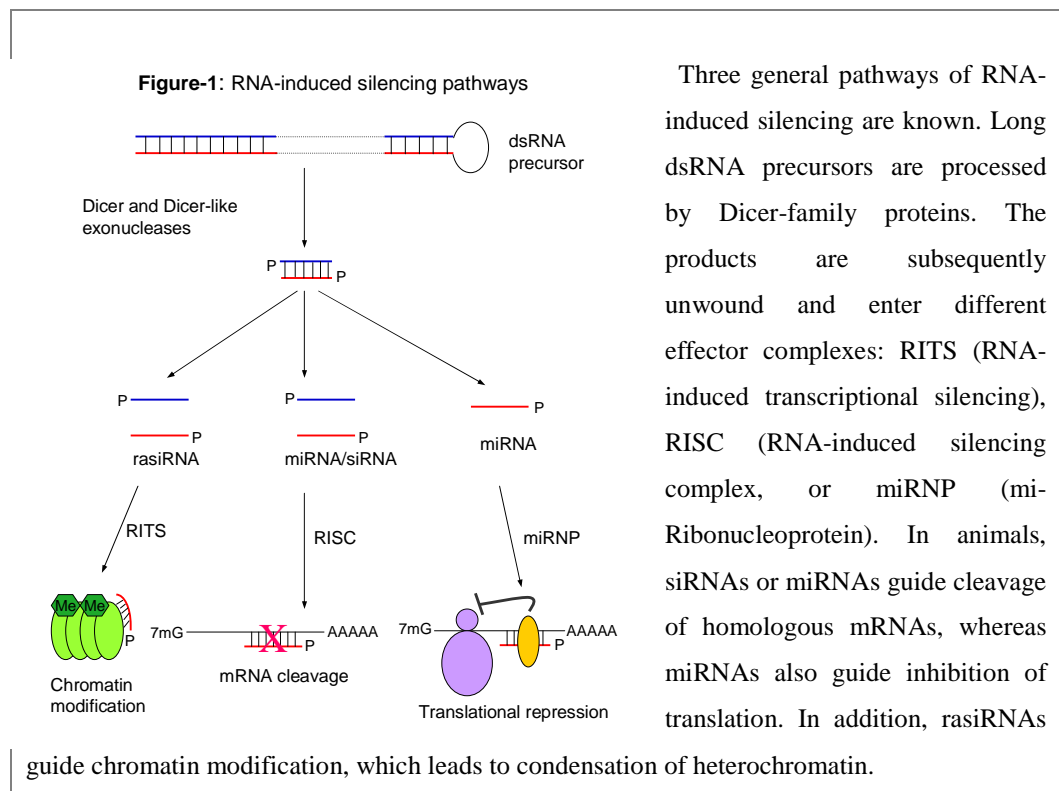
et al. 2007). The key elements that guide all the above processes are small RNAs with size ranges of 20-26nt.

Three major types of small RNAs associated with RNAi have been extensively studied: short interfering RNAs (siRNAs), repeat-associated short interfering RNAs (rasiRNAs) and microRNAs (miRNAs) (Meister and Tuschl 2004). In nature, endogenous dsRNAs are produced by hybridization of complementary RNA transcripts, especially from repetitive sequences such as transposons (Meister and Tuschl 2004). These RNAs are processed to give siRNAs and rasiRNAs and are generally involved in mRNA degradation or chromosomal modifications. There is a possibility that the self-cleaving dsRNA reported in this chapter is a new or modified form of RNAi. miRNAs, which usually function as translational repressors, are produced from transcripts that contain 20- to 50-bp complementary or near-complementary inverted repeats that fold into hairpins. There may well be other forms of RNAi, and recently up-regulation of mRNA expression has been found at stages in the cell cycle (Vasudevan et al. 2007). Recently another type of small RNA named piRNA (Lau et al. 2006) has been found in animals (Aravin et al. 2006; Brennecke et al. 2007; Houwing et al. 2007). piRNAs function in transposon silencing in a similar way to RNAi (Hartig et al. 2007).

RNA interference was first recognized as an anti-viral mechanism to protect organisms against RNA viruses and also to prevent random integration of transposable elements (Waterhouse et al. 2001). Natural siRNAs have predominantly been found in plants and guide cleavage of complementary mRNAs. miRNAs are mainly found in animals and function predominantly to inhibit translation by targeting partially complementary sequences at the 3'-untranslated regions (UTRs) of mRNAs. Finally, artificial long dsRNAs or siRNAs have been used as tools for inactivating target gene expression in both cultured cells and living organisms.

Maturation of small RNAs involves multiple steps catalysed by dsRNA-specific RNase-III-type endonucleases Droscha and Dicer, which generally

contain the catalytic RNase-III domains and dsRNA-binding domains (Bernstein et al. 2001; Lee et al. 2003). Droscha is specifically required for processing miRNA precursors but not long dsRNAs (Lee et al. 2003). The processed or unprocessed precursors are exported from the nucleus to cytoplasm by the nuclear export receptor, exportin-5 (Bohnsack et al. 2004). Once in the cytoplasm, the precursor RNAs are further cleaved by Dicer to give short dsRNAs of 21- to 26-nt with 5' phosphates and 2-nt 3' overhangs (Lee et al. 2003).



Dicer homologues have been found in most eukaryotes including deep-branching unicellular parasites such as *Giardia* and *Trichomonas* (Finn et al. 2006). Some organisms, such as *Drosophila* and *Arabidopsis*, have more than one Dicer paralogue, which process dsRNA precursors of different origins (Lee et al. 2004).

After cleavage by Dicer in the cytoplasm, the short dsRNAs are then incorporated into ribonucleoprotein particles which assemble the RNA-induced silencing complex (RISC) (Hammond et al. 2001b). The components of RISC vary between organisms, and have a molecular mass ranging from 130 - 160

kDa in human (Martinez and Tuschl 2004) and up to 500 kDa in *Drosophila* (Pham et al. 2004). However, every RISC contains a member of the Argonaute (Ago) protein family. The assembly of RISC also requires energy-driven unwinding of the siRNA or microRNA duplexes plus conformational changes of pre-assembled RNPs. Several ATPases have been implicated in RNA silencing mechanisms, and one DEAD-box RNA helicase in *Drosophila*: Armitage has been characterized in detail (Tomari et al. 2004). Naturally occurring small RNAs show a strong bias for only one strand accumulating into the RISC (Schwarz et al. 2003), possibly caused by the rate-limiting unwinding step, which allows the weakly-paired 5'-end of the dsRNA to enter RISC first.

The single-stranded siRNA or miRNA in the RISC is strongly bound to the Ago protein (Martinez and Tuschl 2004). Ago proteins are characterized by two conserved domains: the PAZ domain and the Piwi domain (Carmell et al. 2002). The Piwi domain has been shown to interact with Dicer (Tabbaz et al. 2004), and the crystal structure of an archaeal Ago protein showed that the Piwi domains is strikingly similar to members of the RNase-H family (Song et al. 2004). Since RNase-H cleaves the RNA strand of RNA/DNA hybrids, it has been suggested that the Ago proteins may cleave the target RNA as the siRNAs guide the RISC to the cleavage positions. The PAZ domain was also shown to be involved in protein-protein interaction with Dicer as Ago proteins co-immunoprecipitate with Dicer (Hammond et al. 2001a). More biochemical and structural studies indicated that the PAZ domain is an RNA-binding domain that specifically recognizes the terminus of the short dsRNAs processed by Dicer (Ma et al. 2004). Hence, the PAZ domains of many Dicer proteins have been suggested as a docking place for long dsRNAs. However the way in which single-stranded siRNA or miRNA binds to the Ago protein after unwinding is not fully understood.

In some organisms such as *Neurospora crassa* (Forrest et al. 2004), *C. elegans* (Smardon et al. 2000), *S. pombe* (Martienssen et al. 2005) and plants (Gazzani et al. 2004), an RNA-dependent-RNA-polymerase (RdRp) is also

essential for dsRNA-triggered gene silencing. The RdRp is likely to use the siRNA as primers and convert the target RNAs into dsRNAs and a second wave of gene silencing is initiated. However, RdRp is not found in insects and mammals.

In 1998, it became clear that the unicellular protist *Trypanosoma brucei* have the machinery to degrade mRNAs upon exposure to homologous dsRNAs (Ngo et al. 1998). RNAi has been extensively used to down-regulate gene expression in *T. brucei* (Tschudi et al. 2003). The mechanism of RNAi in *T. brucei* is essentially the same as that of other eukaryotes. Dicer activity was detected in cell-free extracts of *T. brucei* (Ullu et al. 2004), and later an unusual Dicer-like protein with distinct RNase-III domain arrangement was identified (Shi et al. 2006). *T. brucei* genome also contains one protein homologue (TbAgo1) of the Ago gene family (Finn et al. 2006). Biochemical studies showed that the TbAgo1 was a cytoplasmic protein and it bound directly to siRNAs (Shi et al. 2004). The TbAgo1-siRNA complexes have been found to associate with translating ribosomes (Djikeng et al. 2003), and it was proposed that the association between the TbAgo1-siRNA complexes and polyribosomes could facilitate recognition of target mRNA by RISC (Djikeng et al. 2003). An alternative pathway suggested that the TbAgo1-siRNA complexes might also directly associate with ribosome-free mRNAs and the cleavage reaction was not dependent on the interaction between translation and RNAi machineries (Ullu et al. 2004).

The evidence above indicated that siRNAs cloned from *T. brucei* contained a high proportion of sequences derived from retro-transposons, suggesting that the RNAi mechanism in *T. brucei* acts as a genome-wide defence to silence retro-transposons (Djikeng et al. 2001). Inhibiting TbAgo1 led to complete disappearance of retro-transposon-derived siRNAs and increase in transposon levels (Shi et al. 2004). Therefore, it was suggested that the RNAi machinery in *T. brucei* might function in chromatin remodelling* (Ullu et al. 2004) because

* Chromatin remodelling: dynamic structural changes to the chromatin occurring throughout the cell division cycle, so that certain regions of the chromatin can be loosened and exposed for active transcription and others condensed.

retro-transposons are usually found in heterochromatic* regions. This is also the case in *S. pombe* (Volpe et al. 2002). Furthermore, the existence of an RNAi mechanism in *T. brucei* has also been suggested by the finding of dsRNA homologous to snoRNAs, and that these dsRNAs could induce specific silencing of the corresponding snoRNAs (Liang et al. 2003). The studies of RNAi in *T. brucei* led to more investigation of possible RNAi mechanisms in other unicellular parasites, representing deep-branching groups of eukaryotes.

Several protozoan parasites have been subjected to extensive study in searching for evidence of RNAi in these organisms. In case of the *Trypanosomatid* family, RNAi activity was found in *T. congolense*, but not in *T. cruzi* and *L. major* (Ullu et al. 2004). However, database searching has revealed a protein with a solo Piwi domain from *T. brucei*, *L. major*, *T. vivax* and *T. cruzi* (Ullu et al. 2004). Because the Piwi-domain containing proteins are present even in organisms that may lack RNAi, and also in certain prokaryotes (Cerutti et al. 2000), their functions may not be related strictly to gene silencing and still remains unknown. In species of *Plasmodium*, the presence of RNAi is uncertain. Database mining (Finn et al. 2006) for proteins with domains homologous to Dicer, Paz, Piwi and RdRp did not identify candidates in any of the *Plasmodium* species, despite evidence showed the accumulation of siRNA-like molecules in *P. falciparum* cells treated with dsRNAs (Malhotra et al. 2002). However, the possibility of the existence of a non-classical RNAi pathway in *Plasmodium* is not ruled out.

The presence of RNAi has been apparent in the deep-branching eukaryote *Giardia*. Detailed biochemical and structural studies have been carried out for the *Giardia* Dicer protein homologue, showing that recombinant *Giardia* Dicer could cleave dsRNA into 25nt short fragments *in vitro* (Macrae et al. 2006). The latest *Giardia* genome (Morrison et al. 2007) contains protein homologues for Argonaute and RdRp. In addition, earlier studies have shown the presence of 20-30nt long RNAs derived from sense and antisense sequences of the abundant retrotransposon elements in *Giardia* (Ullu et al. 2005), and there is

* Heterochromatic region: Regions of chromosome that are tightly coiled throughout cell cycle, and for the most part, genetically inactive.

unpublished indication that RNAi might be involved in controlling expression of the variant-specific surface proteins (VSPs), and also that the function of RdRp was important (Ullu et al. 2004). The transcriptome of *Giardia* contains numerous sterile antisense transcripts as shown by random cDNA sampling (Elmendorf et al. 2001). It is not yet known whether antisense and retrotransposon transcription is an integrated component of the potential RNAi mechanism in *Giardia*, but the presence of the unusual RNA transcripts reported above strongly suggests special molecular machinery of early-branching eukaryotes.

Following the analysis of the cDNA library discussed in Chapter 3, several unusual ncRNAs with potential functions in RNA-induced silencing were discovered in this study and named Girep RNAs (abbreviation of *Giardia* repetitive RNAs). These sequences consist of seven to eleven direct tandem repeats and are transcribed at both sense- and antisense- directions, therefore a fraction of these transcripts are likely to form long dsRNAs *in vivo*. In addition, regions within these transcripts are homologous to a number of VSP genes which are believed to be regulated by a putative RNAi mechanism in *Giardia* (Ullu et al. 2004). Sequence and structural comparison shows highly conserved regions within these transcripts, however no sequence homology to ncRNAs from other organisms has been observed. Four out of the five Girep RNAs undergo clear self-cleavage at the presence of Mg^{2+} , and this unusual feature is currently not fully understood. To get an overview of the putative RNAi pathway in *Giardia*, protein components of RNAi were studied by comparing the putative *Giardia* proteins with homologous proteins from eukaryotic and prokaryotic organisms. From analyses it appears that the reduced number of functional motifs on a single protein is a common feature of proteins from single-cellular eukaryotes. *Giardia* and other deep-branching eukaryotes exhibit strong similarity with archaea at the protein level, and it is expected that complex RNA-processing pathways such as RNAi in deep-branching eukaryotes involve more dynamic protein-protein interactions. Finally, a truncated Dicer transcript found at early stages of this study shows a number of unusual features, and a small follow-up study was carried out.

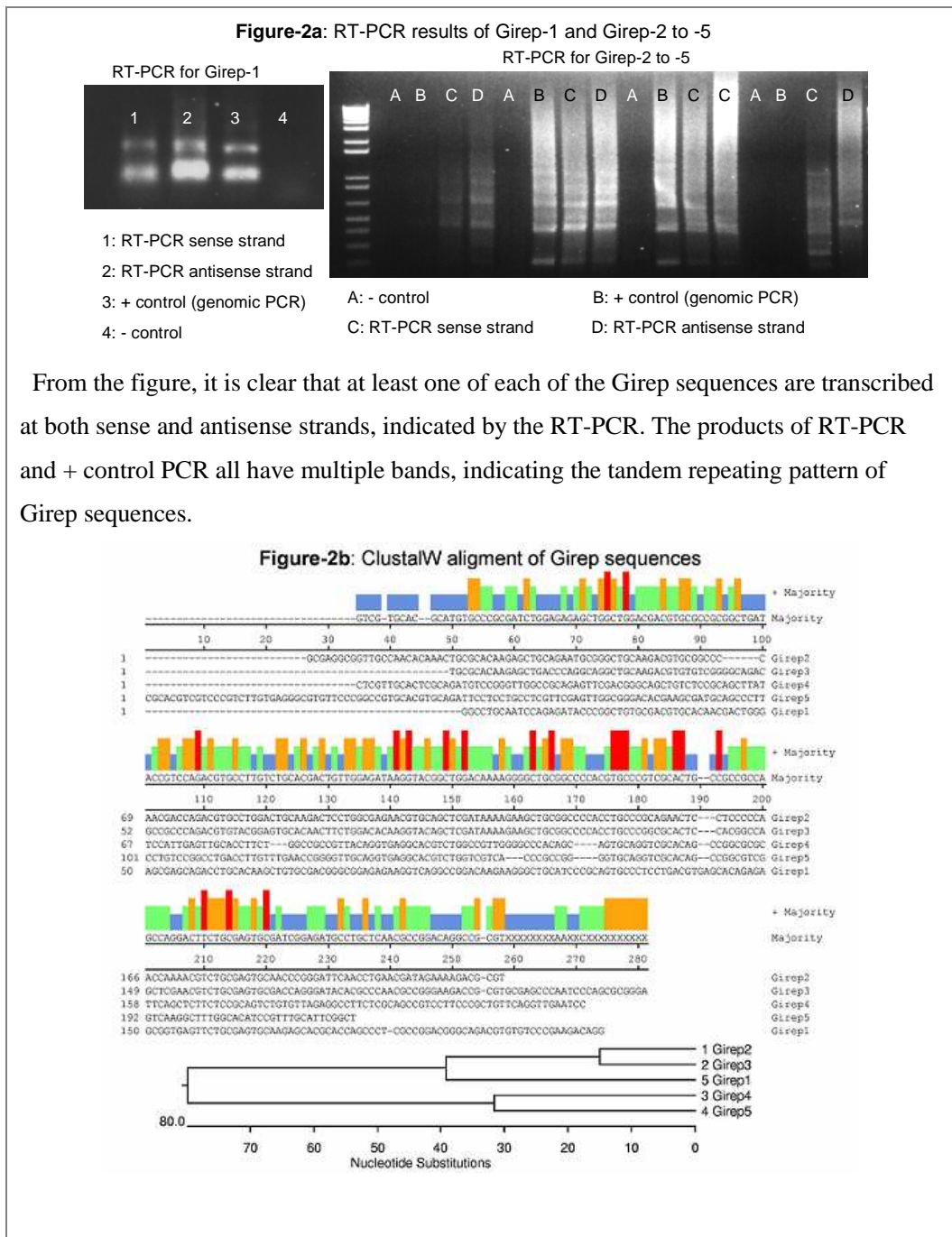
5.2 The unusual ncRNA repeats in *Giardia*

The previous study of ncRNAs from *Giardia* (Chapter-2) has revealed a number of unusual ncRNAs. A fragment of the variant-surface-protein (VSP) seen in the *Giardia* ncRNA library (Chen et al. 2007) raised the question of this fragment being the product of the putative RNAi gene silencing mechanism in *Giardia*. VSP gene expression is crucial for the surface antigenic variation of *Giardia* trophozoites (Nash et al. 1988). The sequences and structures of VSP proteins are highly similar, however in a single trophozoite only one VSP is expressed out of a total of 150 to 200 VSP genes (Nash et al. 2001). The mechanism underlining VSP switching is unknown although both RNAi (Ullu et al. 2004) and epigenetic mechanism (Kulakova et al. 2006) involving histone acetylation/deacetylation have been suggested. A number of studies have suggested the potential presence of RNAi pathway in *Giardia* in various aspects, which include the earlier prediction of RNAi in *Giardia* (Ullu et al. 2004), the expression study of the Dicer protein (Macrae et al. 2006), and the study of sense and antisense small RNAs derived from telomeric repeats (Ullu et al. 2005).

Blasting the fragment of the VSP fragment from the ncRNA library against *Giardia* genome has revealed one unusual long tandem repeated sequence (Girep-1) (Chen et al. 2007) in the *Giardia* genome. Re-blasting this Girep-1 sequence in the *Giardia* genome then identified a group of similar sequences (Girep-1 to Girep-5). This group of Girep sequences are all direct-repeat sequences located at different positions of the genome. In addition to the long tandem repeats, there are also a number of shorter homologous sequences located at non-coding regions of the genome. RT-PCR showed that these repeat sequences were all expressed on both sense- and antisense- strands (Figure-2a). With exception of the antisense strand of Girep-1 being a hypothetical mRNA transcript (GL50803_227577), all the other Girep sequences are non-coding. Genomic information of the five long-tandem repeated sequences studied here are listed in Table-1.

Table-1: Expressed direct-repeat sequences in *Giardia*

Name	Number of repeating units	Length of repeating unit	Location
Girep-1	9	222	Contig2: 327758-328858
Girep-2	9	222	Contig2: 392343-393229
Girep-3	7	228	Contig54: 1296-2416
Girep-4	8	228	Contig111: 1-1810
Girep-5	11	225	Conting98: 1644- Contig50: 735



Sequence alignment of all Girep candidates (Figure-2b) revealed considerable homology among the five sets of sequences, and also shared motifs between sequence pairs, with Girep-2 and Girep-3 being a closely related pair and Girep-4 and Girep-5 being another. The shared sequence motifs and tandem-repeated pattern suggest that these ncRNAs belong to one group. All five Girep sequences show close relationship with a number of VSP genes. Genomic BLAST results indicate that each of the Girep sequences contains sequence fragments that correspond to several VSP genes. The patterns of sequence match are variable, but all involve the repeating units of Girep sequences being partially homologous to repeating units of VSP genes (Figure-3). The details of matching between Girep sequences and mRNAs are shown in Table-2.

It appears that each of the Girep sequences has more than one match to different VSP genes or to other open-reading frames that are not yet characterised. This is likely due to the high degree of sequence similarity among members of the VSP gene family. The matching pattern between the Girep RNA sequences and VSP genes suggest that the function of these unusual RNAs may relate to regulation of VSP gene expression.

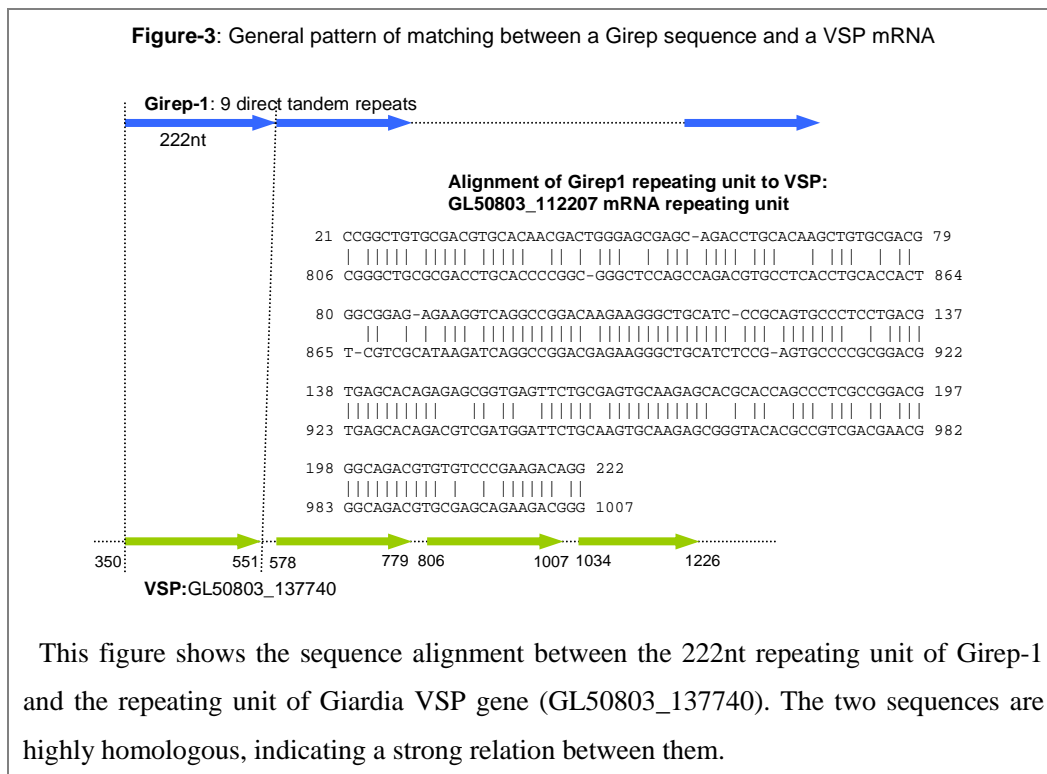


Table-2: Matching of Girep sequences to mRNAs:

Each Girep sequence is homologous to more than one VSP gene. The homologous VSP genes for each Girep sequences are listed in order so that the degree of homology decreases from the top one to the last one in each cell. Overlapping of matches is common.

Girep sequences	VSP or hypothetical genes	Sense/Antisense	Number of repeated matching units
Girep-1	VSP: GL50803_137740	Sense	4
	VSP: GL50803_112207	Sense	2
	VSP: GL50803_137717	Sense	4
	VSP: GL50803_112693	Sense	3
	VSP: GL50803_8595	Sense	1
	Hypothetical: GL50803_87110	Antisense	1
	Hypothetical: GL50803_106057	Sense	1
Girep-2	VSP: GL50803_26894	Sense	1
	VSP with INR: GL50803_101010	Sense	1
	VSP: GL50803_8595	Antisense	1
	Hypothetical: GL53803_87110	Sense	1
	VSP: GL50803_101496	Sense	1
	VSP: GL50803_102178	Sense	4
	VSP: GL50803_137740	Sense	1
	Hypothetical: GL50803_99660	Sense	1
	Hypothetical: GL50803_38998	Sense	1
	VSP: GL50803_137681	Sense	1
	VSP: GL50803_137721	Sense	1
Girep-3	VSP: GL50803_26894	Sense	1
	VSP: GL50803_8595	Sense	1
	Hypothetical: GL50803_87110	Antisense	1
	VSP: GL50803_101496	Sense	1
	VSP with INR: GL50803_101010	Sense	1
	VSP: GL50803_112693	Sense	3
	Hypothetical: GL50803_13197	Antisense	1
	VSP: GL50803_137717	Sense	4
	VSP: GL50803_112647	Sense	1
	Girep-4	VSP: GL50803_101010	Antisense
VSP: GL50803_26894		Antisense	1
VSP: GL50803_102178		Antisense	1
VSP: GL50803_137740		Antisense	4
VSP: GL50803_137717		Antisense	4
Hypothetical: GL50803_227577 (Girep-1 antisense transcript)		Sense	1
VSP: GL50803_101496		Antisense	1
Girep-5	VSP: GL50803_137717	Antisense	4
	VSP: GL50803_137740	Antisense	4
	VSP: GL50803_112207	Antisense	2
	Hypothetical: GL50803_227577 (Girep-1 antisense transcript)	Sense	9
	VSP: GL50803_101010	Antisense	1
	VSP: GL50803_102178	Antisense	1
	VSP: GL50803_101496	Antisense	1
	VSP: GL50803_26894	Antisense	1

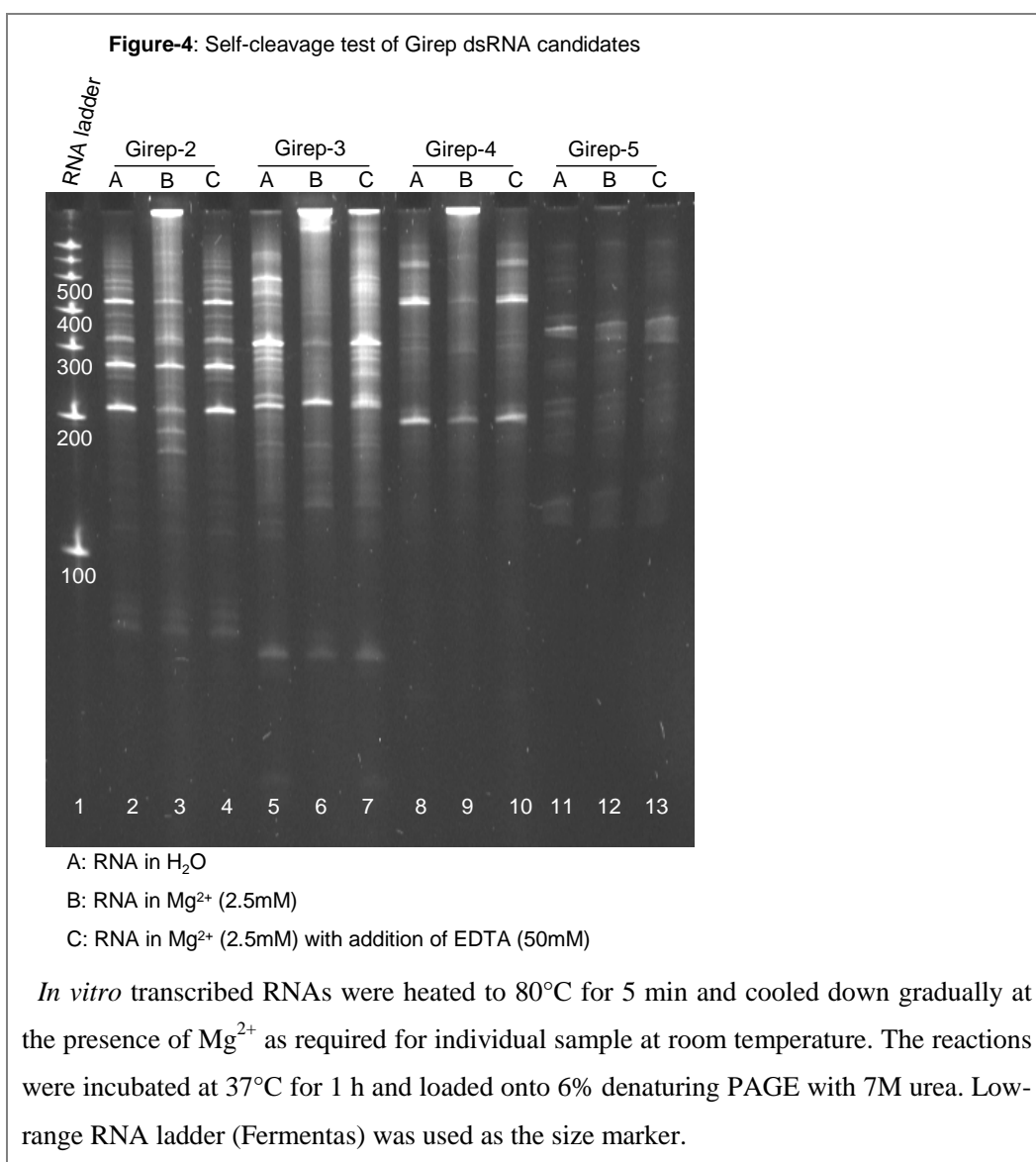
As shown in Figure-3, the general matching pattern indicates a highly homologous, however not completely complementary, alignment of the Girep repeating units and the repeating units of a VSP gene. This is an updated result from previous study (Chen et al. 2007). Both sense and antisense transcripts of Girep sequences have matching mRNA partners, supporting the observed sense- and antisense- transcription of Girep sequences. In all there are 18 mRNAs (Table-2) that have regions homologous to Girep sequences.

Searching the *Giardia* genome revealed additional sequences (not repeats) that are homologous to the Girep sequences. It is highly likely that these shorter sequences can match to additional VSP genes. Comparing Girep sequences with the latest *Giardia* EST database (Morrison et al. 2007) has revealed a large number of homologous hits. This observation suggests that there may be a large portion of the total VSP genes covered by expressed homologous non-coding sequences.

In order to look for potential promoter sequences that may reveal information about the expression of these unusual ncRNAs, the upstream sequences at both sense- and antisense- directions were extracted for all five Girep sequences. Compared with the standard *Giardia* Pol II promoter consensus sequences of cytoskeleton genes (Holberton and Marshall 1995; McArthur et al. 2000), the upstream sequences from both sense- and antisense- directions do not show either a conserved A-rich motif or a likely “TATA” box. Motif analysis does not indicate any consensus regions either. By comparisons with the other ncRNAs found in *Giardia*, this lacking of conserved upstream promoter sequences is not unusual. Results in Chapter 3 show that the potential promoter regions of ncRNAs in *Giardia* are highly variable and suggests the possibility of many non-coding transcripts being generated by a loosely controlled expression cascade.

Based on the observation of self-cleaving feature of two dsRNA candidates Genie-1 and Girep-1 (Chen et al. 2007), the additional four Girep dsRNAs (Girep-2 to Girep-5) have been tested for potential ability to self-cleave at the presence of Mg^{2+} . Results (Figure-4) show partial self-cleaving activity in three

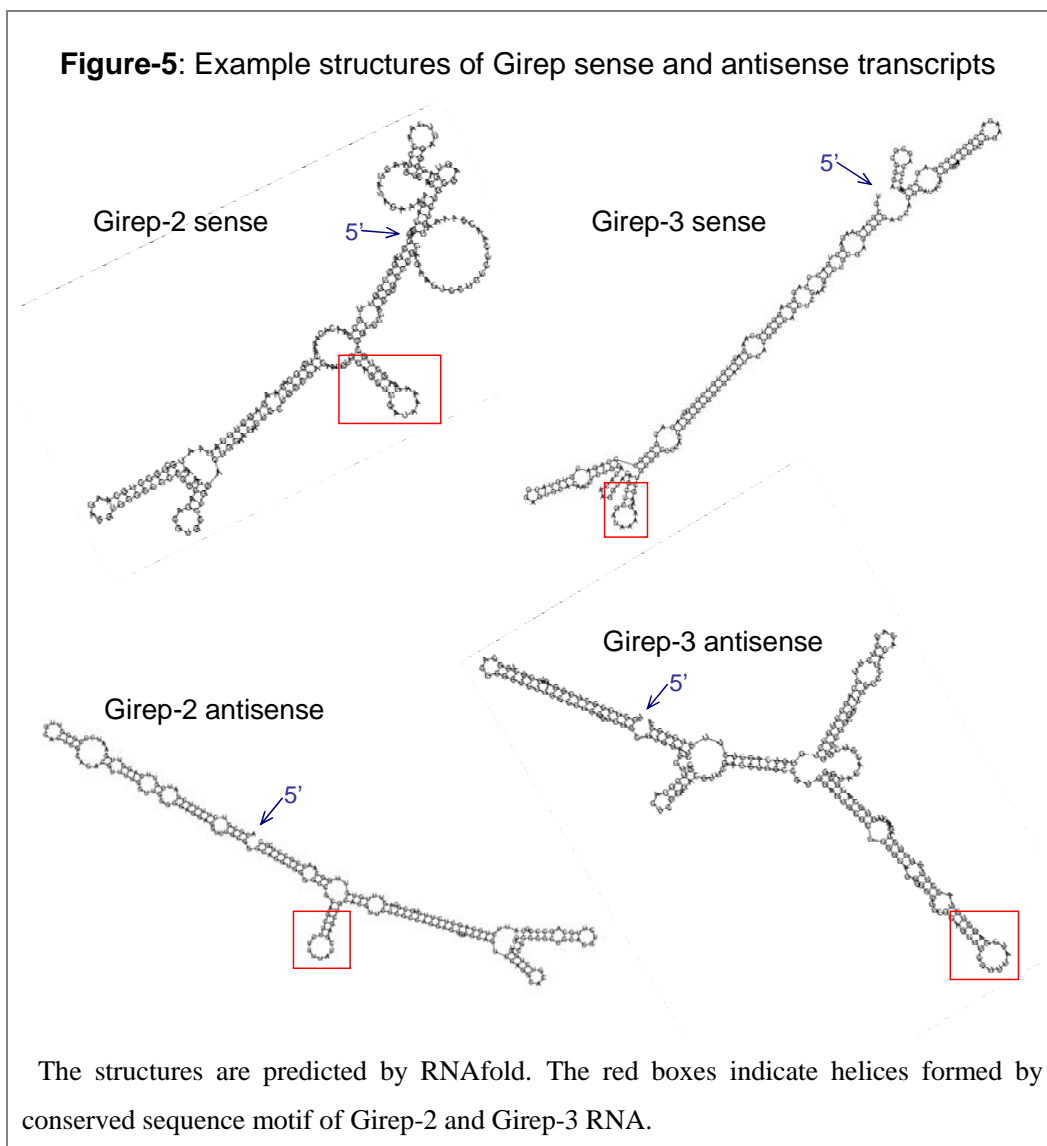
of the four candidates tested. Comparing with RNA incubated in distilled H₂O, addition of Mg²⁺ has caused partial self-cleavage of Girep-2, Girep-3 and Girep-4 as shown in lane 3, 6, and 9. Addition of EDTA as a metal chelator before addition of Mg²⁺ prevented self-cleavage. Therefore the self-cleavage observed is the result of Mg²⁺ addition. It is assumed that Mg²⁺ causes change of RNA-folding thus forming the structures facilitating self-cleavage. Girep-5 candidate did not show apparent self-cleavage, suggesting that self-cleavage may not be a conserved feature of all tandem RNA repeats in *Giardia*. However, this might also due to other variable factors in sample preparation.



It has been shown that the Girep-1 RNA only self-cleaves when both sense- and antisense- strands are present at the same time (Chen et al. 2007). This should hold true for all Girep RNAs because they belong to the same class. However, *in vitro* assays cannot always represent the true situations *in vivo*. The role of dsRNAs has only been found to provide substrates for Dicer protein and subsequently acting in RNAi gene silencing mechanisms to suppress homologous gene expression. It is possible that Girep sense- and antisense- transcripts bind to form long dsRNAs, which are processed by *Giardia* Dicer protein. If this is the case, the amount of sense- and antisense- transcripts should be roughly level. However in Figure-2a, it can be visualized that the amounts of sense- and antisense- RT-PCR products derived from same amount of RNA are obviously not equal at least for Girep-1 and Girep-2. This observation shows that although some of sense- and antisense- transcripts may interact to form dsRNAs, whereas the excess ones may function as single stranded RNAs, which may regulate expression of various VSP genes through a yet uncharacterised antisense mechanism.

Structures generated using RNAfold (Hofacker 2003) of the sense- and antisense- RNAs of Girep-1 to Girep-5 can fold into extensive helices (Appendix-4). The current RNA-folding software algorithms only give putative RNA structures with minimum free energy, and may not represent the true *in vivo* folding of RNA. However, computational prediction of RNA structures is an efficient way for comparing structural similarity among different RNAs. Figure-5 shows the folding of Girep-2 and Girep-3, which are the most closely related pair from the alignment of all Girep sequences (Figure-2a).

As shown in the following figure, the putative structures of the complementary RNA transcripts are different. Therefore it is likely that the sense and antisense transcripts do not completely complement each other to form DNA-like dsRNA or symmetrical structures *in vivo*, instead *in vivo* folding of these RNAs can be variable with some sense and antisense segments binding together and other regions may remain as helices. In addition, the shared sequence motifs among Girep RNAs are likely to fold in to the same helical structure as indicated by red boxes in Figure-5.



Sequences of the Girep RNAs identified here do not show sequence similarity with other known eukaryotic ncRNAs from BLAST (Ye et al. 2006) search against either the current NCBI (<http://www.ncbi.nlm.nih.gov>) or Rfam (Griffiths-Jones et al. 2005) databases. However there is one report from a study of RNAs from *Leishmania infantum* (Dumas et al. 2006), where a class of ncRNAs ranging from 300 to 600 nucleotides were identified, that were expressed as tandem head-to-tail repeats, and were involved in developmental regulation. These ncRNAs from *L. infantum* are transcribed at both sense- and antisense- orientations and are encoded as clusters of 270bp repeats (Dumas et al. 2006). The same study also showed that similar repeated sequences existed in different *Leishmania* strains at relatively similar chromosomal locations; however no expression was detected for *L. major*. So far the *Leishmania* study

is the only report on expression of repetitive ncRNAs, apart from this present study on *Giardia*.

Repetitive ncRNAs have not been widely found in eukaryotes. However the current knowledge on eukaryotic ncRNAs does not exclude the possibility that similar RNAs exist in other eukaryotes. The function of *Giardia* Girep RNAs is yet unknown, but the homology between Girep RNAs and VSP genes gives a hint that these RNAs are likely to be involved in antisense or RNAi regulation of VSP gene expression, which is probably important in the development of *Giardia*.

The following section discusses the putative RNAi pathway in *Giardia*. Functional characterization of *Giardia* Dicer protein (Macrae et al. 2006) has led to further investigation of the RNAi pathway. So far the native RNA substrates for Dicer have not been identified in *Giardia*, but it is highly likely that some of the currently known putative dsRNAs including Genie RNAs (Ullu et al. 2005) and Girep RNAs discussed above are processed by *Giardia* Dicer protein. In addition, *Giardia* also has most of the RNAi-associated protein components which are highly conserved in eukaryotes (see next section).

5.3 Protein components of the putative RNAi pathway in *Giardia*

Apart from the well characterized *Giardia* Dicer protein, other protein components of the potential RNAi mechanism in *Giardia* have not been studied. A large number of proteins, which act in different levels of small-RNA induced gene-silencing have been identified from animal, insects and plants, as shown in Table-3. Unicellular eukaryotes tend to have a smaller number of protein homologues. For example, *T. brucei* only possesses one Ago protein whereas human has 4 homologues, each of which is likely to have slightly different functions.

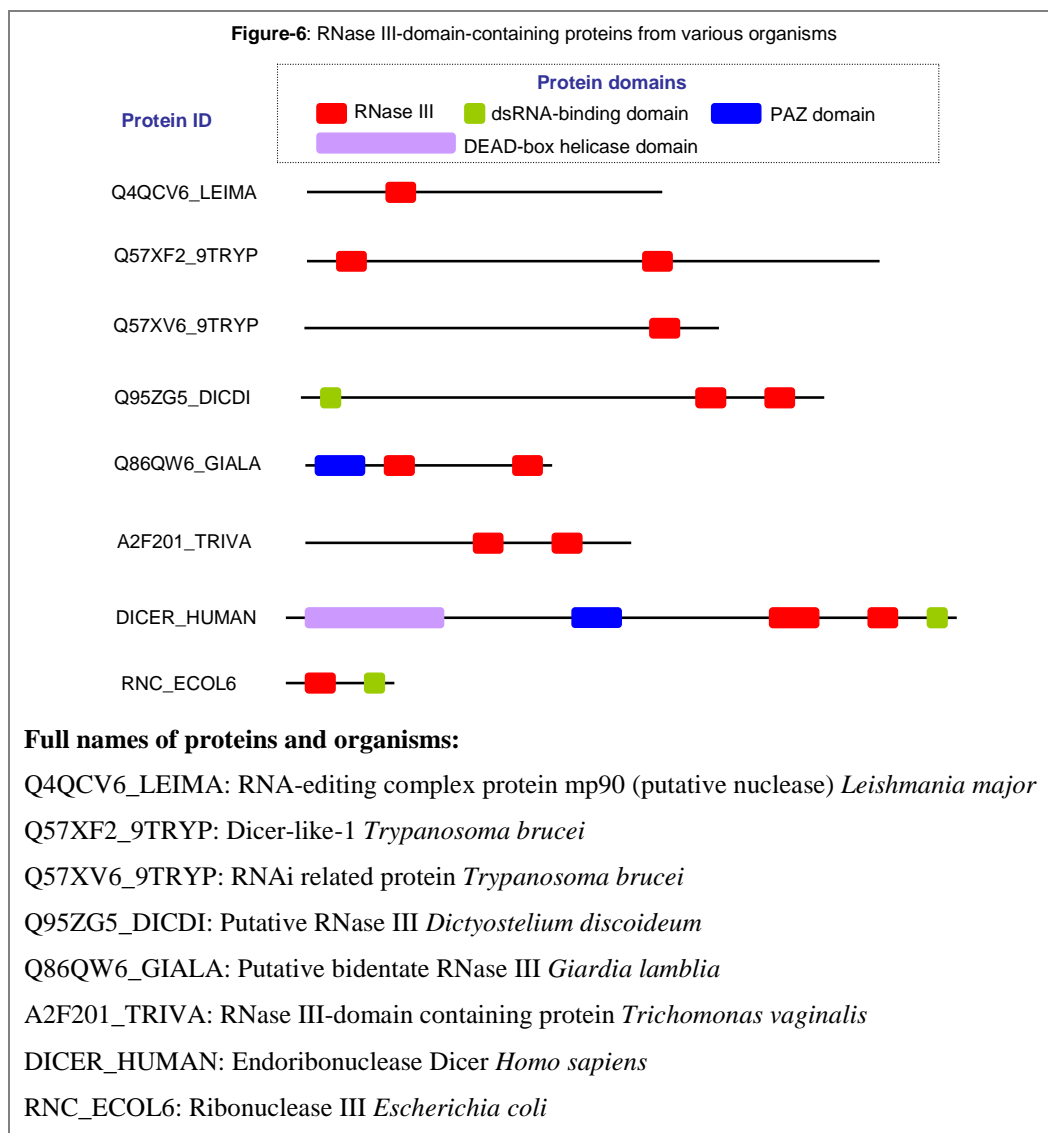
My search of the protein-coding genes of *Giardia* by protein-domain homology revealed several proteins containing the characteristic domains described above. It is clear that *Giardia* contains most of the key protein components required for small-RNA induced gene silencing. However it is not certain that whether the Piwi-domain containing protein has the equivalent function to an Ago-family protein. Database mining has revealed a unique class of proteins which, unlike Ago-family proteins, only contain the Piwi domain without the PAZ domain. The existence of these Piwi-domain containing proteins as a distinct class suggested that these Piwi proteins have separate functions to the Ago-family proteins. A number of studies showed that Piwi proteins bind to a distinct group of small RNAs termed piRNAs (Aravin et al. 2006; Lau et al. 2006; Brennecke et al. 2007; Houwing et al. 2007). Recent studies suggested that Piwi proteins and piRNAs have roles in germ-line maintenance and silencing of transposons (Brennecke et al. 2007; Houwing et al. 2007; Seto et al. 2007). Therefore it is likely that the putative Piwi protein in *Giardia* differs in function to the classical RNAi mechanism.

Table-3: Conserved key proteins in small-RNA induced gene silencing.

Protein	Characteristic domains	Functions	<i>Giardia</i> homologue
Dicer and Dicer-like proteins	PAZ, RNase III	Long dsRNA processing	Dicer (Macrae et al. 2006) (GL50803_103887)
Argonaute and Ago-family proteins	PAZ, Piwi	Short RNA binding	None (However a protein with a solo Piwi domain is present. GL50803_2902 putative Piwi protein)
Putative RNA helicases	DEAD-box	RISC assembly	32 proteins have the putative DEAD domain
RNA-dependent RNA polymerases (RdRp)	RNA polymerase	RNA amplification	Putative RdRp (GL50803_102515)
Other factors (variable across different organisms)	Other domains		
	dsRNA-binding	RISC assembly and initiation of RNAi	weak hits generated from HMMsearch, may be absent
	Exonuclease	siRNA degradation	positive
	DNA helicase	Unwinding DNA	positive
	Chromo	Heterochromatin association	weak hits generated from HMMsearch, may be absent

It is certain that RdRp is present in *Giardia*, which is one of a few unicellular parasites with homologues of RdRp found. RdRp is not a universal component of the RNAi pathway, and is absent in mammals. The other factors in Table-3 (last of the first column) are mostly organism-specific and are accessory factors for specific silencing mechanisms.

Unlike higher eukaryotes, the components of classical small-RNA induced silencing are reduced to various degrees in unicellular and deep-branching eukaryotes. Comparison of key proteins of RNAi from a number of unicellular eukaryotes does not show a universal pattern. Proteins from several deep-branching eukaryotes with RNase III domains, PAZ and Piwi domains were obtained from Pfam (Finn et al. 2006). They were then compared with models of animals, e.g. human as an example, as well as some bacteria and archaea (Figure-6 and 7). It is interesting to notice that although there has been no evidence of RNAi mechanism in *L. major* (Ullu et al. 2004), a protein with single RNase III domain is present, as well as a Piwi-domain containing protein. In most Dicer homologues, the two RNase III domains are arranged close to each other. In eukaryotes, Dicer functions as an intramolecular dimer of the two RNase III domains, assisted by the PAZ domain and dsRNA-binding domains (Zhang et al. 2004; Macrae et al. 2006). An unusual arrangement of RNase III domains is seen in *T. brucei* (Shi et al. 2006), where one RNase III domain is located near the N-terminus and another is located at middle of the protein. It appears that the PAZ domain and dsRNA-binding domain are not absolutely required for RNase III activity in single-celled eukaryotes, suggested by the absence of PAZ domain in *Trypanosome*, *Dictyostelium* and *Trichomonas*. Similarly dsRNA-binding domain is also absent in the Dicer-like proteins of most protists.



Mutation studies of human Dicer protein showed that deletion of the dsRNA-binding domain made Dicer's interaction with the substrate more dependent on the specific structural features of the binding interface, also that mutations in the PAZ domain strongly inhibited Dicer activity (Zhang et al. 2004). Compared with human Dicer, structural studies of *Giardia* Dicer indicated similar PAZ domain structure and RNase III-domain arrangement in the active centre (Macrae et al. 2006). The overall conservation of domain structures and also the fact that *Giardia* Dicer can substitute for *S. pombe* Dicer *in vivo* (Macrae et al. 2006) suggest that Dicer catalysed dsRNA cleavage is a conserved mechanism in eukaryotes.

Recent studies of *Giardia* Dicer with site-directed mutagenesis* revealed that the PAZ domain could be replaced with other RNA recognition domains, which could direct Dicer protein to specific substrates (Macrae et al. 2007). This finding provides a possible explanation to the absence of PAZ domains in some Dicer proteins, including the putative Dicer proteins from several protists shown in Figure-6, as well as Dicer from *S. pombe* and *T. thermophila*. It is likely that either these proteins may contain yet unrecognised RNA-binding motifs, or other proteins with PAZ domains may interact with the RNase III-domain containing proteins and direct cleavage of target RNAs. The latter assumption is supported by the fact that the nuclear RNase III enzyme Drosha interacts with DGCR8-RNA-binding protein, which provides direct and specific recognition of miRNA precursors (Han et al. 2006).

As shown in Figure-6, the RNase III domain has a wide distribution across the three kingdoms of life. The Dicer family of exonucleases generally contains two RNase III domains, which process the dsRNA substrate to give the siRNA 5'- and 3'- ends, respectively (Zhang et al. 2004). The presence of RNA-binding motifs such as the PAZ and dsRNA-binding domain provides specificity of the enzyme. The structure of *Giardia* Dicer protein indicates that Dicer acts as a ruler which measures from the 3'-end of the dsRNA substrate, and the length of the siRNA produced is determined by the distance between PAZ and RNase III domains (Macrae et al. 2006). However, some possibly Dicer-related proteins in unicellular eukaryotes only contain one RNase III domain, as seen in Pfam (Finn et al. 2006), such as the *L. major* RNA-editing protein and the *T. brucei* RNAi-related protein (Figure-6). Proteins with a single RNase III domain are usually seen in fungi, bacteria and archaea. These proteins function as homodimers (Nagel and Ares 2000), and form single processing centres where each RNase III domain cleaves one strand of the dsRNA substrate (Zhang et al. 2004). In *T. brucei*, the single RNase III-domain-containing, RNAi related protein (Q57XV6_9TRYP, Figure-6) is enriched in nuclei whereas the Dicer protein (Q57XF2_9TRYP, Figure-6) is predominantly cytoplasmic (Shi et al. 2006). The functions of single-RNase

* Site-directed mutagenesis: This technique creates mutations at defined sites on DNA molecules, usually plasmids.

III-domain proteins in eukaryotes are yet unknown, however it is likely that variants of RNase III-domain-containing proteins are involved in cleaving different dsRNA substrates. Also, the reduced Dicer and putative RNAi-associated proteins in unicellular eukaryotes may enable flexible association with different RNA-binding domains to suit specific functions.

Unlike the RNase III exonucleases, the Ago protein family, characterised by the RNA-binding PAZ domain and Piwi domain, is only found in eukaryotes, suggesting that the Ago family of proteins is specifically associated with RNAi mechanism. However while the PAZ domain is eukaryote-specific, the Piwi domain is found in all three kingdoms of life, suggesting an early origin of Piwi-family proteins before the divergence of modern life. As shown in Figure-7, Piwi proteins are found in most of the organisms but Ago proteins are only found in RNAi positive eukaryotes. The RNAi-negative *L. major* does not have an Ago protein homologue as expected. However, *Giardia* does not appear to have an Ago protein homologue despite strong evidence of RNAi. Instead, there is a putative Piwi protein, which is the only Piwi-domain-containing protein found in the genome. Aligning the putative Piwi domains from the Piwi family proteins listed in Figure-7, including putative Piwi proteins from eukaryotes, bacteria and archaea show low overall degree of sequence conservation, but highly conserved secondary structure. It is shown in Figure-8 that the arrangement of helices and beta-sheets within Piwi domains from various Piwi proteins are highly similar despite low conservation in primary sequence.

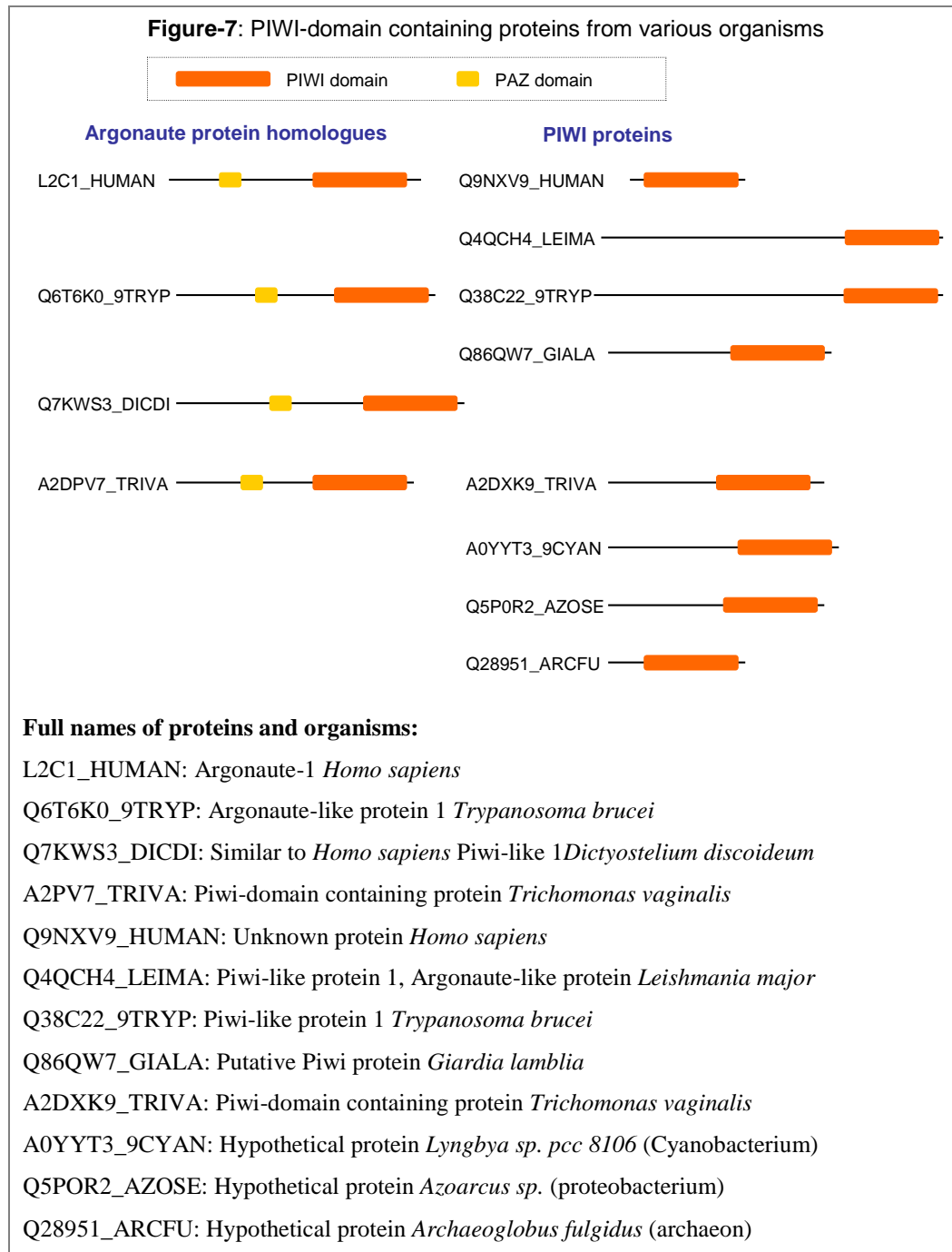
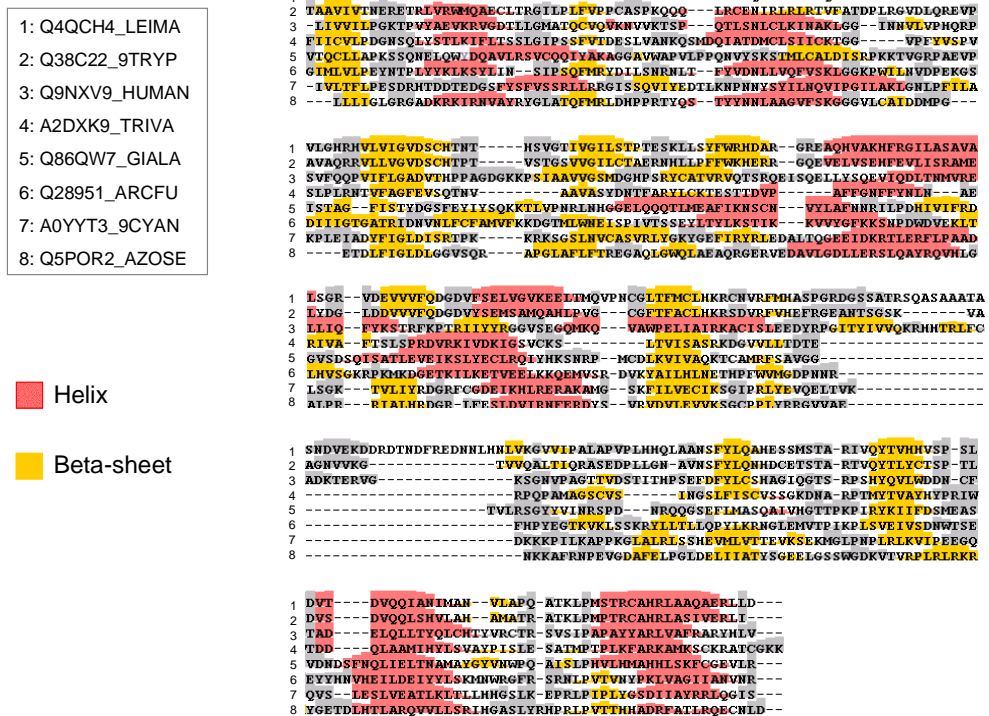


Figure-8: Secondary structural alignment of Piwi domains from various Piwi family proteins

The crystal structure of an archaeal Piwi protein AfPiwi (Q28951_ARCFU *Archaeoglobus fulgidus* in Figure-7) has been determined (Parker et al. 2004); the overall Piwi fold represented novel protein architecture with the individual domain A and B displayed structural similarities to the *lac* repressor and RNase HIII-type fold respectively. The same study (Parker et al. 2004) showed that the protein AfPiwi bound to siRNA *in vitro*. Therefore Piwi family proteins may adopt a conserved model of siRNA binding and mRNA cleavage through the RNase HIII-type domain. It has also been shown that an archaeal Ago protein from *Pyrococcus furiosus* has an RNA slicer activity associated with the Piwi domain (Song et al. 2004). However in Ago proteins, RNA-binding is a feature of the PAZ domain (Ma et al. 2004) which is absent in Piwi family proteins. Comparison of the *Archaeoglobus* Piwi protein and the *Pyrococcus* Ago protein with some eukaryotic Ago proteins also revealed that a region which constituted the docking site for Dicer (Tahbaz et al. 2004) was absent in the both proteins (Parker et al. 2004). This may be a common feature of archaeal Ago/Piwi proteins. The mechanism of RNA-induced silencing in archaea is still unclear. Only three proteins from the family of methanomicrobia contain a single RNase III domain as recorded in Pfam database (Finn et al. 2006) and

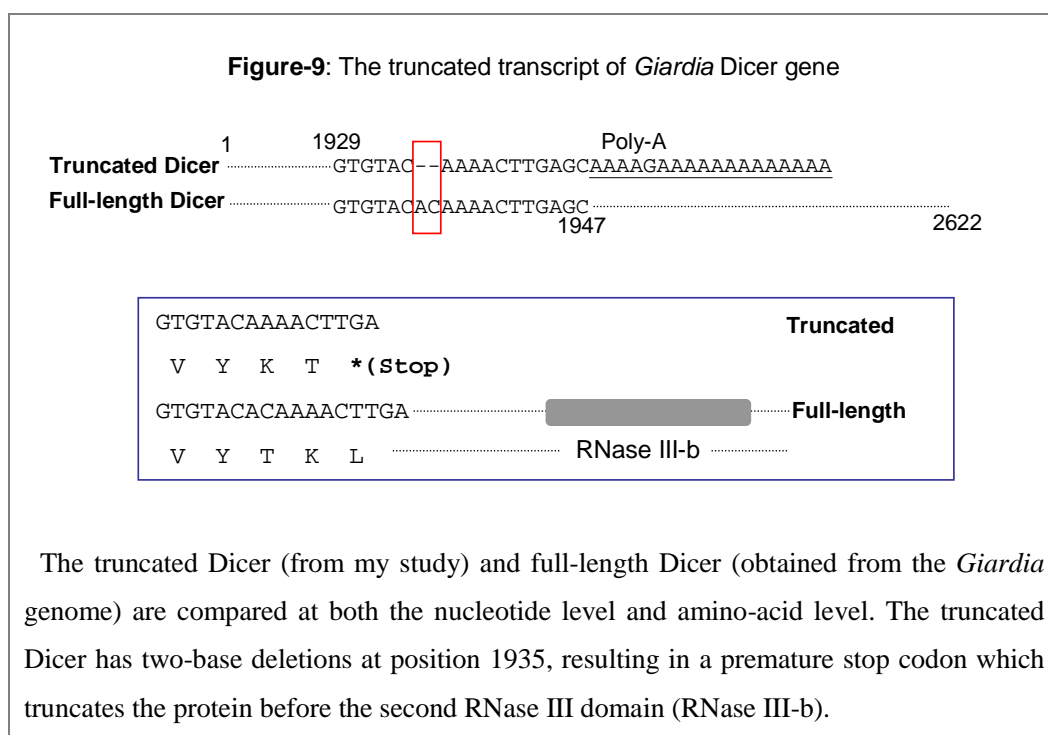
RdRp is not found in archaea. However the presence of Piwi and Ago family proteins in archaea suggest that RNA-induced silencing may have originated before the divergence of archaea and eukaryotes, and subsequently evolved into various silencing mechanisms that involve small silencing RNAs in eukaryotes.

It is unusual that the putative RNAi positive organism *Giardia* does not have an obvious Ago protein homologue. The protein Q86QW7_GIALA in Figure-7 (*Giardia* ID: GL50803_2902) is the only *Giardia* protein that contains a Piwi domain, thus is likely to be part of the putative RNAi pathway. Lacking of PAZ domain in this protein indicates that this protein may function in a similar way as an archaeal Piwi protein however more experimental evidence is needed to support this assumption. The highly reduced genome and cellular architecture of *Giardia* holds the possibility that many types of RNA-processing machinery in this organism exert archaea-like features, which may be represented as reduced number of protein components and reduced protein domains.

In summary, the protein components of the putative RNAi pathway in *Giardia* suggest that *Giardia* has relatively reduced RNAi machinery compared with higher eukaryotes, and some proteins such as the *Giardia* Piwi protein (GL50803_2902) may function in a similar way as archaeal Piwi proteins. Other uncharacterised proteins may be also involved in RNA-induced silencing in *Giardia* through interactions with the *Giardia* Dicer, Piwi and RdRp proteins, however their functions may be more general than RNAi-specific. This type of dynamic protein-protein association is likely to happen frequently in deep-branching eukaryotes with small genomes, such as *Giardia*, where a protein often does not have a set of domains which specify the function of the protein; instead a protein with reduced number of domains can interact with a number of other proteins. In this way, different pathways may form through flexible protein-protein interactions. Formation of the putative RNAi pathway of *Giardia* may reflect some general features of large RNA-processing machinery in early eukaryotes where protein domains were not yet fused into large proteins.

5.4 The possible existence of a truncated Dicer protein

My early study of *Giardia* Dicer mRNA revealed a truncated transcript. 3'-RACE analysis showed that this truncated transcript was poly-adenylated and was the result of a base-deletion, which terminated the mRNA before the second RNase III domain (Figure-9). The full-length Dicer mRNA was not detected in the same assay. My result showed the possible presence of a truncated Dicer protein, which contains only one RNase III domain (Figure-9). It is unknown whether the truncated transcript resulted from a partial degradation of the total RNA. However, other RT-PCR reactions did not show degradation of mRNAs (Appendix-4).

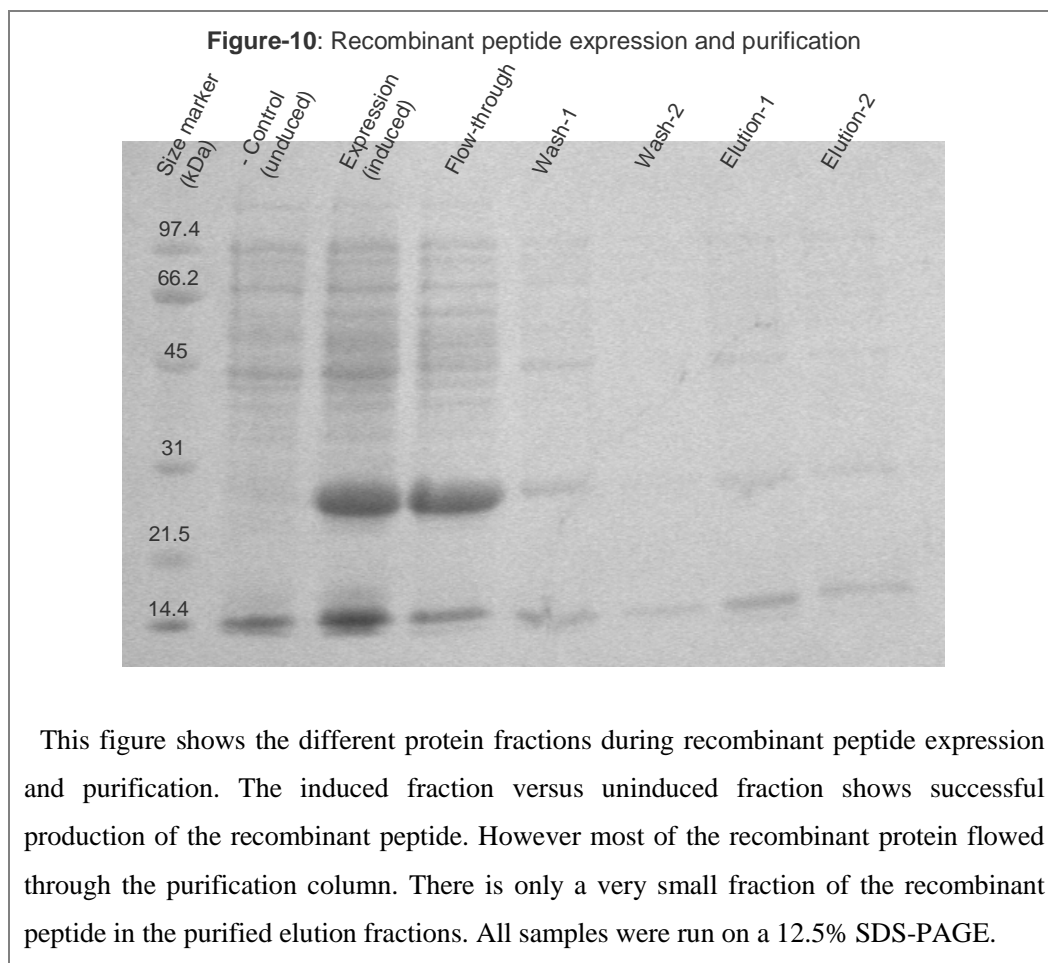


As shown in the above figure, the truncated transcript has a two-nucleotide (“AC”) deletion (boxed in red) at nucleotide position 1935. Deletion of “AC” has resulted in a frame shift which produces a stop codon at nucleotide position 1941 on the truncated transcript, which now lacks the second RNase III domain. Therefore a putative truncated Dicer protein with only one RNase III domain may be produced. The result of the RT-PCR (Appendix-4) was puzzling because the expected full-length transcript was not seen and the reason for this still remains unclear. It is possible that the deletion occurs in this specific

Giardia strand used, leading to silencing of Dicer function. The discovery of the truncated Dicer transcript led to a small investigation of the possible function of the putative truncated Dicer protein with a single RNase III domain. Due to time limitation and the unexplained nature of this truncated transcript, investigation of this problem was not in-depth, but provides useful information for future analysis.

My results show that the sequence of the RNase III domain of the *Giardia* Dicer protein is more similar to the bacterial RNase III domain than to eukaryotic RNase III domain. Despite the well characterized structure of the full-length recombinant *Giardia* Dicer protein (Macrae et al. 2006), it remains an interesting question in this study whether a single RNase III domain of *Giardia* Dicer protein is capable of cleaving native dsRNAs from *Giardia*.

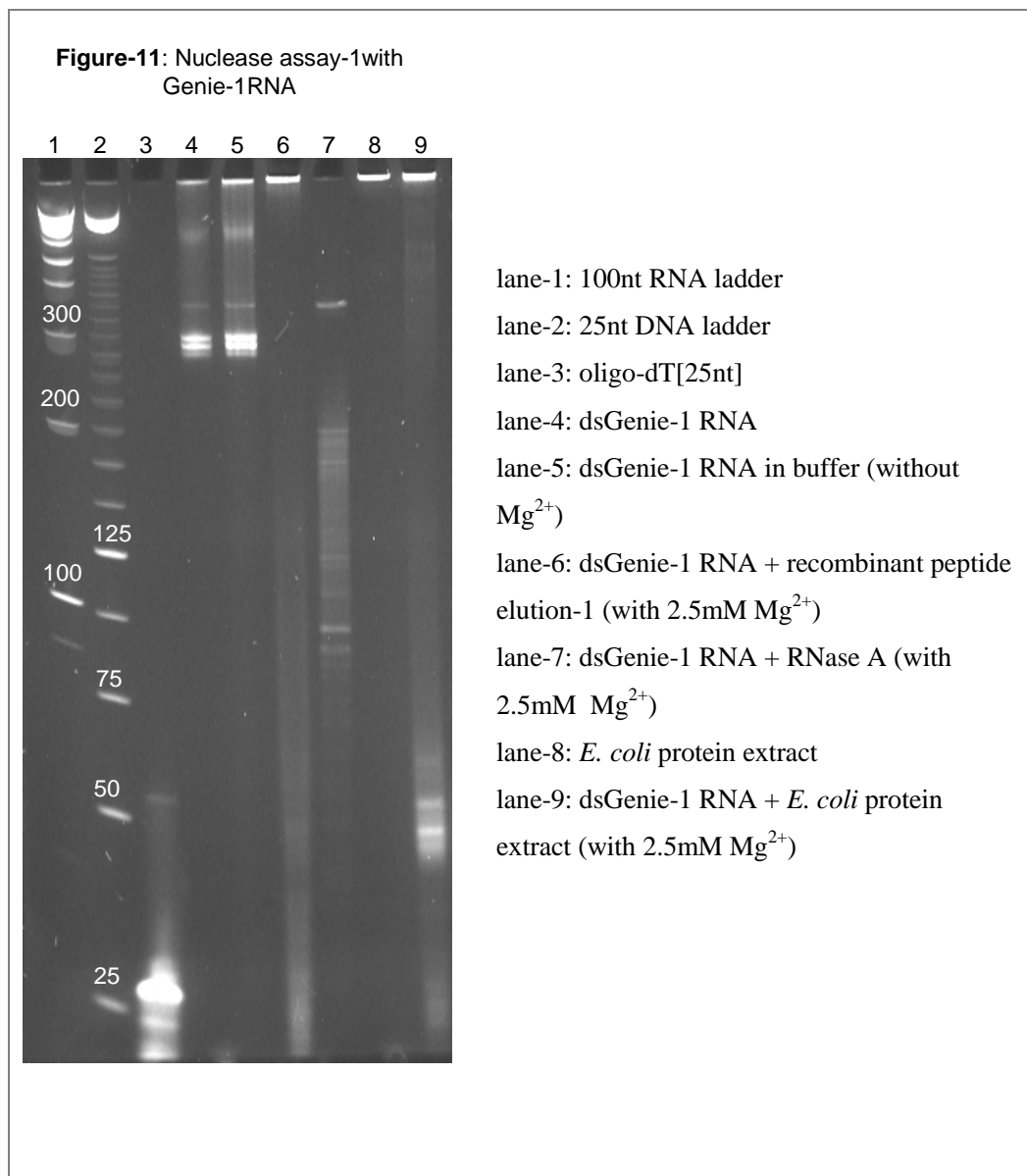
In order to test the possible function of the putative truncated Dicer protein, a recombinant peptide containing the PAZ domain and the first RNase III domain was cloned and expressed in *E. coli* with an N-terminal 6×His tag. Expression yielded significant amount of recombinant protein, however purification using Ni²⁺-charged column resulted in the majority of recombinant protein being in the flow-through fraction and only a small amount of the recombinant protein present in both the washing fraction and final elution fractions. Also a distinct amount of *E. coli* native protein was seen in the final elution fractions. These results are shown in Figure-10. Different buffer conditions were applied but did not improve the result of the purification. This result could be due to the native purification condition required for retaining protein structures, and the spin-column was not fully capable for this purpose. Therefore in the following tests, eluted protein samples were used and the whole *E. coli* extract was used as a control.



The first test of the recombinant peptide was carried out using *in vitro* generated *Giardia* Genie-1 (Ullu et al. 2005) dsRNA as substrate. Results of tests with Genie-1 dsRNA are shown in Figure-11. Using the elution fraction of the recombinant extract (containing the truncated Dicer peptide) resulted in a major group of short RNA products of size around 25 nucleotides as well as a light smearing indicating incomplete cleavage of the RNA substrate (lane-6 in Figure-11). Adding the native *E. coli* extract also resulted in disappearance of the Genie-1 band, but the majority of cleavage product was around 50nt and the minority around 25nt (lane-9 in Figure-11). These short RNAs seen in lane-9 were likely to be produced by *E. coli* native RNase III. (The RNase A reaction was included as a control. RNase A cleaves the ssRNA 3' of pyrimidine residues.)

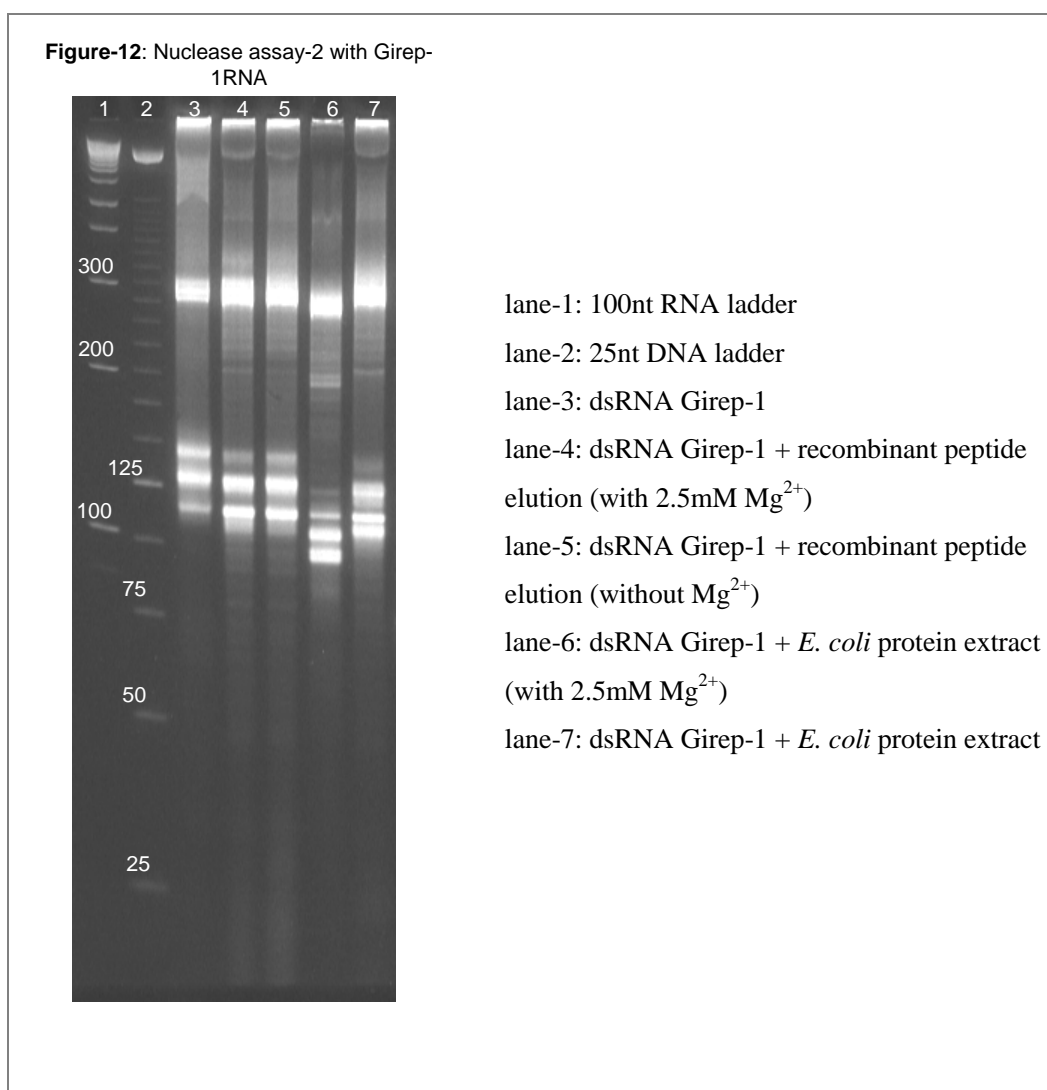
In lane-7 of Figure-11, the presence of a degraded fraction is likely due to incomplete dsRNA formation, thus leaving single-stranded regions unpaired.

The top band in lane-7 should correspond to complete base-paired dsRNAs of Genie-1. Results in the first nuclease assay show that both the recombinant peptide elution and *E. coli* protein extract have ability to cleave a mixed solution of complete and incomplete dsRNAs of Genie-1, and they result in different cleavage products. This result suggests that the single RNase III domain within the recombinant peptide is likely to bind and cleave *Giardia* native dsRNAs but the function is less specific than the full-length Dicer protein which measures and cleaves dsRNAs into equal-length 25nt small RNAs (Macrae et al. 2006).



Due to being unable to obtain a pure expression product, this nuclease assay may not accurately show the possible function of the truncated *Giardia* Dicer

peptide. However, the clear difference between the addition of the recombinant peptide elution and the *E. coli* protein extract suggests a distinguishable nuclease activity of the Dicer peptide on the RNA substrate, which was a mixed population of fully complementary-bound dsRNAs and partially bound sense and antisense RNAs of *Giardia* native RNA Genie. It is not known whether the recombinant peptide could function as dimers during cleavage of RNA substrate like *E. coli* RNase III (Zhang et al. 2004), however this possibility is high. To test the consistency of the nuclease assay on other *Giardia* native RNAs, a second experiment was done using *in vitro* transcribed *Giardia* Girep-1 dsRNA repeats (see Figure-12).



Unexpectedly, both the recombinant-peptide elution and the *E. coli* extract had little effect on dsRNAs of Girep-1 under conditions with or without 2.5mM Mg²⁺, except in lane-6, where a small reduction of RNA size was observed

when *E. coli* protein extract was added with addition of Mg^{2+} , suggesting a possible short cleavage from one end of the RNA substrate. Resistance to nucleases was likely due to the structure of Girep-1 RNA under the reaction conditions. It is also known that bulges in dsRNA structure can cause RNase III resistance in *E. coli* (Calin-Jageman and Nicholson 2003). Therefore resistance to the *E. coli* nuclease observed with Girep-1 RNA could be explained by the possible formation of bulges within the dsRNA structure. Studies of various eukaryotic Dicer substrates showed that the 2-3nt 3'-overhang of dsRNA was important for recognition by PAZ domain, whereas in the absence of 3'-overhang (e.g. blunt end), sequence of the dsRNA substrate had a more pronounced effect on Dicer cleavage (Vermeulen et al. 2005). However a recent study showed that wild-type *Giardia* Dicer could cleave both perfect and bulged dsRNA substrate from either 5'- or 3'- end (Macrae et al. 2007). Therefore at this point it is not clear whether the structure of RNA substrate or the nature of recombinant peptide with single RNase III domain was the cause of nuclease resistance of Girep-1 dsRNA. Tests using purified *Giardia* native Dicer protein on the same RNA substrate may reveal the reason behind variable results observed here, but this is beyond the scope of this study. It is hoped that future study will look into the detailed mechanism of the putative *Giardia* RNAi mechanism.

Results of this study suggest possible nuclease activity of the truncated Dicer peptide with PAZ domain and a single RNase III domain. Using *in vitro* transcribed *Giardia* native RNAs, different results were seen. Appearance of small cleavage products around 25nt in the experiment with Genie RNA suggests that the truncated Dicer peptide may function in a similar way as the full-length Dicer protein, although the defect during purification process reduces accuracy of the assay. The inconsistency seen in the two nuclease assays strongly suggests some structural impact of the RNA substrates on enzymatic activity of RNases. Current findings show that Genie-1 RNA is very likely to be processed by *Giardia* Dicer, but properties of Girep RNAs are less certain. The occurrence of a truncated Dicer transcript may be caused by an unusual RNA-processing mechanism, which resulted in a two-nucleotide base

deletion and early termination of the mRNA. However it is not known whether this unusual transcript only exists in this particular strand of *Giardia*.

5.5 Conclusion and an overview of the unusual RNAs found in this study

Five unusual tandem-repeated ncRNAs are described in this study with assigned names as Girep RNAs. All five Girep RNAs are highly transcribed on both the sense- and antisense- strands and have the potential to form dsRNAs. They consist of repeats that are highly similar in sequence to a number of *Giardia* VSP genes. Four out of five Girep RNAs show self-cleaving property at presence of Mg^{2+} . A high degree of sequence and structural conservation is seen among the Girep RNAs. Current knowledge on RNA tandem repeats is limited, but it is highly likely that these RNAs are involved in regulating homologous VSP gene expression, and possibly related to development of *Giardia*. It is also possible that the putative dsRNAs found in *Giardia* may serve as substrate for Dicer protein during RNAi regulated gene silencing which is a highly conserved mechanism in eukaryotes. *Giardia* proteins of the putative RNAi pathway are compared across different eukaryotic, bacterial and archaeal organisms. Results from sequence and protein domain comparison show that *Giardia* and other protists generally have fewer protein domains clustered on a single protein; therefore dynamic protein-protein interactions are likely to be an important feature of RNA-processing pathways in deep-branching eukaryotes. A primary study, which was triggered by the discovery of this truncated Dicer transcript, looked into potential functions of a truncated Dicer peptide. Results indicate possible nuclease activity of this truncated peptide on one *Giardia* native RNA substrate. However due to the quality of purification, at this stage it is not certain whether a single RNase III domain is sufficient to cleavage dsRNAs into 25nt small RNAs.

In summary, the unusual RNAs described in this chapter have suggested a link to the putative RNAi pathway in *Giardia*. How the putative RNAi mechanism regulates gene expression in *Giardia* is not yet clear, but it is likely to involve protein-protein interactions similar to the putative RNA-induced silencing mechanism in archaea. It is hoped that future studies will reveal the details of ncRNA-regulated gene expression in *Giardia*.

5.6 Experimental materials and methods

5.6.1 Primers used in generating *in vitro* transcription template for Girep RNAs and Girep-1RNA and primers used for recombinant truncated Dicer peptide:

Girep1_F	TGCAGCCCTTCTTGTCGGG
Girep1_R	GATACCCGGCTGTGCGACGT
Girep2_F	AGACGTGCCTGGACTG
Girep2_R	TTGGGGCCGCACGTC
Girep3_F	GCGCACAAGAGCTGAC
Girep3_R	GCGCTGGGATTGGGCT
Girep4_F	GGGTTGGCCGCAGAGT
Girep4_R	CCTGAACAGCGGGAAG
Girep5_F	GCACGTTCGTCCTCT
Girep5_R	ACCGACGCCGGCTGT
Genie1_F_T7F	TAATACGACTCACTATAGGGAGACGACCCTCTTCTCCAGCA
Genie1_R_T7R	TAATACGACTCACTATAGGGAGAGGAGCGCAAAGAGGATGA
GiDcrShort_F	ATAGGCGGCCGCACTCCCGAAATAAGCTGGT
GiDcrShort_R	TCGAGTCGACTACTGTCTTGGAGTGTTTCG

5.6.2 Recombinant truncated Dicer peptide expression and purification

The DNA sequence of Dcr-short was amplified using Dcr_short- forward and reverse primers and double-digested by BamH1 (Roche) and Not1 (Fermentas) restriction enzymes. The digested DNA was inserted into the pIVEX-2.4d vector (Roche) and cloned in DH5 α cell (Invitrogen). After obtaining purified plasmids containing the Dcr-short sequence with an N-terminal 6 \times His tag, the plasmids were transformed into KRX cells (Invitro technologies) and plated onto LB agar plate with 100 μ g/ml ampicilin.

A single colony was picked from the LB plate next day and inoculated LB media with 100µg/ml Amp, and incubated overnight with shaking at 37°C. The next day a fresh LB media with 100µg/ml Amp was inoculated with the overnight culture at 1:100 dilution and was incubated at 37°C with shaking until OD was around 0.5. The culture was then transferred to the 25°C room for continual incubation until the OD was around 0.7. 0.05% of L-arabinose was then added into the growing culture for induction of recombinant peptide expression. The expression culture was then incubated at 25°C with shaking overnight.

After expression, cells were collected and lysed in lysis buffer (50mM NaH₂PO₄, 300mM NaCl, 0.5% Triton-X-100). The crude extract was obtained by centrifugation of the homogenized cell lysate. Ni-ATA columns (Qiagen) were equilibrated with lysis buffer first before running the crude extract through. Three washes were done with washing buffer (50mM NaH₂PO₄, 300mM NaCl, 50mM imidazol). Finally, proteins were eluted using elution buffer (50mM NaH₂PO₄, 300mM NaCl, 250mM imidazol).

All samples were run on a 12.5% SDS-PAGE gel.

5.6.3 Nuclease activity assay

All reactions were carried out in the reaction buffer (10mM MgCl₂, 10mM Tris-HCl, 300mM NaCl). The RNA and protein mixtures were assembled on ice and incubated at 37°C for 1 hr and loaded onto 6% native PAGE. After running in 1×TBE buffer, the gel was stained with EtBr and visualized under UV.

Final words

The studies described in the five chapters have built an overview of various ncRNAs in *Giardia*. Through the construction and analysis of the *Giardia* ncRNA library, candidates belonging to major categories of eukaryotic ncRNAs have been identified, including C/D-box snoRNAs, H/ACA-box snoRNAs, and the RNase P, as well as uncharacterised novel ncRNAs. Computational search identified candidates for four *Giardia* spliceosomal snRNAs. Despite the majority of the ncRNA candidates found in this study being currently uncharacterized due to the long divergence of *Giardia* from other eukaryotes, it is clear that most of the major RNA-processing pathways exist in *Giardia*, including pre-tRNA processing, pre-rRNA processing and pre-mRNA splicing. Computational analysis of characterised *Giardia* ncRNAs shows that most of these RNAs fold into slightly reduced structures compared with those of other eukaryotic models, but functional sequence motifs are highly conserved.

Results from the *Giardia* ncRNA library of this study represent a random sampling of the overall ncRNAs contained in the library. Analysis of expressional patterns of my ncRNA candidates suggests a possible common feature of *Giardia* ncRNA being transcribed in either intergenic regions or on antisense-strands of protein-coding genes. Potential RNA polymerase II and III promoter sequence elements of *Giardia* ncRNAs have been identified from the analysis. The conserved sequence patterns of these elements exhibit distinct *Giardia*-specific features which differ from several common promoter patterns seen in animal and yeast models, mainly *Giardia* has fewer and less well conserved potential promoter elements than animals and about half of the ncRNA candidates identified in this study do not show detectable promoter signals. This observation suggests that while the basic mechanisms of RNA transcription in *Giardia* are typically eukaryotic, the detailed process may be reduced, such as involving fewer protein factors and less tight expressional control.

The studies described in this thesis involve combining molecular biology methods and computational methods. The molecular biology methods were used for random identification of ncRNA candidates from the *Giardia* ncRNA library and for confirming expressions of computationally identified candidates. The computational methods used here have shown promising ability of searching RNA candidates in a highly diverged protist genome based on collective information of characterised eukaryotic ncRNAs. 60 putative snoRNA candidates identified in Chapter-2 and the 4 U-snRNA candidates identified in Chapter-4 are successful results of the combined experimental and computational approach.

In addition to the typical eukaryotic ncRNAs found in *Giardia*, unusual RNAs are also identified in this study. Chapter-5 has discussed the properties of a *Giardia*-specific group of RNA tandem repeats that are transcribed in both sense- and antisense- directions and their possible relation to the putative RNA-directed silencing mechanisms in *Giardia*. Evidence for this type of RNA is rare. To date, only one other study has identified RNA tandem repeats, that in *Leishmania* (Dumas et al. 2006). However the possibility of similar RNAs being found in other eukaryotes cannot be excluded.

In this thesis, *Giardia* has been used as a model organism representing genomically and cellular-architecturally reduced unicellular protists, which are highly diverged from most other eukaryotes. Studying unicellular protists can provide important information for understanding early eukaryotic evolution, because common features shared between diverged groups of organisms are most likely to present ancestral features of all eukaryotes. Results from analysing various ncRNAs of *Giardia* show both conservation of typical eukaryotic RNA-processing pathways and some unique features of *Giardia* RNA-processing, as expected.

One important outcome of this study is that the conservation of many typical eukaryotic ncRNAs suggests functional continuity of these ncRNAs through eukaryotic evolution, and an early origin of these ncRNAs. Besides continuous descent with ancestral functions being maintained, ncRNAs can

also evolve through hierarchical expansion and acquire new functions. In such cases, changes in protein components in an RNP complex result in adaptive changes in the ncRNA components, and eventually the ncRNAs evolve functions different from their ancestors. In animals, hierarchical expansion of ncRNA functions is common with examples such as the Xist RNA in human and rox RNAs in *Drosophila*. It is possible that some of the uncharacterised *Giardia* ncRNAs identified in my study have evolved to perform specific functions in *Giardia*, such as the tandem repeated RNAs discussed in Chapter-5.

To further investigate the RNA processing in deep-branching eukaryotes, a new project is planned to study the small ncRNAs from *Giardia*, *Trichomonas* and *Cryptosporidium* using the Solexa sequencing technology (Bennett 2004). This technology allows large coverage of short 35-mer sequence fragments without cloning. It is expected that this new project will give very wide coverage of small ncRNAs from the three organisms, including the expected microRNAs.

To conclude, this thesis has identified and analysed a number of different ncRNAs from the deep-branching protozoan parasite *Giardia intestinalis*, and compared the RNA-processing of *Giardia* to that of other major eukaryotic model organisms. The study shows that major RNA-processing pathways in eukaryotes are evolutionarily highly conserved, even though quite diverged in some groups. Results described in this thesis will help future study to further understand the evolution of ncRNAs in eukaryotes. Additional studies in *Giardia*, together with similar studies in early diverging eukaryotes will certainly help understand the role of ncRNAs in the earliest eukaryote.

Appendix-1

1. Information and sequences of the ncRNA candidates identified in the library

Candidate	Copy number	Genomic location	Annotation	Length
1	1	3' to ORF 30474, + strand	possible new C/D-box snoRNA	65
2	1	between ORF 7656 and 7655, + strand	possible new C/D-box snoRNA	61
3	1	5' to ORF 4653, + strand	none	106
4	1	on - strand of ORF 8253	none	73
5	1	on - strand of ORF 13601	none	47
6	1	in heterochromatic region	fragment of a transcript containing RNaseP and GlsR15 snoRNA	121
7	1	3' to ORF 15284, - strand	none	41
8	1	on - strand of ORF 13627	none	43
9	1	on - strand of ORF 87422	none	72
10	1	on - strand of ORF 32022	none	86
11	1	3' to ORF 25296, - strand	none	95
12	1	3' to ORF 16285, + strand	none	62
13	1	5' to ORF 5925, + strand	fragment of new C/D-box snoRNA	58
14	1	3' to ORF 14213, + strand	none	136
15	1	3' to ORF 8513, - strand	none	92
16	1	on - strand of ORF 32036	possible new H/ACA box snoRNA	78
17	1	between ORF 28379 and 95192, + strand	none	166
18	1	3' to ORF 16020, - strand	none	120
19	1	between ORF 7870 and 33685, + strand	none	42

20	1	on - strand of ORF 13930	none	133
21	1	5' to ORF 16954, + strand	none	114
22	1	3' end to ORF 19381, + strand	none	60
23	1	in heterochromatic region	none	54
24	1	on - strand of ORF 8460	none	110
25	2	in tandem, heterochromatic region	none	140
26	7	in tandem, 5' to ORF 11976	none	71
27	4	in tandem	fragment of variant surface protein	50
28	4	2 in heterochromatic regions, 2 near ORFs	none	90
29	1	in heterochromatic region	none	87
30	1	on - strand of ORF 98689	none	96
31	1	on - strand of ORF 33735	none	60

>Candidate_1

AAAAAATAAATGAAGACAGAACCACAGACCTGTACTGACCCTTGATGTTAGTTGTCGCTCTGATA

>Candidate_2

TGATGATTCGAATTACCGCCCGAGGGCCCTCGGGCTCCGCTGAGGACATGCTGGTCTGACT

>Candidate_3

CCGATCGAAGACCAAGCGGTGCTAGGTTCAAGCCAGGGCCAAGACCCGGGCAGTCTGTGCTGTGGGGCGCCGCTGTAGACGTCTTCCGAACACACCTGCGA
TAAAC

>candidate_4

CCACACAAAAGGTGAGCGCGTAAGCAAAAGCCAGAAGCCCGTTGCAGCGCTTGCTTCGCAGCTCTACGGGCGC

>Candidate_5

CCCCATGCATTTTCTTGCCAGTCTGCCTCCATACTAATTTCTCCT

>Candidate_6

GATGCGCCCAGGCTGACGGTAGGACGCCTAACCCGATTCAGACTACTCCTTGGTTCTCGCAGAATGATTATCTGTCTCCGAGCAAGCACGACTATGAGCTT
ACTTATGAGATCTGACTCC

>Candidate_7

AGACAGAAGTAGAGCCGTTCTCCAGTAAATCTGCAGCTCA

>Candidate_8
GCTGAGAACGTCAGGAAGGAGCCTAGAAAAGAAGTTGCTGCAC

>Candidate_9
AGGAACCTATATTAGCAGAATTGGAAACGTTATAAGTGGGCTCCATCTTTTGCAGAATGTTGGAAATGTAGT

>Candidate_10
ACGGGAATAACGCCACAGGATCTCAAGGAAGGGGCCCTCAACTTGAAGCTCTGATCGGGTCCCAAGCACAAGTAAATAATTGCC

>Candidate_11
CTAGGCTGAAGCTGCCAAGGTGCGTGATCCCTCGGTGATGCCTTGTGCTTACCAAAGAACAACCACACGGCACAGCCGAATCTCTCATT

>Candidate_12
CCCGATGACGAATAGCTGTCCTGGCGGAGGCGGTCATGACGACGAAGCCATCACGTAGGATC

>Candidate_13
CACGACGGTCTACTGAGAACCCAGTATCTTTAGACTGCTGAGACAGTGTTATATGATT

>Candidate_14
CAGAGTCGGCTTCGACTTTAGCGTAGTTACTGTTTCGTGCGCTTAACCGCCGATCCACTACATGCAAGGGGCAGCCGGGCTGTGAGGCAGCTGCCAGGATGG
TCCTGCCCTTGTCCCGGCTGGCGCCGTCCACCTT

>Candidate_15
CTTCAACTCAGCCGGACAGCCGGAGGCCGGAGACGGGACCGGTCAGGCGGGCGGGGTGCAGTGCCAGCCCCAGCCGCAGAGCGGCTTCCTT

>Candidate_16
CTGCGCTCTGCCAGATACGCCGACAGAAAGCACCAAGGAAGGATGTGGATCTCCATGTCTGCCGTGTGCGCGCATATC

>Candidate_17
TTCGGGATCAGTTTTGGAGTTAATACCACCAAACCCCTGTGCGTACATGTCGCCCCCTAACCTTCTGATGCGGATACCTTGCCGCAGGGCCGTAAAGCGAGGC
TTGGCCCGTGCGACGATGAGGCTCCCTGCGGGGAAGCCCTGCGGCGCGTCTTAAGGAGGCAAC

>Candidate_18
TCTGGATTCCGGGGAGTGTCTGGTGCCGATCGGACACTCCCTAGCCGCCACACTGACAGTTATGGTTGCAGGACAAGCTTAGCGAGTCCGAACTCGACAGGG
ATACTCTACAGCGTTCCT

>Candidate_19
GCGTAAGGTTTTCTGTCAGACTACCCAGAGTAAACCGGTGAG

>Candidate_20
TGTAGGTCTAACATGCTTGCCACGGCGTCCCCGGACATGGCACCGTCTATGTCCTGCTTGTGGCGAGGATGAGGATGGGAACACCTGAGCTTGGGGCTGTT
AGTACGCCCTCAAGAGCCGTCCGAGCCTCCT

>Candidate_21

TGGACGATGAACTGGAGATGCTGGACACGGCTTTGCTCTCCCACCGGAGCACATATGCTGCAGGATGACCGGCGCCTGTCTCCCACCACGTGCCAGCTAAAC
TGCAGCCACATT
>Candidate_22
GTATGCTGCTATGCTGACATGCCGGTACACTTTTTATGAGAGCGAATGTAAATAGCCCTG
>Candidate_23
TACCACTCTGACCGTGAGGCGCATGCCTAGGGCATGGAGAAGAGCAGACTTGAG
>Candidate_24
TGTGCCACTGTGGCTTCGAGCTCTATAATGCGCGACTTAAGAACCTCTGCTTCTACAGACTTTACTTCAAGTAAAGATGTCGCAGTTAGTGCCTCCTCAACAC
AGCTTTC
>Candidate_25
GAGGTAATAGACCAGGCTGCCAGCCCGGCGAAGGTCTGCAAGTGTGACGGAGACAATGGCTACACGCTCCAGGGCGACGCGTGCACCAAGGCGGCTCCTGA
CAACGCGTGCCAGACCCTGGGAACCGCCGGGTGTGCCAC
>Candidate_26
CTGAACGATAGAAAAGACGCGTGCGAGGCGGTTGCCAACACAAACTGCGCACAAGAGCTGCAGAATGCGGC
>Candidate_27
TGCCTCACCTGCACCACTTCGTCGCATAAGATCAGGCCGGACGAGAAGGGC
>Candidate_28
CAGAAGAATGCCAGCAAGTCATGCAATGCCTGTGGATCCGTCCTTCGACCTTCTCCTGACAGACATGTGTCTTTTGGCATGCAGCCCTGC
>Candidate_29
GCTCCTAGAGGAAGAGGCAGGCATGCAGGATATTTTTGGATGGACAGCCCTCATAAGGGCAGCAGAGAGTGGCCACGCAGGCTGCAC
>Candidate_30
CCCACCGGCGTTCCAGTGCTGGCCAGGGGCAAGGAGGCCTGCTCTCCCTGGCCTCTGCGGAAACGGGCAGCTGCGTGATCCACTGACAGCCACCAC
>Candidate_31
GGCGCAGACAACAGCAAGAGTCCAGATGGAGTACCTGCACTCCGCCAAGGTTTAGCGTAG

2. Information of the computationally identified putative snoRNA candidates from *Giardia*

Candidate	Length	C-box	D-box	rRNA complementary length	Mismatch	Predicted 2'-O-methylation site	Total score in Snoscan
U0004	77	AUGAUAU	CUGA	10	0	LSU_rRNA_Am2223	10.76
U0005	97	AUGAUGU	CUGA	9	1	SSU_rRNA_Gm966	10.9
U0007	89	AUGAUGG	CUGA	10	0	LSU_rRNA_Um658	10.53
U0011	124	AUGAUUA	CUGA	11	1	SSU_rRNA_Cm463	10.52
U0013	90	GUGAUCG	CUGA	10	0	SSU_rRNA_Gm457	11.1
U0020	110	AUGAUGG	CUGA	10	1	SSU_rRNA_Am1060	13.53
U0022	90	AUGAUGC	CUGA	12	3	LSU_rRNA_Cm1021	10.15

U0023	63	AUGAUAC	CUGA	9	1	SSU_rRNA_Cm1381	14.15
U0025	160	AUGAUCU	CUGA	17	1	LSU_rRNA_Cm1191	10.13
U0027	129	AUGAUAU	CUGA	10	2	SSU_rRNA_Gm855	10.07
U0029	111	GUGAUGA	CUGA	10	1	LSU_rRNA_Gm406	11.67
U0032	160	AUGAUGC	CUGA	12	2	LSU_rRNA_Am2128	10.41
U0033	118	GUGAUCG	CUGA	11	0	LSU_rRNA_Um1575	17.27
U0034	117	GUGAUCG	CUGA	13	2	LSU_rRNA_Gm1109	10.67
U0035	117	AUGAUCC	CUGA	13	1	SSU_rRNA_Cm726	10.3
U0036	125	AUGAUCG	CUGA	11	3	LSU_rRNA_Cm1650	10.06
U0037	117	AUGAUAA	CUGA	12	2	SSU_rRNA_Cm61	16.8
U0041	123	AUGAUCG	CUGA	11	1	SSU_rRNA_Gm1177	10.31
U0048	115	AUGAUGG	CUGA	23	2	LSU_rRNA_Gm2031	23.98
U0050	105	AUGAUGC	CUGA	10	1	5.8S_rRNA_Um18	16.2
U0052	98	GUGAUCC	CUGA	14	1	LSU_rRNA_Am407	12.39
U0054	93	AUGAUAG	CUGA	12	1	SSU_rRNA_Cm262	12.65
U0055	96	AUGAUUC	CUGA	18	3	LSU_rRNA_Gm790	17.57
U0056	73	AUGAUAG	CUGA	11	1	SSU_rRNA_Cm1059	10.3
U0058	85	GUGAUGU	CUGA	10	1	LSU_rRNA_Cm210	11.15
U0060	89	AUGAUCU	CUGA	9	1	SSU_rRNA_Gm1343	12.11
U0061	90	AUGAUGC	CUGA	12	1	LSU_rRNA_Gm2352	10.02
U0063	80	GUGAUUG	CUGA	13	1	SSU_rRNA_Cm150	14.59
U0065	120	GUGAUGC	CUGA	12	0	LSU_rRNA_Cm771	18.23
U0066	72	GUGAUCG	CUGA	14	1	LSU_rRNA_Cm23	12.63
U0067	162	AUGAUUC	CUGA	15	3	LSU_rRNA_Cm1244	14.65
U0068	148	AUGAUGG	CUGA	14	2	LSU_rRNA_Gm1584	11.8
U0069	75	AUGAUUA	CUGA	14	1	LSU_rRNA_Cm241	12.29
U0070	129	AUGAUUG	CUGA	17	3	SSU_rRNA_Um1423	15.37

U0077	90	GUGAUCA	CUGA	13	1	SSU_rRNA_Gm586	10.21
U0080	74	GUGAUGC	CUGA	10	1	5.8S_rRNA_Cm21	10.31
U0081	73	AUGAUGC	CUGA	15	2	LSU_rRNA_Cm2132	12.17
U0082	85	GUGAUGG	CUGA	13	2	LSU_rRNA_Am_2016	11.99
U0083	84	AUGAUGC	CUGA	11	1	SSU_rRNA_Cm1231	13.48
U0084	123	AUGAUGU	CUGA	15	2	LSU_rRNA_Um1837	10.67
U0086	142	AUGAUUU	CUGA	12	1	SSU_rRNA_Cm701	11.64
U0087	73	GUGAUCG	CUGA	17	3	LSU_rRNA_Cm1099	11.32
U0089	96	AUGAUCC	CUGA	11	1	LSU_rRNA_Gm790	10.72
U0090	130	GUGAUUG	CUGA	19	3	5.8S_rRNA_Cm7	14.37
U0092	90	AUGAUGA	CUGA	12	1	LSU_rRNA_Gm2826	15.77
U0094	106	AUGAUAG	CUGA	11	2	LSU_rRNA_Cm918	10.79
U0098	97	AUGAUAA	CUGA	12	1	LSU_rRNA_Cm2863	10.9
U0100	135	GUGAUGU	CUGA	11	0	SSU_rRNA_Gm473	10.08
U0101	128	AUGAUGA	CUGA	12	1	SSU_rRNA_Am549	11.89
U0104	88	AUGAUUU	CUGA	10	2	LSU_rRNA_Am324	10.85
U0105	83	AUGAUCC	CUGA	11	1	LSU_rRNA_Cm1586	11.14
U0106	92	AUGAUUG	CUGA	10	1	LSU_rRNA_Um831	11.27
U0107	100	GUGAUGU	CUGA	14	1	LSU_rRNA_Cm636	10.62
U0108	101	AUGAUGA	CUGA	13	1	SSU_rRNA_Cm547	11.86
U0109	104	AUGAUUU	CUGA	17	1	LSU_rRNA_Um863	25.38
U0114	78	AUGAUGC	CUGA	10	1	LSU_rRNA_Cm1375	12.5
U0115	89	AUGAUAG	CUGA	11	1	LSU_rRNA_Am941	11.49
U0121	85	AUGAUUG	CUGA	18	0	LSU_rRNA_Cm619	10.33
U0122	110	AUGAUGG	CUGA	13	1	LSU_rRNA_Gm83	12.03
U0123	105	AUGAUGA	CUGA	13	2	LSU_rRNA_Gm88	12.05

3. Sequences of the computationally identified putative snoRNA candidates from *Giardia*:

```

>U0004
AACUAUGAUUUUCCCCUUUCUGCCAGAGCUGGACAGUGUCCGAAUUUCCGAUUCCAAUUGCUGGACACACUGAGC
>U0005
AAGGAUGAUGUAAAUUACUAUAGCGUGUAAAAGGUACUCCGCCUUGAUCUGCCUUAACUUUACUCUACUAUUUUACACUAAGGGCCUCUCUCUGAGU
>U0007
ACAGAUGAUGGUUCAUUUCAUCUAUUUACUAGAAGGCAACUAGCCGAUAGUCUGAUAAAGUAGCCUCUUCUUAUGCAUGUUAACUGAAU
>U0011
ACUCAUGAUUACAAGCAGCUCGAGCCGAAUCCAGGGCAAAAAUAAUUUUAAACCAUGUUUUUGUCCCCUUGAAGCAGAGCGAUUUACUCCAUCCCUAUUCUAGUAG
AAGUUACCACUGAGU
>U0013
AGAAGUGAUCGAU AUGUCGAAUUAACGCUCUGUGGCUUGGUUGCCUUCUUGGAUCACGCAAAGGCAAGGCACUGUAUCCUAAACUGAAU
>U0020
AGCGAUGAUGGCAGGUGCGGUGCUCUCGUUAUCGCUGACGGAUCCACAAGGGUGAGCAGCCUCCCGCCUGGCCCCCGUCAGGAAGUGACACCGGCGUACUGUCCCUGAGA
UGUCCUUGAGAGCCGAGACCCCGCUGACA
>U0022
AGCUAUGAUGCACAUGCAUAAGUUACAGCUGUUACUUUCUGAUAGUCUUCUUAAGCUUAUGAAGCCACUCGCGAGUCAGUGGCUCUGAGU
>U0023
AGGAAUGAUACCCACCUGCUCGCGUCCUGAGGCUUAGUUGCUGGUC AACCCUGGCCACUGACU
>U0025
AGUGAUGAUCUAGUGGCAGAAUACCAGGCAGCAAGCACCAGAAUAAGGUGUGUUGGCCUCCAAU AUUCGAAGAGCAGAGGUUGCUGCCUCGGCUAGAUGCAGGCCAAUUG
AAGGAAGAGGAUCGUGAAUAGUAGCAAUGUGCGAAACUUCGUUUGCUGAGG
>U0027
AUCUAUGAU AUUGGUCGCUUCCCAGUAGCAAUACGCCCCCGAUUCU AAGCUUCUUAAUUAACUAAGUUACAGAAAGACGCUGCCAAUUUAAGCUCUUCUCAUUCAUGG
ACCAUCUAUACAUCUGAGU
>U0029
CAAGGUGAUGACCACAGGCGAUAGUGAGCCGUGGCUGCUCACUGGUCUAGGUCUCAAUACCAACAUUAUCCACGUCACGGACAGCAUCCCCAGCGAAUGGACCCUGAU
C
>U0032
CAGAAUGAUGCGAGCCAUGUAAUUGAAAUGUGAAUGCAGAUUUUAUUUCAGCUGCAACCUGAAGGGCUCAGUGGCCCGAAGCCUGGGGCGGCAUGCACUGGAAGGCAUC
GAACCUUUGGAUGUCGACGGUGUCUGCAGCGGGGGCAGGCUUGCUGAC

```


>U0033
 CAGAGUGAUCGCGGGCGCCCGUCAGAGCCUGAGCCGACAGGCCCGCAGUGAAGUGCGCCUGGAUCCACGGAGAUCGUUCACUCCCCAGUGGGCCGCAGCUUCCCCUGCCC
 UGCUGACA
 >U0034
 CAGAGUGAUCGUGGUACCCAUCAGAGGCCGAGCCGACAGGCCCGCAGUGCAGCGCCCCUGGACCCACGGAGAUCGUUCACUCCCCAGUGGGCCGUAGCUUCCUCCGCUCU
 GCUGACA
 >U0035
 CAGGAUGAUCCAGAGGGGAGGAUAAGAGUCCUGAUCCAUCAGCUCAGGGGGUGGCACAGCUUGGGCAGGGUUUCCUCCUUGUCUAAAUUUAGGAUUGCUGCCUACGCC
 CUCUGACC
 >U0036
 CAGGAUGAUCGACUGCGGGGUUGGUACUGUGCAGAGUGCAGCCCAGUGACUUGUGUCCUUCAUUCAACACCGUCCAGUACUACGUUCGUUAGGAGAGGUGAUGGGUGC
 CUGGAGUAGUCUGACG
 >U0037
 CAGUAUGAUAAAACAUAUCCGAUCCGUGUGAGAGGGGCUGAUUGGCAUCAGUUGCGCCAAUUGCAACAGAUCCUCUAUCGUCAGAUACGGAAUAUCUCCUAGUUAGAUG
 GACUGAGA
 >U0041
 CCUGAUGAUCGCCGCCUCCCUUGAUUCCCGAGGAGAAUCUAGUUGAUAGCGAGGGCAGUACGGCCCAGCCAGUACGGACAGAGCCGUGAUACUCGUACGAGCAGCACUGC
 ACAGCCACUGAAA
 >U0048
 CGGGAUGAUGGUACUGCAGGCGAAGAGGGCCUCGCGCUCGCCGCCCCAGUCGAUCUCCUACUAAAGAGCAAGCGUGUGGGGACUUCUAAGACGGUUGCCGCUAUCAC
 UGAAC
 >U0050
 CUACAUGAUGCGUGUCUCUGUGGGACCGAUAGAGGCACCCGCUGAUCUGGAAAGACCAUCAGCAGUGGCCACACAAGAUAGACACGCAGAGGACGAGAGCUGAUCAGAU
 CACCCUGGAUGAUGAGC
 >U0052
 CUCAGUGAUCCAUCUGAGCCCGCACCCUUCAGGCUGGCGCUUGGCUCUUGGUUAGUUCCAUCUCCUGGGCUGUGACUCAACCAGACUCUGCUCUGAGG
 >U0054
 CUCCAUGAUAGGGAGAUUCAGCGGAGAGAGGUGUGGGGCCCCUACUCAUUCUAUCGGAUGCUUCCAGACUCCAUCUUCGAGCCCCUGAGA
 >U0055
 CUCUAUGAUUCACUCCACUCGCCGGCCAUGCCCAGCCCAGCCAGCGGGCGUCUCACUCAGUCCUGGCCGUCCGAAGGCGAGGGCUGGUCGCUGAUU
 >U0056
 CUUUUAUGAUAGUUACUGUGGCAAGGGAGUAACGCUCGUGGUCGACAGAGCACAGCUGACCGGGCGAACUGAAG
 >U0058
 GAAUGUGAUGUCCGAGCGAUGAACACGUCGGCUCCUUGUAUGGGUAACGUGUAACCUCCUGGGCACCCGCAGGCCUUGACC

>U0060
GACGAUGAUCUUAACCAGCGUGAGACCCGCCAGUGCGCAACGACUCACACGAGAGAAUGUACCCAAACAGCAGCUAUCAAACUGACC
>U0061
GAGAAUGAUGCUGUAGGAUGACGGCUUUUGAGGUACAUAACCAUUUAUCACCAAGAGAGAAGCUUUGGACCUCGUCUCUUGCCCCUGAAA
>U0063
GAGCGUGAUUGUAGCAAAAUAUGCACCACGCCUUGGGGCCUCAAAAGGUCUGUGUGGUGUCAGCUAAGCAGUGAGCUGAGA
>U0065
GAGGGUGAUGCAUAGCUCUCUGCUAGCUAGGAAGCGGGUGGCCACUGACGGGACGACUGGACGUAACCACGUUAAGCUUGAUAGGCACGUUGAGCGCAGCGUUUUUAU
GGCCUCUGAAU
>U0066
GAUCGUGAUCGUGUGUCUUCGCUCCUCGCCCAAGACAGACCGCUCGGAAGGGGCGAGGGGUCUCUGAAC
>U0067
GAUGAUGAUUCACUCCGCCACCGGCUGCCCCGGCGCUAGCUAUUGAGCUCGCGACGGACCCUCCAUCUUGGCCGUGGAGCGGGGCUCCUCAAGGCAGCAGGGGGCC
AGCGCGGUCGCCGCUCCAUCCCCAGCCGCGGGCUAGCACGCCACUGAGC
>U0068
GCAGAUGAUGGGCUCUGCCACGUACUGUAUUCUCAUCACGUUCCGACGCGGCGGACAGAGACUACGGUUCUUUCAGUUUGUGCUGCCGCAGGCCUUGUAGUCACUCA
GCAGCCUGCUGUUUUUACCGAUCGGAUGUUUCCUGAAG
>U0069
GCAUAUGAUUAGCACCACGCAUUUGGUUUAACAAACUAUUUAGGGCGUUAGACCCUUUGCACUGCACACUGAGU
>U0070
GCCAAUGAUUGCACAAGCUGUAAGUCUGGAAGCACGCUCACCUACGGAUCCACAGGAAACACUGGCACAUGUGGGGCCGAGUGCGCGGGGCACAGGCACAGGCAAGUG
CAGGGAGUGCGGUCUGACU
>U0077
GGACGUGAUCAAAGGCGCGCCCGCUGUAACGAGCUACCUCGCCUUGCUAAUAACGAUAGUCCCAAUCGUGGGAAUUGGCUCGCACUGAAC
>U0080
GGAUGUGAUGCAACCUAGCCGAGACAUCUGCGCUUCUAGGGAGUGGCUGGUCAGUACCUAGAGGCAUUCUGAUG
>U0081
GGCGAUGAUGCGCACGGGAAAGCACUGGUCGGGCGGGCAAUGCCGGCGACGCCUUAGUGGACCCGCACUGAUU
>U0082
GGCGGUGAUGGAGUAUUCACGUGGGUUUUUGCAGGGGCGGGUCUUUCUGAAGCAUACCGCAGGCAGAUUCUGCGAUACACUGACCCACUACCAUUGACGGGUGCUCUUG
CGCUAGUCUUGUAUCUGGUUCGUGAUG
>U0083
GGCUAUGAUGCAUCUGACUAAGCGUAUGUAAGAAGCCCGCAGUUCACUGACUCUCCCCGAGAAACGCAGCUGCUGACUGAGG
>U0084

GGGCAUGAUGUCCGGUAUGCAAUAAGCACCACCUCACAGUGCACGCACACGCCGGCAAGCUUAUGCGGCCAUUACAUAGACUUGGAGUGAAUGAGCGAUGUGACGCUCG
UGGAGGAGCUGAUG
>U0086
GUCUAUGAUUUUACCUACUCCGUCAUUCAUCUUCGCAACCUCUCAGGCCUCCUCUAGGAGCUUACUGGCACUGUAGCAUGCGGGCUUGGCAGUUGACAGGAUGUGCCCA
UCCAUGGUUCAUGCAAACUAGUCUGCUGAAU
>U0087
GUGAACGAUCUCAGUGGGUCCAGGGCACUGCACUGCGGGCCCCGUCGGGCCGGUCUCUGACGGGCGCCGGAUCACUCUGCCCCGGUCUGGCAGAUCACGAGGGGUCUGA
GG
>U0089
GUGAAUGAUCCCACCACAAGAUUGCCUGUCUGGAGUAGGGACCUGGGAGUACUCGACCUAGGUGUCUCCAAACCUCACGACCAGGCCGGUCUGACG
>U0090
GUGCGUGAUUGUAUUGACCGACGCAUUCUAGCCAUUUGCCGGCGCGGCUCGGACAGGAUCAUUCACCUCUCGUCCGCGUUGAUAAAUAUGACAAGAUACACGUCAAGA
UGGACCCUUAUGAUCUGAUC
>U0092
UAAAAUGAUGAAAAAGAGCAUACAGUGUGCAAGGCAAAAAUGAAACCUCAACAUAUAAAAAGGCAAGGUGCAGUGUCCCCGAUGUCUGAUC
>U0094
UACGAUGAUAGAGUUCAAAGCCUAGUAAAACACGAGUAAGUUUACAAUAAUAAACUCGCUAACGAGCCAAUGGUUCUUUUGUGAAAAGCUGUCCAGACCUCUGAGG
>U0098
UCACAUGAUAAAGGUGUCAGAUGGUCAGAGGGCGUAGGCCAGCAAUCCUAAAUUUAGACAAGGAGGAAACCCUGCCCAAGCUGUGCCACCCCCUGAGCUGAUG
>U0100
UCAGGUGAUGUAGCGUCUUCAGGCCCAACGACGCUCGUCACUGUCUCAUCGUGCUUUGUGACUUAACUCGUGUGGCCCUUGGCUUGUCCUUGCCACAGAAUGCACUCU
GAAGUCGGAUGAGCGGAAACUGAUG
>U0101
UCUAAUGAUGAAAACUUGGUCUUCUGGAGUUGAAAUAAGGAGGUCUUCUGCUUGCUACAUGGGUAGAAGGAAGGAAGCUC AACACAUGUAGUAAAUUUAAAGCGUG
GCUAGAUGGUGCAUCUGAUU
>U0104
UCUGAUGAUUUACAGUGGCAUUUGCCUCCUUGUCAGUUGAAACAGACACAAGUUAAGACGGAGUAAAUCUCUAAUUUAAAAACUGAUC
>U0105
UGAGAUGAUCCUUUGCGAUGACCCAGCGCGGGAUGUUCUUGUUCGUCUUGUCUUCUUCGCGAGCCUCAACUCCUACUGAGC
>U0106
UGAU AUGAUUGGCAAAACGGAAGUUUGCACACAGUGUAAAACCGAAACAGACCACCUUAUUGAUGGAGAAUGUGUACCAGCAGGGACUGACC
>U0107
UGCAGUGAUGUGUGGAGUGCUCGCGGGUCGAUGGGUCCACAGUCUGUGAGGAAAAGGUGUUACUUUUUCUCUGUAAAUAAAAAAGCCAUGUCUGCUGACC
>U0108

UGCCAUGAUGACCCUUCGAGGCUUCUCUGGAGCACUGCUCGACUCUCUGCCACCUACGUGCUGGACACCUACAGGAUCAGUGAGACGCACGAGCUGAAG
 >U0109
 UGCCAUGAUUUUUAUUUCCCGGCGUCCUGAUCGUCAGAUAUUUUGAUUUUUUGAUUUUCGGAAGGCAACAUUCGCACGGGGUCGCAGCGACUGAUU
 >U0114
 UGGAAUGAUGCAGUAGCAAGCUAAUCUCUGUCCGUGGGCUCUGGGAGUUUACUUACCUGUUGCGUAAAAACUCUGAGC
 >U0115
 UUAGAUGAUAGCUAUGUUCAGCAGGUUCAGACUUUAGGGCCGAAAUUGUUACUACACUUAAGAAGCGAUCUUGGGGUUCUAGCUGAAAUAUUCUAAGCAUUGUAUCCA
 UGUCUUGGAGCUGAGA
 >U0121
 UUGUAUGAUUGCUUUUGUUCUUUCGUUCCCCGUGCUCCAUGCCUCUGCACCUUGAGGCCACUAGUCAGGUACUUCAUUGGCUGACA
 >U0122
 UUUCAUGAUGGCCGUCUCGCCGCUACAGGACAGGCGACUGCCUCGUUUGGCCUCCA AUUGAUGAAAUAUUCAUUGCAAUUUUCUUUAAAAAAUAAAUACGAAUGCUGAA
 A
 >U0123
 UUUGAUGAUGAGCGUGUGGCUAUAGGGGAUCUUGUAUGGCAGAAUUGUGGUCUCACCUUGCAGUUUUUCUUCGUUCACAGUGUGCGCCGCUCUUCGCUUCUGAAG

4. Giardia rRNA sequences

>'Giardia_5.8S_rRNA'
 AACGCCCCGCGGCGGATGCCCTCGGCCCGGGCGGCGACGAAGAGCGGGCGGAGCGCGAGACGCGGTGCGGACCCGCC
 GCCCGAGAAGCACCGACCCTCGAACGCAGCGGCCCGGGCGCCGCCCTCGGCGCCC
 >'Giardia_LSU_rRNA'
 GCGCGCCCCGAGGCGGGCGGGGCGACGGGCGGAACTTAAGCATATCAGTACGCCCCGAGGAGAAAACCAACCGGGATT
 CCCGTAGCGGCGAGCGACGCGGGAGGAGCCCGCCCCGAAGGCGCGCTGTGGGGCGCAGGCGCAGGCCCGCCGCGAGGGG
 GCCGAGGGCCCCGCCGAGAGGGTGCAAGCCCCGTACGGCGGCCCGGGCCTGCGCGGCGAGTAGCGCTGCTTGAGC
 GTGCAGCGCGAAGGGAGGCGCGCCCTTCCAAGGCTAAATACGCCCGGGACCGATAGCGGACCAAGTAGCGCGAGCGA
 ACGGTGAAAAGGACGCCCTGCGGCCGCTCAAAAGACCTGAACCCGGCCGGCCCGCGCCCGCCGGCCCCGCTCTCGAAAC
 ACGGACCGAGGAGCCACGCGCCGCGGCGAGCCCGAGGGAGCCCCGCGGCGGAGCGAGCGGAGACGCCCGGGCCCGC
 CGCGCCCTGCGGGCGTGCGCGKCCGAGCCCGGCGCGTGGGCCCGAAAGCGGTGATCTATGCCCGGCGAGGGCGAG
 GCCGGGCGAAAGCCTGGTGGAGGCCCGCGCGGTGCTGACGCGCAGATCGCTCGTCCGAGCCGGGCATGGGGCGAAAG
 ACTCATCGAACCCTGGTAGCTGGTTGCCCTCCGAAATGCTCCCAGGACAGCCCGCCCGCCCGAGTTGCGGCCCGTAG
 AGCGCTGGCCGGGCGGAGCGGGGCGCTGCCCTCGCCCCCCCCAAACTCCGAAGGGCCGCGCCCGCCCGCGCTGG

CCTGGGCGGGGCGGGCGAATGCGGGCGGGCGTGGGCCCTCTGGTAAGCAGGACGGGCGAGGCGGGACGATCCGGAC
 GCCGGGCGAGGTGCGCCGCGGGGCCCGGGAACGGCGTCCGCGGTCGCCGACAGCTGGAAGGTGGCCCCAGAAGTCG
 GCATCCTCCAGGAGTGTGTAACAACCCACCAGCCGAATCGGCCGGCCCCGAAAATGGAGCGCGCCGGAGCCCCGGACC
 CGCGCCCCGGCCCGCGCGCGGGTAGGAGCCCGCAGAGGCCCGGGGGCGAAGGCGCGCGCAGGCCCGCCGGAC
 CGGCCTCTGGTGACAGATCTCGGCAGCAGTAGCCGCTACTCCGCGCCCCGGAGGACTGAGGGGGAGACGGGTTCGCGGGC
 GCCTGCATCTGGCCGCGGGTACTCGGGCCTAAGCGGCGGGTGAAGACCGGGAAGGGGCGTGCCCGCCGTCGAACGGG
 GAGCCGGCGGAGACTCCGGCAGGCGCGGCCCCCGGAGACGCCCGCCCCGGCGACGCGCACGGGGACCGCGCGGG
 CGGCGCCCCGGCCCGCAACGCCCGCAGCCCCGGACGCTTGCGCGGAGAGGGGGGCCCGGGGGCGACCCCGCGCG
 TCCCCGGCCGCCCTGAAAAGCCGGGGGGCGCCGGCCGCGCGCGTACCGACCGCAGCAGGACTCCGGGGTCAGCAGCC
 TCTAGCGCGGGAGCGAACCGCGCTCAGGGAAGTCGGCAAGCCGGCTCCGTAACCTCGGGAAAAGGAGTGGCTCTGACGG
 CGCGCCGGGTACAGAACTGGAACGCGACGCGGGATCCCGACTGTTTACTAGAAACACAGCGTCGCGAGGGCCCGACCCGG
 CGTGCGCGACGTGATTTCTGCCAGTGCCACGACCGTCACCGTGAAGCGATCCGCCGAAGCCCTGGTAAACGGCGGG
 AGTAACTAGACTCTCTTAAGGTAGCCAAATGCTCGTCCGGCAATTTCCGACGTGCATGAATGGACCAACGAGGATCC
 CACTGTCCCGAGCCGCGCCTCCGCGAGCCTCAGCCTCGGGAACCGGGCAGGGCCCGCCAGCGGGCAAGAAGACCCTT
 TTGAGCTTACTCCAGCCCGGGCCTGTGGGGCGGGCGGCCGCGCAGCGCACAGGGGAGGCCGCGCCCCCTGAGACACC
 CTGACGGCCCGCCCGCCCGCTCACCCGGTCGCGCGGGGACCCCGCCGGGCGGGGAGTTCCGGCTGGGGCGGCGCGCT
 GCTACACCGGACCGCAGGCGTCCACGGCGGGCTCAGCGAGGACGGAGACCTCCCGCGGAGCAGAAGGGCACAAGCCCG
 CCCGACCCGCGCCCCCGTGCCGGCGCGGGCCGCGAAAGCGGGGCTACCGATCCTTCGCCGCCCGGCcGCgGGCGCG
 GAGGTGGCAGAAAAGTTACCACAGGGATAACTGGCTTGTGGCCCGGAGCGCCCGCAGCGACGCGGCYTTTGTATCCTT
 CGATGTGGCTCTTCTACCGTCCGCGCGCACCGGCGCGGAAGCGTCGGATTGTTACCCGTTCAAGGGATCGTGAGCT
 GGGTTTAGACCGTCTGTAGACAGGTTAGTTTTACCTACTGGCCCCGGGGCCAGAGCACGGCGGGYAGTACGAGAGGA
 ACGCCCGCGGGCCCGCAGCCCGCGGTTGCCCGCCGGGCAGYGCCGYGCCCGCGCCCGGGGGYCTTGCGCTGA
 CCGMCTTAAGCGCGACCCCGCTCGCGCCCCCGCCGGCCGCGCGCCCCAGCCCCGTGCCCGTCCCGGAGCGGCCCC
 CGCCCGGGGAGACCACCCGGCGGGCGCTCCTGTACGGCGCAGAGCCCTGCGATCGCTGAGGGACGCGCCTGCAGAGC
 GCGGGGCGGGGCGCGGCCACTTGCTCTGGGGGGTGGCGGGCAGACAGACAGGCAGAGCGCGAAAAGAGAAGATTG
 AGGGAGTGCAGGGTGCCTCAAGGGTGGCCAGGGGGCAGTGACAGCCACCACCGGGTCTGCCTTGCACAGAGgAGaCR
 CCCGTGTGCGCAGGGGGCGGCGCAGGACCGCAGGGGGCCCCGGGGAGGCGGCCCGGGGA

>'Giardia_SSU_rRNA'

CATCCGGTTCGATCCTGCGGAGCGGACGCTCTCCCAAGGACGAAGCCATGCATGCCCGTCAACCGGGACGCGGCGG
 ACGGCTCAGGACAACGGTTGCACCCCCGCGGGTCCCTGCTAGCCGGACACCGCTGGCAACCGGCGCAAGACGTG
 CGCGCAAGGGCGGGCGCCCGGGCGAGCAGCGTACGCGAGCGACGGCCCCCGGGCTTCCGGGCATCACCCGGTCCG
 GCGCGGTTCGCGCGCGCCGAGGGCCCGACGCTGGCGGAGAATCAGGGTTCGACTCCGGAGAGCGGGCTTGCAGACGG
 CCCGCACATCCAAGGACGGCAGCAGGCGCGGAACCTTGCCCAATGCGCGGGCGCGGAGGCAGCGACGGGGAGCGCGGAG
 CGAGGCGGGCCACAGCCCCCGCGGAGCCGAGGGCAAGGTCTGGTGCAGCAGCCGCGTAATTCAGCTCGGCGA
 CCGTCCGCGGGCGTGTGCAGTTGAAACGCCGTAGTTGGCCCCCGCCGCCACGAGGAAACGGGAGCGCTCCAGGCA

GGCCCGTTGGACCCGCGCGTGGGACCGCGCAGCGGGCGCGCGCGCCGCGGACCCCGAGGAGAGCGGGCGGGGGCA
CCGGTACCGGCCGGGGACGGGTGAAACAGGATGATCCCGCCGAGACCGCCGGCCGCGCAGGCGCCTGCCAAGACCGCCT
CTGTCAATCAAGGGCGAAGGCCGGGGGCTAGAAGGCGATCAGACACCACCGTATTCCCGGCCGTAACGGTGCCGCCCC
GCGGCCGGCGCGCGTCCCGCCGGCCGCCAGGAAACCGGGAGGCTCCGGGCTCTGGGGGAGTATGGCCGCAAGGC
TGAAACTTGAAGGCATTGACGGAGGGGTACCACCAGACGTGGAGTCTGCGGCTCAATCTGACTCAACGCGCGCACCTCA
CCAGGCCCGGACGCGCGGAGGACCGACAGCCGGGCGCGCTTTCGCGATCGCGCGGGCGGTGGTGCATGGCCGCTCCAG
CCCGTGGCGCGAGCCGTCTGCTCCATTGCGACAACGAGCGAGACCCCGGCCGCGGGCGCCGCGGGACGGCCCGCGAG
CGGGAGGACGGCGGGCGATAGCAGGTCTGTGATGCCCTCAGACGCCCTGGGCCGACGCGCGCTACACTGGCGGGGCC
AGCCGGCGCCCGGAGGACGCGCGGAGCCCCCGCGTGGCCGGGACCGCGGGCTGGAACGCCCCCGCGCACCAGGAATG
TCTTGTAGGCGCCCGCCCCACCGCGCGCCGACGCGTCCCTGCCCTTGTACACACCGCCCGTCCGCTCCTACCGACTG
GGCGGGCGCGAGCGCCCCGACGCGCAAGGGCCGCGAGCCCCCGCGCTGGAGGAAGGAGAAGTCGTAACAAGGTA
TCCGTAGGTGAACCTGCGGATGGATCC

5. Alignment of LSU rRNA with methylation sites highlighted

```
Human_LSU_rRNA      CGCGACCTCAGATCAGACGTGGCGACCCGCTGAATTTAAGCATATTAGTCAGCGGAGGAA
Yeast_LSU_rRNA      -----
Giardia_LSU_rRNA    -----
```

```
Human_LSU_rRNA      AAGAAACTAACCAGGATTCCCTCAGTAACGGCGAGTGAACAGGGAAGAGCCCAGCGCCGA
Yeast_LSU_rRNA      -----
Giardia_LSU_rRNA    -----
```

```
Human_LSU_rRNA      ATCCCCGCCCCGCGGGGCGCGGGACATGTGGCGTACGGAAGACCCGCTCCCCGGCGCCGC
Yeast_LSU_rRNA      -----
Giardia_LSU_rRNA    -----
```

```
Human_LSU_rRNA      TCGTGGGGGGCCCAAGTCCTTCTGATCGAGGCCAGCCCGTGGACGGTGTGAGGCCGGTA
Yeast_LSU_rRNA      -----
Giardia_LSU_rRNA    -----
```

```
Human_LSU_rRNA      GCGGCCGGCGCGCGCCCGGGTCTTCCCGGAGTCGGGTTGCTTGGGAATGCAGCCCAAAGC
Yeast_LSU_rRNA      -----
Giardia_LSU_rRNA    -----
```

Human_LSU_rRNA GGGTGGTAAACTCCATCTAAGGCTAAATACCGGCACGAGACCGATAGTCAACAAGTACCG
 Yeast_LSU_rRNA -----
 Giardia_LSU_rRNA -----

Human_LSU_rRNA TAAGGGAAAGTTGAAAAGAACTTTGAAGAGAGAGTTCAaGaGGGCGTGAAACCGTTAAGA
 Yeast_LSU_rRNA -----
 Giardia_LSU_rRNA -----

Human_LSU_rRNA GGTAACGGGTGGGGTCCGCGCAGTCCGCCCGGAGGATTCAACCCGGCGGGCGGGTCCGGC
 Yeast_LSU_rRNA -----
 Giardia_LSU_rRNA -----

Human_LSU_rRNA CGTGTCGGCGGCCCGGCGGATCTTTCCCGCCCCCGTTCTCCCGACCCCTCCACCCGCC
 Yeast_LSU_rRNA -----
 Giardia_LSU_rRNA -----

Human_LSU_rRNA CTCCTTCCCCCGCCGCCCTCCTCCTCCTCCCCGGAGGGGGCGGGCTCCGGCGGGTGCG


```

Yeast_LSU_rRNA -----
Giardia_LSU_rRNA -----

Human_LSU_rRNA GGGGTGGGCGGGCGGGGCCGGGGTGGGGTCGGCGGGGACCGTCCCCGACCGGCGACC
Yeast_LSU_rRNA -----GTTTGACCTCAAATCAGGTAGGAGTACCCGCTGAACTTAAGCAT
Giardia_LSU_rRNA -----

Human_LSU_rRNA GGCCGCCGCCGGGCGCATTTCACCGCGGCGGTGCGCCGCGACCGGCTCCGGGACGGCTG
Yeast_LSU_rRNA ATCAATAAGCGGAGGAAAAGAAACCA-----ACCGGATTGCCTTAGTAACGGCGA
Giardia_LSU_rRNA -----

Human_LSU_rRNA GGAAGGCCCGGCGGGGAAGGTGGCTCGGGGGCCCCGTCCGTCCGTCCGTCCCTCCTCCTC
Yeast_LSU_rRNA GTGAAGC--GGCAA--AAGCTCAAATTTGAAATCTGGTACCTTCGGT-GCCCGAGTTGTA
Giardia_LSU_rRNA -----

Human_LSU_rRNA CCCCCTCTCCGCC--CCCCGGCCCCGCGTCCTCCCTCGGGAGGGCGCGCGGGTCCGGGGCG
Yeast_LSU_rRNA ATTTGGAGAGGGCAACTTTGGGGCCGTTCTTGTCT----ATGTTCTTGGAACAGGACG
Giardia_LSU_rRNA -----

Human_LSU_rRNA GCGGCGGCGGCGGCGGTGGCG-GCGGCGGCGGGGGCGGCGGGACCGAAACCcCCCCGAG
Yeast_LSU_rRNA TCATAGAGGGTGAGAATCCCCTGTGGCGAGGAGTGCGGTTCTTTGTAAAGTGCCTTCGAA

```

```

Giardia_LSU_rRNA      -----GCGCGGCCCGAG
                                                                ***

Human_LSU_rRNA        TGTTACAGCCCCCGGCA-GCAGCACTCGCCGAATCCCGGGGCCGAGGGAGCGAGACCC
Yeast_LSU_rRNA        GAGTCGAGTTGTTTGGGAATGCAGCTCTAAGTGGGTGGTAAATTCCATCTAAAGCTAAAT
Giardia_LSU_rRNA      GCGGCGGGGGcGACGGGCG---GAACTTAAGCATATCAGTACGCCCCGAGGAGAAACCA
                        *      **      *      *      *      *      *
Human_LSU_rRNA        GTCGCCGCGCTCTCCCCCTCCCGGCGCCACCCCGCGGGGAATCCCCGCGAGGGGGG
Yeast_LSU_rRNA        ATTGGCGAGAGA--CCGATAGCGAACAAGTACAGTGATGGAAAGATG-----AAAAGAAC
Giardia_LSU_rRNA      A---CCGGGATT--CCCCgTAGCgGCGAGCGACGCGG-GAGGAGCCCCCGCCCGAAGGCGC
                        ** *      **      *      *      *      *
Human_LSU_rRNA        TCTCCCCCGCGGGGGCGCGCCGGCGTCTCCTCGTGGGGGGCCGGGCCACCCCTCCACG
Yeast_LSU_rRNA        TTTGAAAAGAGAGTGAAAAAGTACGTGAAATTGTTGAAAGGGAAGGCATTTGATCA--G
Giardia_LSU_rRNA      GCTGTGGGGCGCAGGCGCAGGCCCGC-----CGCGAGGGGgCCCGAGGGCCCCGCCCGAG
                        *      * *      *      **      *      **      *      *
Human_LSU_rRNA        GCGCGACCGCTCTCCACCCCTCCTCCCCGCGCCCCCGCCCCGGCGACGGGGGGGGTGCC
Yeast_LSU_rRNA        ACATGGT-GTTTTGTGCCCTCTGCTCCTTGTGGGTAGG---GGAATCTCGCATTTCACT
Giardia_LSU_rRNA      AGGGTGCAAGCCCCGTACGGCGGC-CGCCGGGcCTGCG---CGGCGAGTAGCGCTGC-TT
                        *      * *      * *      *      *      **      *
Human_LSU_rRNA        GCGCGCGGGTCGGGGGGCGGGGCGGACTGTCCCCAGTGCGCCCCGGGCGGGTTCGCGCCGT
Yeast_LSU_rRNA        GGGCCAGCATCAGTTTTGGTGGCAGGATAAATCCATAGGAAT----GTAGCTTGCCCTCGG
Giardia_LSU_rRNA      GAGCGTGcAGCGCGAAGGGAGGCGCGGCCCTTCCAAGGCTAA-----ATACGCCCG-

```

```

* ** * * * * * ** *
Human_LSU_rRNA CGGGCCCGGGGAGGTTCTCTCGGGGCCACGCGCGCGTCCCCGAAGAGGGGGACGGCGG
Yeast_LSU_rRNA TAAGTATTATAG-----CCTGTGGGAATAC-----TGCCAGCTGGGACTGAGGACTGCGA
Giardia_LSU_rRNA --GGACCGATAG-----CGGACCAAGTAGCGCGAGCGAACGGTGAaAAGGACGCCCTGCG
* * * * * * * * *
Human_LSU_rRNA AGCGAGCGCACGGGGTCGGCGGCGACGTCGGCTA-CCCACCCGACCCGTCtTGAAACACG
Yeast_LSU_rRNA CGTAAGTCAAGGATGCTGGC---ATAATGGTTA-TATGCC--GCCCGTCTTGAAACaCG
Giardia_LSU_rRNA GCCGCTCAAAGACCTGAACCCGGCCGGCCGGCCCGCCGGCCCGTCTCGAAACACG
* * * * * * * * * * * * * * * *
Human_LSU_rRNA gACCAAGGAGTCTAaCACGTGCGGAGTCGGGGGCTCGCACGAAAGCCGCCGTGGCGCAA
Yeast_LSU_rRNA GACCAAGGAGTcTAACGTCTATGCGAGT-----GTTTGGGTGTAAAACCCATACGCGTAA
Giardia_LSU_rRNA GACCGAGgaGCCACGCGCCGCGGCGAGCC-----CGAGGGAGCCCCGCGGCGGAG
**** * * * * * * * * * * * * * * *
Human_LSU_rRNA TGAAGGTGAAGGCCGGCGCGCTCGCCGGCCGAGGTGGGATCCCGAGGCCTCTCCAGTCCG
Yeast_LSU_rRNA TGAAAGTGAACGTAGGT-----TGGGGCCTCGCAAG----
Giardia_LSU_rRNA CGAGCGCGA-----
** * **
Human_LSU_rRNA CCGAGGGCGCACCAACGGCCCGTCTCGCCCGCCGCGCCGGGGAGGTGGAGCACGAGCGCA
Yeast_LSU_rRNA ---AGGTGCACAATCGACCGATCCTGATGTCTTCG--GATGGATTTGAGTAAGAGCATA
Giardia_LSU_rRNA -----GACGCCCG--GGCCCGCCGCGCCC-CTGCG--GGCGTGCGCGGKC-CGAGCCGC
* ** * * * * * * * * * * * *

```

```

Human_LSU_rRNA      CGTGTTAGGACCCGAaAgATGGTGAACtATGCCTGGGCAGGGCGAAGCCAGAGGAAACTC
Yeast_LSU_rRNA      GCTGTTGGGACCCgAaAGATGGTGAaCTATGCCTGAATAGGGTGAAGCCAGAGGAAACTC
Giardia_LSU_rRNA    GGCGCGTGGGCCCCGAAAGGCGGTGATCTATGCCCGGCGAGGGCGAGGCCGGGCGAAAGCC
                    *   ** ***** ***** ***** *   **** ** ** *   **** *

Human_LSU_rRNA      TGGTGGAGGTCCGTAGCGGTCCTGACGTGCAAATCGGTCGTCCGACCTGGGTATAGGGgC
Yeast_LSU_rRNA      TGGTGGAGGCTCGTAgCGGTTCTGaCGTGCAAATCGATCGTCGAATtTGTTATAGgGC
Giardia_LSU_rRNA    TGGTGGAGGCCCGCCGCGGTGCTGACGCGCAGATCGCTCGTCGGAGCCGGGcATGGGGGC
                    ***** ** ***** ***** ** ***** ***** *   *** ** *****

Human_LSU_rRNA      GAAAGACTAATCGAACCATCTAGTAGCTGGTTCCTCCGAAGTTTCCCTCAGGATAGCTG
Yeast_LSU_rRNA      GAAAGACTAATCGAACCATCTAGTAGCTGGTTCCTGCCGAAGTTTCCCTCAGGATAGCAG
Giardia_LSU_rRNA    GAAAGACTcATCGAACCGCCTGGTAGCTGGtTGCTCCGAAATGTCTCCAGGACAGCCG
                    ***** ***** ** ***** ***** *   ***** * ** * ***** ** *

Human_LSU_rRNA      GCGCTCTCGCAGACCCGACGCACCCCGCCACGCAGTTTTTATCCGGTAAAGCGAATGATT
Yeast_LSU_rRNA      AAGCTCGTAT-----CAGTTTTATGAGGTAAAGCGAATGATT
Giardia_LSU_rRNA    CCGCCCC-----GCAGTTGCGGCCCGTAGAGCGCTGGCCG
                    ** *                               *****           *** **** *

Human_LSU_rRNA      AGAGGTCTTGGGGCCGAAACGATCTCAACCTATTCTCAAACTTTAAATGGGTAAAGAAGCC
Yeast_LSU_rRNA      AGAGGTTCGGGGTTCGAAATGACCTTGACCTATTCTCAAACTTTAAATATGTAAGAAGTC
Giardia_LSU_rRNA    GCGGGAGCGGGGGGCCTGCC--CCTCGCCCCCCCCCAAACCTCCGAAGGGcCGCGCCGCC
                    **   **** *           ** ** * ***** **           * * *

```

Human_LSU_rRNA CGGCTCGCTGGCGTGGAGCCGGGCGTG-GAATGC-GAGTGCCTAGTGGGCCACTTTTGGT
 Yeast_LSU_rRNA CTTGTTACTTAATTGAACGTGGACATTTGAATGAAGAGCTTTTAGTGGGCCATTTTGGT
 Giardia_LSU_rRNA CCGC-CGCTGgCCTGGGCG-GGGCGGGCGAATGC-GGGCGGCGCGTGGGCCCCtCCTGGT
 * ** ** ** * * ** * ** *

Human_LSU_rRNA AAGCAgAACTGGCGCTGCGGGATGAACCGAACGCCGGGTTAAGGCGCCCGATGCCGACGC
 Yeast_LSU_rRNA AAGCaGAACGGCGATGCGGGATGAACCGAACGTAGAGTTAAGGTGCCGAATAC-ACGC
 Giardia_LSU_rRNA AAGCAGGACGGGCGAGGCGGGACGAtCCGGACGCCGGGCCAGGGTGC-----GC
 ***** ** ***** ***** ** ** * * * * ** *

Human_LSU_rRNA TCATCAGACCCCAGAAAAGGTGTTGGTTGATATAGACAGCAGGACGGTGGCCATGGAAGT
 Yeast_LSU_rRNA TCATCAGACACCACAAAAGGTGTTAGTTTCATCTAGACAGCCGGACGGTGGCCATGGAAGT
 Giardia_LSU_rRNA CGCCGGGGCCCGCGGAACGGCGTCCGGCCGGTcCCGACAGCTGGAAGGTGGCCCCaGAAGT
 * * * ** ** * * * ***** ** ***** *****

Human_LSU_rRNA CGGAATCCGCTAAGGAGTGTGTAACAACCTCACCTGCCGAATCAACTAGCCCTGAAAATGG
 Yeast_LSU_rRNA CGGAATCCGCTAAGGAGTGTGTAACAACCTCACCGGCCGAATGAACTAGCCCTGAAAATGG
 Giardia_LSU_rRNA CGGCATCCTCCAGGGAGTGTGTAACAACCCACCAGCCGAATCGGCCGGCCCGGAAAATGG
 *** ***** * * ***** ***** ***** * ***** *****

Human_LSU_rRNA ATGGCGCTGGAGCGTCTGGGCCATACCCGGCCGTCGCCGGCAGTCGAGAGTGGACGGGAG
 Yeast_LSU_rRNA ATGGCGCTCAAGCGTGTACCTATACTCTACCGTC-----AGGGT-----
 Giardia_LSU_rRNA AGCGCGCCGGAGCCcCGGACCCGCGCCCGGCCGCGC-----
 * **** ** ** * * ** *

Human_LSU_rRNA CGGCGGGGGCGGCGCGCGCGCGCGTGTGGTGTGCGTCGGAGGGCGGCGGCGGCGGC

```

Yeast_LSU_rRNA      -----TGATATG-----
Giardia_LSU_rRNA    -----GCG-----
                    *

Human_LSU_rRNA      GGCGGCGGGGGTGTGGGGTCCTTCCCCGCCCCCCCCCCCCACGCCTCCTCCCCTCCTCCC
Yeast_LSU_rRNA      -----
Giardia_LSU_rRNA    -----

Human_LSU_rRNA      GCCCACGCCCCGCTCCCCGCCCCCGGAGCCCCGCGGACGCTACGCCGCGACGAGTAGGAG
Yeast_LSU_rRNA      -----ATGCCCTGACGAGTAGGCA
Giardia_LSU_rRNA    -----CGCGGCGGGTAGG-A
                    * * ** *****

Human_LSU_rRNA      GGCCgCTGCGGTG--AGCCTTGAAGCCTAGGGCGCGGGCCCGGGTGGAGCCCGCCGAGGT
Yeast_LSU_rRNA      GGC-GTGGAGGTC--AGTGACGAAGCCTAGACCGTAAGGTCGGGTCGAACGGCCTCTAGT
Giardia_LSU_rRNA    GGCCGCAGAGGCCCCCGGGGGCGAAGGC-GGCGCGCAGGCCcGCCGGACCGgCCTCTGGT
                    *** * * ** * ***** * * ** * * * ** * ** * **

Human_LSU_rRNA      GCAGATcTTGGTGGTAGTAgcAAATATTCAAACGAGAACTTTGAAGGCCGAAGTGGAGAA
Yeast_LSU_rRNA      GCAGATcTTGGTGGTAGTAgcAAATATTCAAATGAGAACTTTGAAGACTGAAGTGGGGAA
Giardia_LSU_rRNA    GCAGATCTCGGCAGCAGTAGCCGCTACTC---CGCGC-CCCGGAGGACTGAGGGGGAGAC
                    ***** ** * ***** ** ** * * * ** * * ** *

Human_LSU_rRNA      GGGTTCCATGtGAACAGcAgTTGAACATGGGTTCAGTCGGTCCTGAGAGATGGGCGAGCGC
Yeast_LSU_rRNA      AGGTTCCACGTCAACAGCAGTTGGACGTGGGTTAGTCGATCCTAAGAGATGGGGAAGCTC

```

```

Giardia_LSU_rRNA      GGGTTCCGCGGGCGCCTGcATCTGGCCGCGGGTGACTCGGGCCTAAGCGGCGGGTGAAGAC
***** * * *** ** * ***** * *** ** * ** * *
Human_LSU_rRNA        CGTTCCGAAGGGACGGGCGATGGCCTCCGTTGCCCTCGGCCGATCGAAAGGGAGTCGGGT
Yeast_LSU_rRNA        CGTTTCAAAGGCCTGA-----TTTTATGCAGGCCACC-ATCGAAAGGGAATCCGGT
Giardia_LSU_rRNA      CGG---GAAGGGGcG-----TGCCCGCCCGTCTGAACGGGGAGCCGGC
**          *** * * * * * ** * * * * * * *
Human_LSU_rRNA        TCAGATCCCCGAATCCGGAGTGGCGGAGATGGGCGCCGCGAGGCGTCCAGTGCGGTAACG
Yeast_LSU_rRNA        TAAGATTCCGGAACCTGGAT-----ATGGATTCTTCA-----CGGTAACG
Giardia_LSU_rRNA      GGAGACTCCGGCAGGCG-----CGGCCCCCGCG-----GAGACG
*** ** * * * * * * * * * * * * * * *
Human_LSU_rRNA        GCACCGATCCCGGAGAAGCCGGCGGGAGCCCCGGGGAGAGTTCTCTTTTCTTTGTGAAGG
Yeast_LSU_rRNA        TAACTGAATGTGGAGACGTCGGCGCGAGCCCTGGGAGGAGTTATCTTTTCTTCTT-AACA
Giardia_LSU_rRNA      CC-CGCCCCCGGCGACGCGCACGGGGACCGCGCGGGCGGCGCCCCGGCC-CGCGAACG
*          ** ** * * * * * * * * * * * * * *
Human_LSU_rRNA        GCAGGGCGCCCTGGAATGGGTTCCGCCCCGAGAGAGGGGCCCGTGCCTTGGAAGCGTCGC
Yeast_LSU_rRNA        GCTTATCACCCCGGAATTGGTTTATCCGGAGATGGGGTCTTATGGCTGGAAGAGGCCAGC
Giardia_LSU_rRNA      CCCCAGCCcCCGGAC-GCCTTGCGCGGAGAGGGGgCCCGGG-----
*          *** * * * * * * * * * * * * *
Human_LSU_rRNA        GGTTCGGCGGGCGTCCGGTGAGCTCTCGCTGGCCCTTGAAAATCCGGGGGAGAGGGTGTA
Yeast_LSU_rRNA        ACCTTTGCTGGC-TCCGGTGCCTTGTGACGGCCCCGTGAAAATCCACAGGAAGGAAT--A
Giardia_LSU_rRNA      -----GGCGGA-CCCCGCGCTCCCCGGCCGCCCTGAAAAGCCGGGGGGCGCCG---

```

```

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

Human_LSU_rRNA      AATCTCGCGCCGGGCCGTACCCaTaTCCGCAGCAGGTCTcCAAGGTGAACaGCCTCTGGc
Yeast_LSU_rRNA      GTTTTTCATGCCAGGTCTGACTGATAACCGCAGCAGGTCTCCAAGGTGAACAGCCTCTAGT
Giardia_LSU_rRNA    ----CCGCGC---GCCGTACCGA---CCGCAGCAGGACTCCGGGGTCAGCAGCCTCTAGC
                    *   **   *   *****   *   ***** *****   **   *   *****   *

Human_LSU_rRNA      ATGTTGGAACAAtGTAGGTAAGGGAAGTCGGCAAGCcGGATCCGTAACCTTCgGGATAAAGG
Yeast_LSU_rRNA      -TGATAGAATAAtGTAGATAAGGGAAGTCGGCAAAATAGATCCGTAACCTTCGGGATAAAGG
Giardia_LSU_rRNA    GCGGGAGCG-AACGCGGCTCAGGGAAGTCGGCAAGCCGGCTCCGTAACCTCGGGAAAAGG
                    *   *   ** *   * *   ***** *****   *   ***** ***** *****

Human_LSU_rRNA      ATTGGCTCTAAGGGCTGGGTCGGTCGGGCTGGGGCGCGAAGCGGGGCTGGGCGCGCGCCG
Yeast_LSU_rRNA      ATTGGCTCTAAGGGTCGGGT-----AGTGAGG-----GCCT
Giardia_LSU_rRNA    AGTGGCTCtGA-----
                    *   *****   *

Human_LSU_rRNA      CGGCTGGACGAGGCGCGCGCCCCCCCCACGCCCGGGGCACCCCCCTCGCGGCCCTCCCCC
Yeast_LSU_rRNA      TGGTCAGACGCAGCGGGCGT-----
Giardia_LSU_rRNA    -----CGGCGcgC-----
                    ***   **

Human_LSU_rRNA      GCCCCACCCGCGCGCGCCGCTCGCTCCCTCCCCACCCCGCGCCCTCTCTCTCTCTCTC
Yeast_LSU_rRNA      -----GCTTGT-----
Giardia_LSU_rRNA    -----

```


Human_LSU_rRNA CCCCCTCCCCGTCTCCCCCTCCCCGGGGGAGCGCCGCGTGGGGGCGCGGCGGGGGGA
 Yeast_LSU_rRNA -----
 Giardia_LSU_rRNA -----

Human_LSU_rRNA GAAGGGTCGGGGCGGCAGGGGCCGCGGGCGGCCGCGGGGCGGCCGGGCGGGGCAGGTC
 Yeast_LSU_rRNA -----GGACTGCTTGGTGG-----GGCTTGC-----
 Giardia_LSU_rRNA -----

Human_LSU_rRNA CCCGCGAGGGGGCCCCGGGGACCCGGGGGGCCGGCGGGCGCGGACTCTGGACGCGAG
 Yeast_LSU_rRNA TCTGCTAGG-----CGGACTACTTGCGTG--
 Giardia_LSU_rRNA -----

Human_LSU_rRNA CCGGGCCCTTCCCCTGGATCGCCCCAGCTGCGGGCGGGCGTTCGCGGCCGCCCCGGGGAGC
 Yeast_LSU_rRNA -----CCTTGTTGTAGACGGCCTTGGTAG-----
 Giardia_LSU_rRNA -----

Human_LSU_rRNA CCGGCGGGCGGGCGGGCGGCCCCCACCCCCACCCACGTCCTCGGTCGCGCGCGGTCCG
 Yeast_LSU_rRNA -----GTCTC-----
 Giardia_LSU_rRNA -----

Human_LSU_rRNA CTGGGGGCGGGAGCGGTTCGGGCGGCGGCGGTTCGGCGGGCGGCGGGGCGGGGCGGTTCGTC
 Yeast_LSU_rRNA -----
 Giardia_LSU_rRNA -----

Human_LSU_rRNA CCCCCGCCCTACCCCCCGGCCCGTCCGCCCGGTTCCCCCTCCTCCTCGGCGCGCG
 Yeast_LSU_rRNA -----
 Giardia_LSU_rRNA -----

Human_LSU_rRNA GCGGCGGCGGCGGCAGGCGGCGGAGGGGCCGCGGGCCGGTCCCCCGCCGGGTCCGCC
 Yeast_LSU_rRNA -----TTGTAGACCG-----
 Giardia_LSU_rRNA -----

Human_LSU_rRNA CCGGGGCCGCGGTTCCGCGCGCGCCTCGCCTCGGCCGGCGCCTAGCAGCCGACTTAGAAC
 Yeast_LSU_rRNA -----TCGCTTG---CTACAATTAACGATCAACTTAGAAC
 Giardia_LSU_rRNA -----CGGGTCAGAAC
 * * *****

Human_LSU_rRNA TGGTGCAGGACCAGGGGAATCCGACTGTTTAATTAACAACAAGCATCGCGAAGGCCCGCGG
 Yeast_LSU_rRNA TGGTACGGACAAGGGGAATCTGACTGTCTAATTAACAACATAGCATTGCGATGGTCAGAAA
 Giardia_LSU_rRNA TGGAACGGACGCGGGGATCCCGACTGTTTACTAGAAACACAGCGTCGCGAGGGC-GCAC
 *** ***** ***** * ***** ** * ***** ** * ***** ** * *

Human_LSU_rRNA CGGGTGTTGACGCGATGTGATTTCTGCCcAGTGCTCTGAATGTCAaAGTGAaGAAATTCA

```

Yeast_LSU_rRNA      GTGATGTTGACGCAATGTGATTTCTGCCcAGTGCTCTGAATGTCAAAGTGAaGAAATTCA
Giardia_LSU_rRNA    CCGGCGCTGGCGCGACGTGATTTCTGCCAGTGCCACGACCGTCACCGTGAAGCGATCCG
                    * * ** *** * ***** ** **** ***** ** *

Human_LSU_rRNA      ATGAAGCGCGGgTAAACGGCGGGAGTAaCTATGACTCTCTTAAGGTAGCCAAaTGCCTCG
Yeast_LSU_rRNA      ACCAAGCGCGGGTAAACGGCGGGAGTAaCTATGACTCTCTTAAGGTAGCCAAaTGCCTCg
Giardia_LSU_rRNA    CCGAAGCCCTGTAACGGCGGGAGTAaCTATGACTCTCTTAAGGTAGCCAAATGCCTCG
                    **** * *****

Human_LSU_rRNA      TCATCTAATTAGTGAcGCGCATGAAtGGATGAaCGAGaTTCCCACTGTcCCTACCTACTA
Yeast_LSU_rRNA      TCATCTAATTAGTGACGCGCATGAATGGATTAACGAGATTCCCACTGTcCCTATCTACTa
Giardia_LSU_rRNA    TCGGGCAATTTCCGACGTGCATGAATGGACCAACGAGGATCCCACTGTCCCGAGCCGCGC
                    ** **** ***** ***** ***** ***** * * *

Human_LSU_rRNA      TCCAGCgAAACCACAGcCAAGGGAACGGGCTTGGcGGAATCAGCGGGGAAAGAAGACCCT
Yeast_LSU_rRNA      TCTAGCGAAACCACAGCCAAGGGAACGGGCTTGGCAGAATCAGCGGGGAAAGAAGACCCT
Giardia_LSU_rRNA    CtCCGCGAGCCTCCAGCCTCGGGAACGGGCGAGGGCCGGCCAGCGGGGCAAGAAGACCCT
                    **** * ***** ***** ** ***** *****

Human_LSU_rRNA      GTTGAGCTTGACTtCTAGTCTGGCACGGTGAAgAGACATGAGAGGTGTAGAATAAGTGGGA
Yeast_LSU_rRNA      GTTGAGCTtGACTtCTAGTTTGACATTGTGAAGAGACATAGAGGGTGTAGAATAAGTGGGA
Giardia_LSU_rRNA    TTTGAGCTTGACTCCAGCCCGGGCTGTGGGGCGGGGCGGCCGGCGCAGCGCACAGGGGA
                    ***** ** * ** * ** * ** *

Human_LSU_rRNA      GGCCCCCGGCGCCCCCGGTGTCCCCGCGAGGGGCCCGGGGCGGGGTCCGCGGCCCTGC
Yeast_LSU_rRNA      G--CTTCGGCGCC-----

```

```

Giardia_LSU_rRNA      GG-----CCGCGCCCC-----
*           * *****

Human_LSU_rRNA        GGGCCGCCgGTGAAATACCAcTACTCTGATCGTTTTTTTCACTGACCCGGTGAGGCGGGGG
Yeast_LSU_rRNA       -----AGTGAAATACCACTACCTTTATAGTTTCTTTACTTATTCAATGAAGCGGAGC
Giardia_LSU_rRNA     -----TGAGACACCCTGACGGCCCGCCCGCCCCGCTCACCCGGTCGCGCGGGG-
                    *** * *** ** * ** * * * * *
                    * * * * *

Human_LSU_rRNA        GGCGAGCCCGAGGGGCTCTCGTTCTGGCGCCAAGCGCCCGCCCGGGCGCGACCCG
Yeast_LSU_rRNA       TGGAATTC----ATTTTCCACGTTCTAGCATTCAAGGTCC---CATTCGGGGCTGATCCG
Giardia_LSU_rRNA     -----aCCCG
                    * ***

Human_LSU_rRNA        CTCCGGGGACAGTGCCAGGTGGGGAGTTTGACTGgGGCGGTACACCTGTCAAACGGTAAC
Yeast_LSU_rRNA       GGTGGAAGACATTGTCAGGTGGGGAGTTTGGCTgGGCGGCACATCTGTTAAACgATAAC
Giardia_LSU_rRNA     C-----CCGGGCGGGgAGTTCGGCTGGGGCGGCGCGCCTGCTACACCGGACC
                    * ** ***** * ***** * *** * ** * *

Human_LSU_rRNA        GCAGGtgTCCTAAGGCGAGCTCAGGGAGGACAGAAACCTCCCGTGGAGCAGAAGGGCAAA
Yeast_LSU_rRNA       GCAGATGTCTAAGGGGGGCTCATGGAGAACAGAAATCTCCAGTAGAACAAAAGGGTAAA
Giardia_LSU_rRNA     GCAGGCGTCCCACGGCGGGCTCAGCGAGGACGGAGACCTCCCGCGGAGCAGAAGGGCACA
                    ***** * ** * ***** ** * ** * * * * * * *

Human_LSU_rRNA        AGCTCGCTTGATCTTGATTTTCAGtACGAATACAGACCGTGAAAGCGGGGCCTCACGATC
Yeast_LSU_rRNA       AGCCCCCTTGATTTTGATtTTCAgTGTGAATACAAACCATGAAAGTGTGGCCTATCGATC
Giardia_LSU_rRNA     aGCCcGCCCGACCCGCGCCCCCGTGCCGGCGCGGGCCCGCAAAGCGGGGCCTACCGATC

```

```

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

Human_LSU_rRNA      CTTCTGACCTTTTGGGTTTTAAGCAGGAgGTGTCAGAAAAGTTACCACAGGGATAACTGG
Yeast_LSU_rRNA      CTTTAGTCCCTCGGAATTTGAGGCTaGAgGGTGCCAGAAAAGTTACCACAGgGGATAACTGG
Giardia_LSU_rRNA    C TTC-GCCGCCCGGCCGCGGGCGCGGAGGTGGCAGaAAAGTTACCACAGGGATAACTGG
*** * * *          ***** *****

Human_LSU_rRNA      CTTGTGGCGGCCAAGCGTTCATAGCGACGTCGCTTTTGGATCCTTCGATGTCGGcTCTTC
Yeast_LSU_rRNA      CTTGTGGCAGTCAAGCGTTCATAGCGACATTGCTTTTGGATTCTTCGATGTCGGCTCTTC
Giardia_LSU_rRNA    CTTGTGGCCGCCGAGCGCCCCGACGCGCGCYTTTGGATCCTTCGATGTCGGCTCTTC
***** * * ***** * ***** ** ***** *****

Human_LSU_rRNA      CTATCATTGTGAAGCAGAATTCGCCAAGCGTTgGATtgTTCACCCACTAATAGGGAACGT
Yeast_LSU_rRNA      CTATCATAACCGAAGCAGAATTCGGTAAGCGTTGGAttGTTACCCACTAATAGGGAACGT
Giardia_LSU_rRNA    CTACCGTCCGCGCGCACCCGGCGCGGAAGCGTCGGATTGTTACCCgTTCA-AGGGATCGT
*** * *          ***          ***** ***** * * ***** ***

Human_LSU_rRNA      GagGCTGGGTTTAGAcCGTCGTGAGACAGGTTAGTTTTACCCTACTGATGaaTGTGTTGTTG
Yeast_LSU_rRNA      gAggCTGGGTTTAGAaCCGTCGTGAGACAGGTTAGTTTTACCCTACTGATGA-ATGTTACCG
Giardia_LSU_rRNA    GAGCTGGGTTTAGACCgTCGTGAGACAGGTTAGTTTTACCCTACTG-----GCCCGGGG
***** ***** *

Human_LSU_rRNA      CCATGGTAaTCCTGCTCAGTACGAGAGGAACCGCAGgttCAgACATTTGGTGTATgTGCT
Yeast_LSU_rRNA      CAATAGTAATTGAACtTAGTACGAGAGGAACAGTTCATTTCGGATAATTGGTTTTTGCgGC
Giardia_LSU_rRNA    CCAGAGCACGGCGGGYcCAGTACGAGAGGAACGCCCGCCGCGGGCCGcCAGC-CCCgCGGT
* * * *          *****          * *          * *

```



```

Human_LSU_rRNA      GTTCGTGGGGAACCTGGCGCTAAACCATTTCGTAGACGACCTGCTTCTGGGTCGGGGTTT-
Yeast_LSU_rRNA     TTTTGCG-----TGGGGATAAATCATTGTATACTAGATGTACAACGGGGTAT-
Giardia_LSU_rRNA   GGGAGTGCAGGGTGCCCCAAGGGGTGGCCAGGGGGCAGTGCACGCCACCACCGGGTCTg
                   * *                               * *                               * * * * *
Human_LSU_rRNA     -CGTACGTAGCAGAGCAGCTCCCTCGCTGCGATCTATTGAAAGTCAGCCCTCGACACAAG
Yeast_LSU_rRNA     -TGTAAGCAGTAGAGTAGCCTTGTTGTTACGATCTGCTGAGATTAAGCCTTTGTTGTCTG
Giardia_LSU_rRNA   CCTTGCACAGAGGAGACRCCCGTGTGCGCAGGGGGGcGGCGCAGGACCGCAGGGGGgCCCG
                   * ** *** * * * * * * * * * *
Human_LSU_rRNA     GGTTTGTC-----
Yeast_LSU_rRNA     A-TTTGT-----
Giardia_LSU_rRNA   GGGGAGGCGGCCCGGGGA
                   *

```

6. Alignment of SSU rRNA with methylation sites highlighted.

```

Human_SSU_rRNA      TACCTGGTTGATCCTGCCAGTAG-CATaTGCTTGTCTCAAAGATTAAGCCATGCATGTCT
Yeast_SSU_rRNA     TATCTGGTTGATCCTGCCAGTAGTCATaTGCTTGTCTCAAAGATTAAGCCATGCATGTCT
Giardia_SSU_rRNA   CATCCGGTCGATCCTGCCGG-AGCGCGACGCTCTCCCCAAGGACGAAGCCATGCATGCC
* * *** ***** * **      * ***      * *** ** ***** *
Human_SSU_rRNA     AAGTACGCACGGCCGGTACAGTGAAACTGCGAATGGCTCaTTAAATCAGTTATGGTtCCT
Yeast_SSU_rRNA     AAGTATAAGCAATTTATACAGTGAAACTGCGAATGGCTCaTTAAATCAGTTATCGTTTAT
Giardia_SSU_rRNA   -----GcTCACCCGG-GACGCGGCGGACGGCTCAGGACAACGGTTGCACCCCC
* * **      * *** * *****      * * * ***
Human_SSU_rRNA     TtGGTCGCTCGCTCCTCTCCTACTTGG-ATAACTGTGGTAaTTCTAGaGCTAAtAcATGC
Yeast_SSU_rRNA     TTGATAG----TTCCTTTACTACATGGTATAACTGTGGTAATTCTAGAGCTAATACATGC
Giardia_SSU_rRNA   GCGGCGG-----TCCCTGCTAGCCGG-ACACCGCTGGCAACCC-GGCGCcaAGACGTGC
*      *          * * ***      ** * * *      *** **      * * ** ** ** ***
Human_SSU_rRNA     CGACGGGCGCTGACCCCTTCGCGGGGGGGATGCGTGCATTTATCAGATCAAAACCAACC
Yeast_SSU_rRNA     TTTAAATCTC-GACCCTTT----GGAAGAGATGT----ATTTATTAGATAAAAA-----
Giardia_SSU_rRNA   GCGCAA-----GGCGGGCGC-----
*          *      *
Human_SSU_rRNA     CGGTCAGCCCCTCTCCGGCCCCGGCCGGGGGGCGGGCGCCGGCGGCTTTGGTGACTCTAG
Yeast_SSU_rRNA     --ATCAAT--GTCTTCGGACTC-----TTTGATGATTGATA
Giardia_SSU_rRNA   -----

```



```

Human_SSU_rRNA      ATAACCTCGGGCCGATCGCACGCCCCCGTGCGGGCGACGACCCATTCGAACGTCTGCCC
Yeast_SSU_rRNA      ATAACTTTTTCG--AATCGCATGGCCTT-GTGCTGGCGATGGTTCATTCAAATTTCTGCCC
Giardia_SSU_rRNA    ----CCGCGGGCGAGCAGCGTGAC-----GCAGCGACGGCCCGCCGGGCTTCCGGGG
                    *      *      ** * *      **** * * *      ** *

Human_SSU_rRNA      TATCAACTTTCGATGGTAGTCGCCGTGCCTACCATGGTGACCACGGGTGACGGGGAATCA
Yeast_SSU_rRNA      TATCAACTTTCGATGGTAGGATAGTGGCCTACCATGGTTTCAACGGGTAACGGGGAATAA
Giardia_SSU_rRNA    CATCACCCG--GTCGGCGCGGTTCGCGGCGCGCCGAGGGCcCGACGCCTGGCGGAGAATCA
                    **** *      * **      ** ** * * * * * * * * * *

Human_SSU_rRNA      GGGTTCGATtCCGGAGAGgGGAGCCTGAGAAACGGCTACCACATcCAAGGaAGGCAGCAGG
Yeast_SSU_rRNA      GGGTTCGATTCCGGAGAGGGAGCCTGAGAAACGGCTACCACATcCAAGGaAGGCAGCAGG
Giardia_SSU_rRNA    GGGTTCGACTCCGGAGAGCGGGCCTGCGAGACGGCCCGCACATCCAAGGACGGCAGCAGG
                    **** * * * * * * * * * * * * * * * * * * * * * *

Human_SSU_rRNA      CGCGCaAATTACCCACTCCCGACCCGGGGAGgGTaGTGACcGAAAAATAACAATACAGGACT
Yeast_SSU_rRNA      CGCGCaAATTACCCAATCCTAATTCAGGGAGGTAGTGACAATAAATAACGATACAGGGCC
Giardia_SSU_rRNA    CGCGGAaCTTGCCCAATGCGCGGCGCGAGGCAGCGACGGGGAGCG-----
                    **** * * * * * * *      * **** * * * *      *

Human_SSU_rRNA      CTTTCGAGGCCCTGTAATTGGAATGAGTCCACTTTAAaTCCTTTAACGAGGaTCCATTGG
Yeast_SSU_rRNA      CATTTCG-GGTCTTGTAAATGGAATGAGTACAATGTAAATACCTTAACGAGGaACAATTGG
Giardia_SSU_rRNA    -----CGCGAGCGAGGCGGGCCACAGC-----CCCCGCCGCGGAGC---CGA

```

```

*   *   *   *   *   *   *   *   *   *   *
Human_SSU_rRNA      AGgGCAAGTCTGGTGCCAGCAGCCGCGGtAATTCCAGCTCCAATAgCGTATATTAAAGTT
Yeast_SSU_rRNA      AGGGCAAGTCTGGTGCCAGCAGCCGCGGtAATTCCAGCTCCAATAGCGTATATTAAAGTT
Giardia_SSU_rRNA    GGGCAAGGTCTGGTGCCAGCAGCCGCGgTAATTcCAGCTCGGCgAGCGTCGCGCGGCGCT
**   *   *****
Human_SSU_rRNA      GCTGCAGTTaAAAAGCTCGTAGTTgGATCTTGGGAGCGGGCGGGCGGTCCGCCGCGAGGC
Yeast_SSU_rRNA      GTTGCAGTTaAAAAGCTCGTAGTTGAAC TTTGGGCCCGGTTGGCCGTCGATTTTTT-C
Giardia_SSU_rRNA    GCTGCAGTTGAAACGCCCGTAGTTGG-----
*   *****   ***   *   *****
Human_SSU_rRNA      GAGCCACCGCCCGTCC--CCGCCCTTGCCCTCTCGGCGCCCCCTCGATGCTCTTAGCTGA
Yeast_SSU_rRNA      GTGT-ACTGGATTTCCAACGGGGCCTTTCTCTTGCTAACCTTGAGTCCTTGTGGCT--
Giardia_SSU_rRNA    -----CCCCCGCC-----GCC--
**           *   *           **
Human_SSU_rRNA      GTGTCCCGCGGGGCCGAAGcGtTTACTTTGAAAAAATTAGAGTGTTCAAAGCAGGCCCG
Yeast_SSU_rRNA      -----CTTGGCGAACCAGGACTTTTACTTTGAAAAAATTAGAGTGTTCAAAGCAGGC--G
Giardia_SSU_rRNA    -----ACGAGGAAACGGGAGCGCTcCaGGCAGGCCCG
**   ***   ***   *   *   *   *   *
Human_SSU_rRNA      AGCCGCCTGGATACCGCAGCTAGGAATAATgGAATAGGAC-CGCGGTTCTATTTTGTGG
Yeast_SSU_rRNA      TATTGCTCGAATATATTaGCATGGAATAATAGAATAGGACGTTTGGTTCTATTTTGTGG
Giardia_SSU_rRNA    TTGGACCCG-----CCGCGTGGGACCGCGCAgCGGG---CGCGGCGC-GCCGCGGCAG
*   *           **   **   *           *   **           **   *   *

```

```

Human_SSU_rRNA      TTTTCGGAACTGAGGCCATGATTAAGAGGGACGGCCGGGGGCATTCGTATTGCGCCGCTA
Yeast_SSU_rRNA      TTTCTAGGACCATCGTAATGATTAATAGGGACGGTCGGGGGCATCAGTATTCAATTGTCA
Giardia_SSU_rRNA    CCCCGAGGA-----GAGCGGGCGGGGGCACCGGTACCGGCCGGGGA
                    * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
Human_SSU_rRNA      GAGGTGAAATTCTTGGACCGGCGCAAGACGGACCAGAGCGAAAGCATTTGCCAAGAATGT
Yeast_SSU_rRNA      GAGGTGAAATTCTTGGATTTATTGAAGACTAACTACTGCGAAAGCATTTGCCAAGGACGT
Giardia_SSU_rRNA    CGGGTGAACAGGATGATCCCGCCGAGACCGCCGGCCGCGCAGGCGCCTGCcAAGACCGC
                    * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
Human_SSU_rRNA      TTTTCATTAATCAAGAaCGAAAGTCGGAGGTTCTGAAGACGATCAGATACCGTCGTAGTTCC
Yeast_SSU_rRNA      TTTTCATTAATCAAGaACGAAAGTTAGGGGATCGAAGATGATCAGATAcCGTCGTAGTCTT
Giardia_SSU_rRNA    CTCTGTCAATCAAGGGcGAAGGCCGGGGGCTAGAAGGCGATCAGACACCACCGTATTCCC
                    * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
Human_SSU_rRNA      GACCATAAACGATGCCGACCGGCGATGCGGGCGGCGTTATTCCCATGACCCGCCGGGCAGC
Yeast_SSU_rRNA      AACCATAAACTATGCCGACTAGGGATCGGGTGGTGTTTTTTTAATGACCCACTCGGCACC
Giardia_SSU_rRNA    GGCCGTAAACGGTGCCGCCCCGCGGCCGGCGCGCGC-----GTCCCGC-CGGCCGC
                    * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
Human_SSU_rRNA      TTCCGGGAAACCAAAGTCTTTGGGTTCCGGGGGAGTATGGTTGCAAAGCTGAAACTTAA
Yeast_SSU_rRNA      TTACGAGAAATCAAAGTCTTTGGGTTCTGGGGGAGTATGGTcGAAGCTGAAACTTAA
Giardia_SSU_rRNA    CCAGGGAAACCGGGAGGCTCCGGGCTCTGGGGGGAgtATGGCCGCAAGGCTGAAACTTGA
                    * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

```

```

Human_SSU_rRNA      AGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACG
Yeast_SSU_rRNA      AGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACG
Giardia_SSU_rRNA    AGGCATTGACGAGGGGTACCACCAGACGTGGAGTCTGCGGCTCAATCTGACTCAACGCG
*** ***** ** ***** ***** ***** ***** **

Human_SSU_rRNA      GGAAACCTCACCCGGCCCGGACACGGACAGGAtTGACAGATTGATAGCTCTTTCTCGATT
Yeast_SSU_rRNA      GGGAAACTCACCAGGTCCAGACACAATAAGGATTGACAGATTGAGAGCTCTTTCTTGATT
Giardia_SSU_rRNA    CGCA-CCTCACCAGGCCCGGACGCGCGgAGGACCGACAG-CCGGGCGCGCTTTGCGGATC
* * ***** ** ** ** * ***** ***** * ** ***** **

Human_SSU_rRNA      CCGTGGGTGGtGgTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCG
Yeast_SSU_rRNA      TTGTGGgTgGTGGTGCATGGCCGTTCTTAGTTGGTGGAGTGATTTGTCTGCTTAATTGCG
Giardia_SSU_rRNA    GCGCGGGCGGTGGTGCATGGCCGCTCCCAGCCCGTGGCGCGAGCCGTCTGCTCCATTGCG
* *** ***** ***** ** ** ***** * ** ***** * *** **

Human_SSU_rRNA      ATAACGAaCGAGACTcTGGCATGCTAACTAGTTACGCGACCCCGAGCGGTCGGCGTCCC
Yeast_SSU_rRNA      ATAACGAACGAGACCTTAACCTACTAAATAGTGGTGCTA-----GCATTTGCTGGTTA
Giardia_SSU_rRNA    AaACGAGCGAGACCCCGGCC-----GCGG-----GCGCC-----
* ***** ***** * * * * *

Human_SSU_rRNA      CCAACTtCTTAgAGGGACAAGTGGCGTTCAGCCACCCGAGATT--GAGCAATAACAgGTC
Yeast_SSU_rRNA      TCCACTTCTTAGAGGACTATCGGTTTCAAGCCGATGGAAGTTTGAGGCAATAaCAGgGTC
Giardia_SSU_rRNA    -----GCGGGACGGCCCGCG-CGAGCGGGAGGACGGC-GGGGCGATAGCAgGTC
* ***** * *** ** ** ** *****

Human_SSU_rRNA      TGTGATGCCCTTAGATGTCCGGGGCTGCACGCGCGCTACACTGACTGGCTCAGCGTGTGC

```

Yeast_SSU_rRNA TGTGATGCCCTTAGACGTTCTGGGCCGCACGCGCGCTACACTGACGGAGCCAGCGAGT--
 Giardia_SSU_rRNA TGTGATGCCCTCAGACGCCCTGGGCCGCACGCGCGCTACACTgGCGGGGCCAGCCGGC--
 ***** ** * * ***** ** * *

Human_SSU_rRNA CTACCCTACGCCGGCAGGCGCGGGTAACCCGTTGAACCCCATTCGTGATGGGGATCGGGG
 Yeast_SSU_rRNA CTAACCTTGGCCGAGAGGTCTTGGTAATCTTGTGAACTCCGTCTGTGCTGGGGATAGAGC
 Giardia_SSU_rRNA -----GCCCGCGAGG-----ACGCGCGGAGCCCCGCGGTGGCCGGGAcCGCGG
 *** ** * ** * * * * *

Human_SSU_rRNA ATTGCAATTATTCCCATGAACGAGGGAATCCCGAGTAAGTGCGGGTCATAAGCTtGCG
 Yeast_SSU_rRNA ATTGTAATTATTGCTCTTCAaCGAGGAA--TTCCTAGTAAGCGCAAGTCATCAGCTTGCG
 Giardia_SSU_rRNA GCTGGAACG--CCCCGCGCACCAGGAA--TGTCTTGTAGGCGCCCGCCCCACCGCGCG
 ** ** * * ** * * * * *

Human_SSU_rRNA TTGATTaAGTCCCTGCCCTTTGTACACACCGcCCGTCGCTACTACCGATTGGATGGTTTA
 Yeast_SSU_rRNA TTGATTACGTCCCTGCCCTTTGTACACACcGCCCCGTCGCTAGTACCGATTGAATGGCTTA
 Giardia_SSU_rRNA CCGGACGCGTCCCTGCCCTTGTACACACCGCCCCGTCGCTCCTACCGACTgGGCGCGGCG
 * ***** ***** ***** ** *

Human_SSU_rRNA GTGAGGCCCTCGGATCGGCCCGCCGGGGTCGGCCACGGCCCTGGCGGAGCGCTGAGAA
 Yeast_SSU_rRNA GTGAGGCCTCAGGATCTGCTTAGAGAAGGGG-----CAACTCCATCTCAGAGCGGAGAA
 Giardia_SSU_rRNA GCGAGCGCCCCGGACGCGC-----GAAGGG-----CC-----
 * *** * ** * * ** *

Human_SSU_rRNA GACGGTCGAACTtGACTATCTAGAGGAAGTAAAAGTCGTAACAAGGTTTCCGTAGGTGAA
 Yeast_SSU_rRNA TTTGGACAAACTTGGTCATTTAGAGGAACTAAAAGTCGTAACAAGGTTTCCGTAGGTGAA

Appendix-2

1. Sequences of GncR candidates

>GncR1
TTCGGGATCAGTTTTGGAGTTAATACCACCAAACCCCTGTGCGTACATGTCGCCCCCTAACCTTCTGATGCGGATACCTTGCCGCAGGGCCGTTAAGCGAGGCTTGGC
CCGTGCGACGATGAGGCTCCCTGCGGGGAAGCCCTGCGGCGCGTCTTAAGGAGGC

>GncR2
TCCCTGGGCGTCGGGCAGAAAGTGCCGGTCTCTGGATTCCGGGGAGTGTCTGGTGCCGATCGGACACTCCCTAGCCGCCACACTGACAGTTATGGTTGCAGGACAA
GCTTAGCGAGTCCGAACCTCGACAGGGATACTCTACAGCGTTCC

>GncR3
CTTCAACTCAGCCGGACAGCCGGAGGCCGGAGACGGAGCACGGTCAGGCGGGCGGGGTGCAGTGCCAGCCCCAGCCGCAGAGCGGCTTCCTTA

>GncR4
CTAGGCTGAAGCTGCCAAGGTGCGTGATCCCTCGGTGATGCCTTGAGTGTGCTTACCAAAGAACAACCACACGGCACAGCCGAATCTCTCATT

>GncR5_CDsno
CTCCAACACGACGGTCTACTGAGAACCCAGTATCTTTAGACTGCTGAGACAGTGTTATATGATT

>GncR6
CCGATCGAAGACCAAGCGGTGCTAGGTTCAAGCCAGGTCCAAGACCCGGGCAGTCTGTGCTGTGGGGCGCCGCTGTAGACGTCTTCCGAACACACCTGCGATAAAC

>GncR7_P_GlsR15sno
GATGCGCCAGGCTGACGGTAGGACGCCTAACCCGATTGACTACTCCTTGTTTCCCTCGCAGAATGATTATCTGTCTCCGAGCAAGCACGACTATGAGCTTACTTAT
GAGATCTGACTCC

>GncR8_CDsno
TGATGATTGCAATTACCGCCCGAGGGCCCTCGGGCTCCGCTGAGGACATGCTGGTCTGACT

>GncR9
TGGACGATGAAGTGGAGATGCTGGACACGGCTTTGCTCTCCACCGGAGCACATATGCTGCAGGATGACCGGCGCCTGTCTCCACCACGTGCCAGCTAAACTGCAG
CCACATT

>GncR10_U5cand
ACAACCTGCAGATCATTCATCTCTGCGGTGGATGTATCTATCTGGTACGAGATATGTTGGGAGAGGAAATGGCAGACAGTTGCATTTTTTTGGGGTTATGGGCTG

>GncR11
GCGTAAGGTTTTCTGTGACTACTACCCAGAGTAAACCGGTGAG

>GncR12
CAGAGTCGGCTTCGACTTTAGCGTAGTTACTGTTTCGTGCGGCTTAACCGCCGATCCACTACATGCAAGGGGCAGCCGGGCTGTGAGGCAGCTGCCAGGATGGTCTGCG
CCTGTCCCGGCTGGCGCCGTCCACCTT

>GncR13_CDsno
AAAAAATAAATGAAGACAGAACCACAGACCTGTACTGACCCTTGATGTTAGTTGTCGCTCTGATA
>GncR14_likely_U
GCTCCTAGAGGAAGAGGCAGGCATGCAGGATATTTTTGGATGGACAGCCCTCATAAGGGCAGCAGAGAGTGGCCACGCAGGCTGCAC
>GncR15
GTATGCTGCTATGCTGACATGCCGGTACACTTTTTATGAGAGCGAATGTAAATAGCCCTG
>GncR16
CTGAACGATAGAAAAGACGCGTGCAGGGCGGTTGCCAACACAAACTGCGCACAAGAGCTGCAGAATGCGGC
>GncR17
GAGGTAATAGACCAGGCTGCCAGCCCCGGCGAAGGTCTGCAAGTGTGACGGAGACAATGGCTACACGCTCCAGGGCGACGCGTGCACCAAGGCGGCTCCTGACAACG
CGTGCCAGACCCTGGGAACCGCCGGGTGTGCCAC
>GncR18
TACCACTCTGACCGTGAGGCGCATGCCTAGGGCATGGAGAAGAGCAGACTTGAG
>GncR19
CAGAAGAATGCCAGCAAGTCATGCAATGCCTGTGGATCCGTCCTTCGACCTTCTCCTGACAGACATGTGTCTTTTGGCATGCAGCCCTGC
>GncR20_U1cand
aaacatcagCGGCATCGTCatcacgaaGATGAgcAAAAgcATAAAGTTCGAGATCCTCATCGTGTCTGCGAAGAGGAGGTTGACCAGGTTGCCGgcGGCAGaatttgCGGGTGATG
TCCG
>GncR21
CTCTGATCGCCTGGCCGGAGCACATTTGTGATCTCCTATACCA
>GncR22
AGTATTTAGAACTCGCCACGAGGTCAGTATGGTTCACATGGATCACACACTAGTAAATAAAAAGATGA
>GncR23
CCCGATGACGAATAGCTGTCTTGCGGAGGGCGGTCATGACGACGAAGCCATCACGTAGGATC
>GncR24
AGACAGAAGTAGAGCCCGTTCTCCAGTAAATCTGCAGCTCA
>GncR25
CGCGGAGGCAGGGGCCGGCCCGCCTTCAACTCA
>GncR26
TCTGATTAGTCGGGCATTTGCACTGGGCGCCTAACACGGATAACCCGCGGAATAAATCTTAGTTAA
>GncR27
GTGAATAAAACTGGGAAAAA
>GncR28
TCCTACGGCGGAAACT
>GncR29_HACAsno

TCAAAAGCAAGGCTAGAGCCATGGAGCGCGGATCTGCGCTCTGCCAGATACGCCGACAGAAAGCACCAAGGAAGGATGTGGATCTCCATGTCTGCCGTGTGCGCGC
ATATCCT
>GncR30
CACGAGGAAACGAGTGTTCGCCGGGCATAACTGGGCATGCATTTTCCTTGCCCAGTCTGCCTCCATACTAATTTTCCTA
>GncR31
GGCGCAGACAACAGCAAGAGTCCAGATGGAGTACCTGCACTCCGCCAAGGTTTAGCGTAG
>GncR32
ACACAAAAGGTGAGCGCGTAAGCAAAAGCCAGAAGCCCGTTGCAGCGCTTGCTTCGCAGCTCTACGGGCGC
>GncR33
TGTAGGTCTAACATGCTTGCCACGGCGTCCCCGGACATGGCACCGTCTATGTCCTGCTTGTTGGCGAGGATGAGGATGGGAACACCTGAGCTTGGGGCTGTTAGTACG
CCCTCAAGAGCCGTCCGAGCCTCCT
>GncR34
AGGAACCTATATTAGCAGAATTGGAAACGTTATAAGTGGGCTCCATCTTTTGCAGAATGTTGGAAATGTAGT
>GncR35
ACGGGAATAACGCCACAGGATCTCAAGGAAGGGGCCCTCAACTGAAGCTCTGATCGGGTCCCAAGCACAAAGTAAATAATTGCC
>GncR36
GCTGAGAACGTCAGGAAGGAGCCTAGAAAAGAAGTTGCTGCAC
>GncR37
GCGAAGAAGACAAAGACGAACAAGAATCCCGCGCTGGGTCTCGCAAAGGATCATCTCAAGAAGACGTGGAATTA
>GncR38
CCCACCGCGTTCCAGTGTGGCCAGGGGCAAGGAGGCCTGCTCTCCCTGGCCTCTGCGGAAACGGGCAGCTGCGTGATCCACTGACAGCCACCAC
>GncR39
TGTGCCACTGTGGCTTCGAGCTCTATAATGCGCGACTTAAGAACCCTGCTTCTACAGACTTTACTTCAAGTAAAGATGTCGCAGTTAGTGCCTCCTCAACACAGCTTT
C
>GncR40
TACTTCTTGCGTCCACGGCCCTGCACCTGCTCCAGTTGTGTCAATATCTTCTGTCCCTGGCCTTTGTCTCCACTGTGCTCGTCCATATAGCAGACATAGAACTTCAGCA

2. Upstream sequences of *Giardia* tRNA genes with A/T-rich regions highlighted

```

>tRNAAsn_6497_60875_60975
ATAATTGTCGGTACGACGTATGCTGACCCACACAATTCTTGAAGACGACAATCTGCTGCAACACAATTATACTTATTCTGCTATCGCATCCGCGCGCGC
>tRNAAla_7704_192066_192166
GCCCTCCGCGGATCGTCTGCAGAAGCCCCGACACATCCTTAGAGACCTAGAAACCTTTGAAGTACCTTTTCGCGCATGTTGGCATTFTTTTCCCTTGGCAC
>tRNAHis_7649_23432_23532
TGTGGATTGGATCAGCTGTTCCATAGCCGAACGCATTTACACCACAAACCATCTGCCAATACGATTATTTCCGTTGGTTTTAGCTTGTGCGGCCCGCCGG
>tRNAArg_7607_178483_178583
ATTGAGGCGAATCCAGAGCGTGGCCAGTGGCGGGGAGGGCAGAGAGGCCGTGCGGGCGAGGTCTTCGACGGCGGCCGCAAAGTTGGGGGCGGGCATAAAAAG
>tRNAGly_7416_14623_14723
ACCCGTCGATGCTGGTGGAAACCAGCGCGCTGTCGGCGACGTGCAGCGACGTCACCGCGTGGCTGTGAGAAAGGTCGTGGCGGGGTGGGGCGGAAAAGCG
>tRNACys_7069_41236_41336
GTGCGTGGGTCCTGAGCTCCTCGAGGCGCCCGGTGAGGCTGTTGGTGAGCCAGACGGAGGTCGGGAACCGGCCATTTATCGGCGCCGGGCGGAGGCAAAA
>tRNAGlu_7645_129794_129894
GTATGCTAAGACGAAGGAAGGCGGAAACGCTCCGGCGCCGGAATCGAACCCGGATCTGTTGGGTGAGAACCAACCATTCTAGCCGTTATACTACGCCGG
>tRNAAsp_7609_157934_158034
AGAAAAAGTTAGGATAAACCAATAAAAATTACAAATACTCTATTACAAATAATTTCTATTAACCAAATATTTTATATAATCTTTAATTTGTTTCCTCCCTG
>tRNAIle_6481_45711_45811
AGGGCGAGCGTCCTTCAGTCACAGAGCGCCTCGTTCAGCGAGTAAACCCGGCCAATACAATCATACTACCTATACAATTTATATGTTCTCGCGCAGGCGAC
>tRNALeu_3413_23735_23835
CGCGGGCGGTACCGCTCGATGAATGGGAGTGCCTTCATCGGCCACTCTTTTCAAACAATAAACTGCCTACTCCGGCCTCTGGTGAGGTCCCAGCGGGGCGC
>tRNALys_3550_1551_1651
TCTGGAGGTACGTGTAGCCGGACGGCAGAACGGCCATTGATCAAATACCAACACACTGCTAGCAAAATTA CAAGTCATTCTGATATTCTGCGTGCGGTGCG
>tRNAMet_7513_364026_364126
ACGAGCGAGCAGGCGGAGGCGGCCCTCGCCCTAGGCCCCACGGAGTAAACAAACAGAAAGAAGAAATAACACGTGCATGCTGCTGCCTGGCCCTGACGTGT
>tRNAPhe_7520_92453_92553
TGGCCCGATGTGCTGTCTAACCAGAAACCGGCTGCATAAGCCGGCCACGCCACAGACCTGAGCGGCGACAAAGCAACCCTCTGTGACCCCGGCTGGAAAG
>tRNAPro_7513_379598_379698
AGTACGAACGGGTGATCGAACAGCAGAACGAGCTGGTGACAAACCTCTGGGGGATCGTTGGGGAGCTCAACAGGCGCTTGGGGAGAAGGGCGGTAACG
>tRNAThr_7513_666161_666261

```

CTGCCGCGGGCGGGCGTCCGTCCCCGCGCCGTGGGCCGGGGCGCCGGGTCTGAAGAGGCCCGAGGGCTGCGGCGCAGACGTCATTTTGGGCGGCAAAAAG
>tRNASer_3407_294473_294573
GGGGTCCGCTGGTAGCACCAGAGTTGGAGGTGGATGCAGTGCTCGACGGAGAAGTGAGCGCATAGCATGAAGCGTTATCCCGGCGGAAGCCCAGACGAAAG
>tRNATrp_6907_25968_26068
CAGGACCCCTCCACGACGCAGGCACCCGTCCGCGCCAGACACAGCGAAGAGCACGGGGCCGCGCGGCGACGAAGAGGGCAGGCGGGCGCCGAGGCATAAA
>tRNAVal_7704_131532_131632
CGCGCTACAACATCCAGTCCC GCCGGCCCCGTGGCACCGGGACAGGTTCTGTCCCCGATCCCTCGTCCGGTTCCGGCTCGTGGAGAGAGTGAACGGAAGA
>tRNApseudo_7649_74214_74314
ATCACACTCAGACAGCAGGTGACAACCTCACCCCAATGAAACAGAGAGAGTAAACCTTTTATAATAAAGAACTAGCCTTCATTATCATCGGCCGCTCGCTG

3. Upstream sequences of 9 novel ncRNAs which are likely to be transcribed by RNA Pol III class-3 mechanism:

["A/T-elements" are boxed in red and "G/A-elements" are boxed in blue]

```
>GncR1_7513_834858_834558
ATAATCCAGTCCACGATCCGTACAGCGCTGGCAGACCGCGTCGTTATCGCCATTGCACAC
AGGATCAATACCATCATCGACTTCGACAGAATCATAGTTATGGACGCGGGCGAGGTAAAG
GAGTTCGACACCCCCAGGGCGCTCCTGAGCAACCCAGAGTCGCTCTTCAGCAAGCTGGTC
GCAGAGTGCAGGACGCAGAGGAGCTCGCCGGGATTGCCCGTGTCTCGGAGTG[AATAAAA]G
CAAACC[ATTTTAAAT]CGAATTCTCCCG[GAAAAGAAAAAG]GGTGGGGCAAGCGATCATACC
A
```

```
>GncR3_5739_223496_223196
GTATCTGTTTTGAAGAACTCCAACCGGGCGTCTCGCCCGGGCGTCAAGGCTACTGAACTG
CGTCGCAGAGCAAAAAGTAGCCCGGCCAACATGTCCATCGGGGCGTCATCAGCGTGCTCG
AAGGCGCGGAAGTACAGCAGGGCCCCGGAGCGCCGGTAGACGTAGAACTGGAGAGGCGGC
ATAGCAAAG[TTTTTTAAAT]GGATTGCTTCGGA[GAAAGAAAAAG]GACGCCGAAGCAGCC
GCGGCCTCTGGCTTGGACCCCGTGGCGTTCGCCGGCCTCCGCGGAGGCAGGGGCCGGCCCG
C
```

```
>GncR8_7560_43071_42771
GCGCTTGTCAAACGTTTTAGACTTATCAAAGCTCAGTCTATCAAGATATCGTTCGGAAA
GCTGAGTCGGAAGTATCCTTCAGCTTGTCTCCGGGAGCCTTACACGCCGTATGTCTCC
GTTATCCGTATACTCCTTGGCCAACCTGCAGCGACCCCTGCCATGCTCCTACTGGACTTTGT
CCGCGCATCCTTGGGCATTAATTCCTGGGATTGAAGTCTTTTTCCAGAATTTGTTCTTTT
AGTGTTTAGTGTCTTTGTCTTTTATCTTAGCTTTTCTATTAAATTGAAAGTC[GAAAATAAA]
G
```

```
>GncR13_3407_251240_251540
CTAGAAGGTGCTGGTGAATTGGGCTGCGCGTTTCTTCCCACAAAAGCCGCTTGTACTTCT
TCATAAGCTCTTTTCGAGAAGTCTGAGCCAATATCGAAGGGAGCAGTCTGCATCCAGCCAA
CTCAACCTCGCTTTCTAAAAAATCCAAAACAGATTACAGAGATTTATGAGATACGGTAAC
```

GCAGAGGCCGGACCGGTGCTTTTACTGCAAGTTACTAGGCAGCAAGTTCAAGTCTGGGAA
 CCGAGATCGTTTCAAAAACGGTTTTAAAAAGCTCCGAAGCAAATGAGAACAAAAGCAGAC
 G

>GncR23_4036_136457_136757

TTTACTTCAATGCACACTAGAGAAAAGTGTAAAGGAGTGGTACGATGCCAAGAGTGACACC
 GCAGCCCAGGACTGGGCCGCACAATTAGAAGCTCTGATAGTTACGCGGGCGTCTTCCACA
 CCAAATGACTATAGCTTTCCTCACTCAGTGAACATGAACGAGGTCGACTTTGAGGAGGACCTT
 GCGGAGGACGCGATCGATATATCCAGCACCTCTTCTTCTTAGTCTCTTCTTAGCATCCAG
 AATAAATCACATTAATGTATTTTAATTTGAATTTTATCCCCGAGAAAAAGAACCCCA
 A

>GncR25_5739_223519_223219

TCTCGTAAATGTGACAGGTGTAGGTATCTGTTTTGAAGAACTCCAACCGGGCGTCTTCGC
 CCGGCGTCAAGGCTACTGAACTGCGTCGCAGAGCAAAAAGTAGCCCGCCAACATGTCCA
 TCGGGGCGTCATCAGCGTGCTCGAAGGCGCGGAAGTACAGCAGGGCCCCGAGCGCCGGT
 AGACGTAGAAGTGGAGAGGCGGCATAGCAAAGTATTTTAAATTTGGATTTCGTTTCGGA
 AGAAAAAGGACGCCGAAGCAGCCGCGGCCTCTGGCTTGGACCCCGTGGCGTCGCCGGCCT
 C

>GncR29_6593_40418_40718

TCAAGCTGATTGATGACGCGCTTGATGATCGTGTCTTCCCGGCCAACAGTTCAGCTTTG
 CCTCTGAGGAGGCGGCGGATCTGAGCGATCTGGAAGGAGCGCACGTTGTTCGACGCTGACG
 AGGACGATCTTCTGTACTCGGTCAGACACCTCTCCAGCTTCGCAACATACGCCTGTTCG
 CTGGCCTGCTTCGCAGTGAGTGGGGCAGCGGACATTTAACCTACGCACGGAATCTATAGA
 TGTCTCCAGAATCAAAATTAATAATGGATTGCTTCTTAAAAATGATGGCCGGAAGAGAAAAAG
 A

>GncR30_3550_17802_17502

ACGTCTACGGATTGTAGTATGCTCTCTGCTAGATCTATCTGCCTTAGAAGGGTCTGCTTA
 TGCATAGCCAGCTTCAATGCTTCGCGCTGTATGGCTCTAGTAGTCTCCAGTTGATCAGGC
 CCCTTTACCACATCCTTGCTGCCAGAGCCTGGATGCGCAGAGGCCATCTTTGAGGAGCTG
 GCCAACCGCTTGTTCGGGGGAACTTTTGCTGGAACCTTCAGCTGACTGAGGTGAAGGGAA

GCAGACTCCATGGCATAAATAAATGCAAAATTCTTTAACCTGAAAACAAAAITGGCTAGCA
A

>GncR37_6770_20897_20597

AAGCAAGAGCGGCAGGCAAGCAAGCACTCTCTTGCGAGGCTCCTGGGGTATCCCCACTGC
CACGTAGATACAAACGAGTACTATACATAAAGAGAAGCTGGCGAGGACAAGCTGCTTGGA
GGTCAACGGTATTTCTAATCCAGGGCGGATGGGATGTACTGTATAGCCCAAAGAGGACAG
GCTCTGCGGGGCATTGTGTGTGGTTATTCTCATGAATATAAATTGATTCCACTTTTTGGC
TTGGGAAAAAAGCCCCTTCGTCATGACTGCGAATAAGACGCTCAGTAGGAAGTTGAGGCT
C

CAGAGUCGGCUUCGACUUUAGCGUAGUUACUGUUUCGUCGGCUUAACCGCCGAUCCACUACAUGCAAGGGGCAGCCGGGCUGUGAGGCAGCUGCCAGGAUGGUCCUGCCC
 UUGUCCCCGGCUGGGCGCCGUCCACCUU
 .(((((((.....)))))).((((((...((...(((.....))))))..))..))))).(((((((.....((...(((((.....))))))..))))))..)))).. (-55.20)
 >GncR14_likely_U
 GCUCCUAGAGGAAGAGGCAGGCAUGCAGGAUAAUUUUGGAUGGACAGCCUCAUAAGGGCAGCAGAGAGUGGCCACGCAGGCUGCAC
 ..((((.....))....((((((...((...(((.....((...(((.....))))..))))))..))))..)))).. (-26.60)
 >GncR15
 GUAUGCUGCUAUGCUGACAUGCCGGUACACUUUUUAUGAGAGCGAAUGUAAAUAGCCUG
((((.....((((.....))))))..)))).. (-8.80)
 >GncR16
 CUGAACGAUAGAAAAGACGCGUGCGAGGCGGUUGCCAACACAAACUGCGCACAAAGAGCUGCAGAAUGCGGC
((((.....((((.....))))))..)))).. (-18.20)
 >GncR17
 GAGGUAUAGACCAGGCUGCCAGCCCGGCGAAGGUCUGCAAGUGUGACGGAGACAAUGGCUACACGCUCAGGGCGACGCGUGCACCAAGGCGGCUCUGACAACGCGUG
 CCAGACCCUGGGAACCGCCGGGUGUGCCAC
 ..((((.....))..((((.....((((.....(((.....(((.....(((.....(((.....(((.....(((.....))))))..))))..))))..)))).. (-55.10)
 >GncR18
 UACCACUCUGACCGUGAGGCGCAUGCCUAGGGCAUGGAGAAGAGCAGACUUGAG
((((.....((((.....))))..)))).. (-15.00)
 >GncR19
 CAGAAGAAUGCCAGCAAGUCAUGCAAUGCCUGUGGAUCCGUCCUUCGACCUUCUCCUGACAGACAUGUGUCUUUUGGCAUGCAGCCCUGC
 (((.....((((.....(((.....(((.....(((.....(((.....(((.....))))))..))))..))))..)))).. (-22.70)
 >GncR21
 CUCUGAUCGCCUGGCCGGAGCACAUUUGUGAUCUCCUAUACCA
((((.....((((.....))))..)))).. (-8.00)
 >GncR22
 AGUAUUUAGAACUCGCCACGAGGUCAGUAUGGUUCACAUGGAUCACACACUAGUAAAUAAAAGAUGA
 ..((((.....((((.....))))..)))).. (-11.50)
 >GncR23
 CCCGAUGACGAAUAGCUGUCCUGGCGGAGGCGGUCAUGACGACGAAGCCAUCACGUAGGAUC
((((.....((((.....))))..)))).. (-16.90)

```

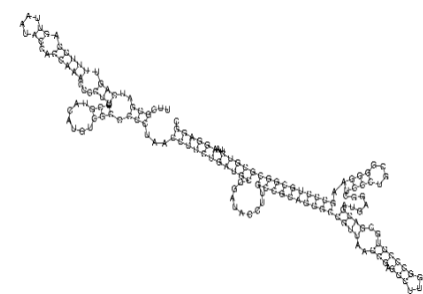
>GncR24
AGACAGAAGUAGAGCCCGUUCUCCAGUAAAUCUGCAGCUCA
.....((((((.....)))))). ( -4.54)
>GncR25
CGCGGAGGCAGGGCCGCCCCGCCUUAACUCA
...((((((.....))))))..... (-14.60)
>GncR26
UCUGAUUAGUCGGGCAUUUGCACUGGGCGCCUAACACGGUAACCCGCGGAAUAAAUCUUAGUUAA
((((.....((.....((.....)))))).((.....))))..... (-11.80)
>GncR27
GUGAAUAAAACUGGGAAAAA
..... ( 0.00)
>GncR28
UCCUACGGCGGAAACU
((((.....)))..... (-2.00)
>GncR30
CACGAGGAAACGAGUGUUUCGCCGGCAUAACUGGGCAUGCAUUUCCUUGCCCAGUCUGCCUCCAUACUAAUUUCUCCUA
...((((((.....((.....((.....((.....((.....)))))))).))))).)))))... (-27.60)
>GncR31
GGCGCAGACAACAGCAAGAGUCCAGAUGGAGUACCUGCACUCCGCCAAGGUUUAGCGUAG
.(((((((.....((.....((.....)))))).)))).))..... (-16.10)
>GncR32
ACACAAAAGGUGAGCGGUAAGCAAAAGCCAGAAGCCCGUUGCAGCGCUUGCUCGAGCUCUACGGGCGC
.....((((.....((.....))))))......((((.....((.....)))))).))..... (-25.00)
>GncR33
UGUAGGUCUAACAUGCUUGCCACGGCGUCCCCGGACAUGGCACCGUCUAUGUCCUGCUUGUUGGCGAGGAUGAGGAUGGGAACACCUGAGCUUGGGGCUGUUAGUACGCC
CUCAAGAGCCGUCCGAGCCUCCU
.(.....((.....((.....((.....((.....)))))))).)).....((((.....((.....)))))).....((((.....((.....))))))
).)))).)))).)))).). (-50.50)
>GncR34
AGGAACCUAUUUAGCAGAAUUGGAAACGUUAUAAGUGGGCUCCAUCUUUUGCAGAAUGUUGGAAAUGUAGU
.....((((.....((.....((.....)))))).)))).))..... (-13.80)
>GncR35

```

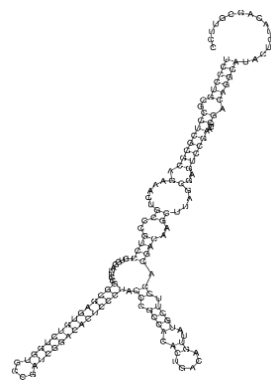
```

ACGGGAUAACGCCACAGGAUCUCAAGGAAGGGGCCCUCAACUUGAAGCUCUGAUCGGGUCCCAAGCACAAGUAAAUAUUUGCC
..(((.....((((((.....((.....))....))))).))....)).....(((.....))).. (-21.90)
>GncR36
GCUGAGAACGUCAGGAAGGAGCCUAGAAAAGAAGUUGCUGCAC
.(((.....)))...(((.....))).. ( -5.30)
>GncR37
GCGAAGAAGACAAAGACGAACAAGAACAUCCCGCGCUGGGUCAUCGCAAAGGAUCAUCUCAAGAAGACGUGGAAUUA
((((.....((.....((.....))))..))).....(((.....)))..... (-11.16)
>GncR38
CCCACCGCGUCCAGUGCUGGCCAGGGGCAAGGAGGCCUGCUCUCCUGGCCUCUGCGGAAACGGGCAGCUGCGUGAUCCACUGACAGCCACCAC
....(((.....((((((.....((.....))))..))).....)).....(((.....)))..... (-37.30)
>GncR39
UGUGCCACUGUGGCUUCGAGCUCUAUAAUGCGCGACUUAAGAACCUCUGCUUCUACAGACUUUACUUCAAGUAAAGAUGUCGAGUUAGUGCCUCCUCAACACAGCUUUC
.....((((.....((((.....((.....((.....))))((.....))))..))).....)).....)).....)).....)).....)).....
(-25.90)
>GncR40
UACUUCUUGCGUCCACGGCCCUGCACCUGCUCAGUUGUGUCAUAUCUUCUGUCCUGGCCUUGUCUCCACUGUGCUCGUCCAUAUAGCAGACUAAGAACUUCAGCA
.....(((.....((((((.....((.....))))..))).....)).....)).....((.....)).....)).....)).....)).....)).....
(-17.70)

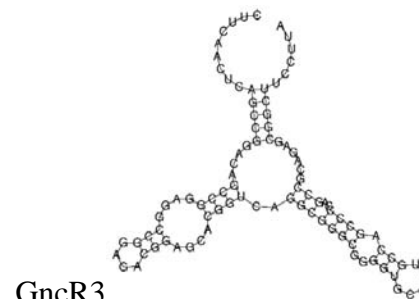
```

Predicted secondary structures of uncharacterised *Giardia* novel ncRNAs

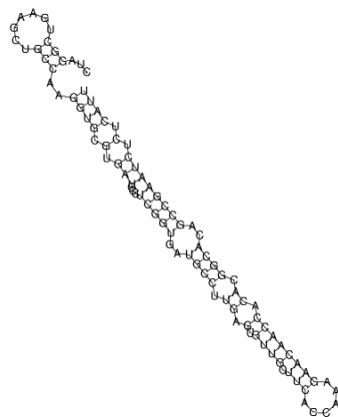
GncR1



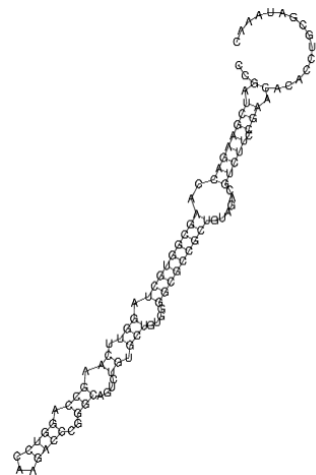
GncR2



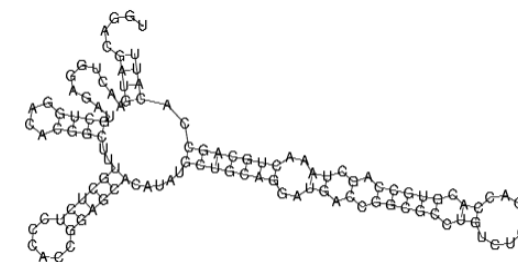
GncR3



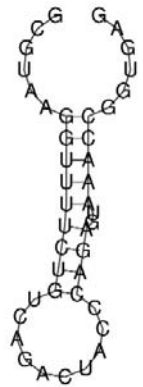
GncR4



GncR6

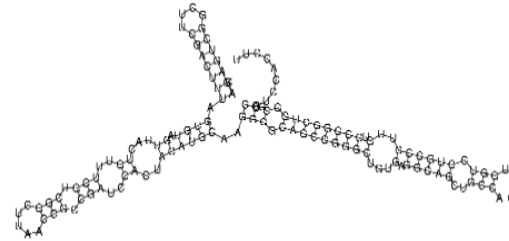


GncR9

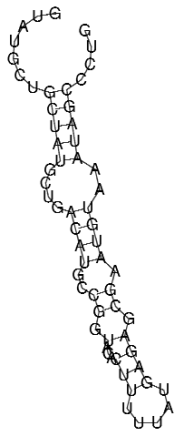


GncR11

GncR12



GncR14



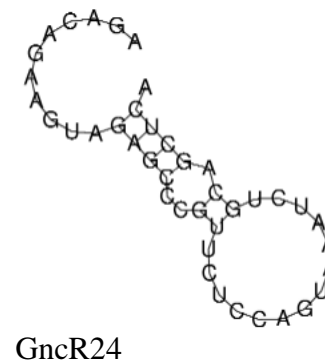
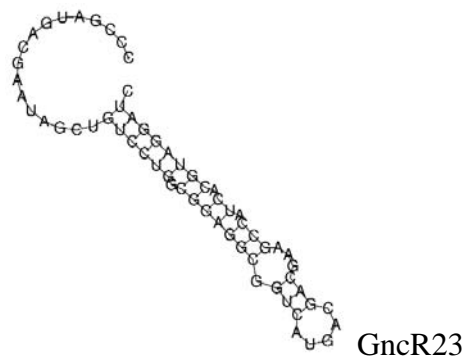
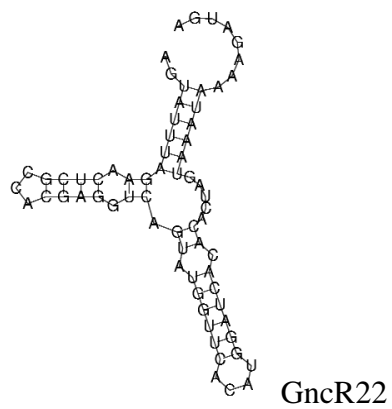
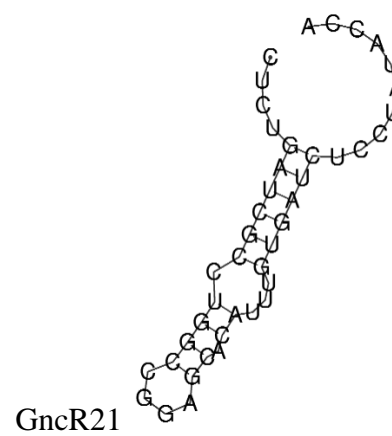
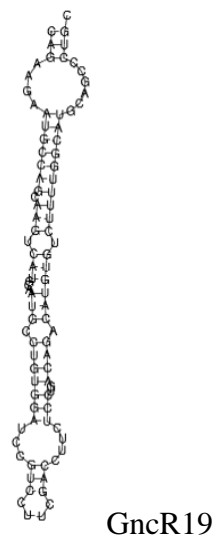
GncR15



GncR16

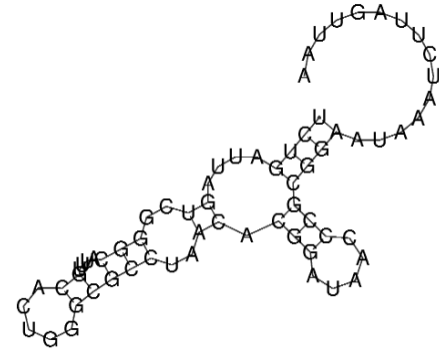


GncR17





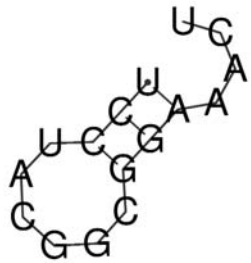
GncR25



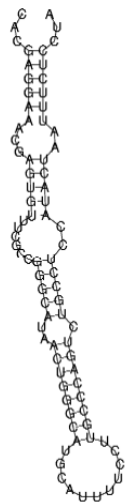
GncR26



GncR27



GncR28

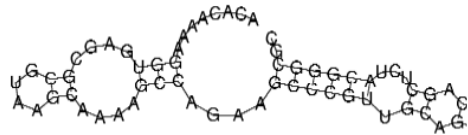


GncR30

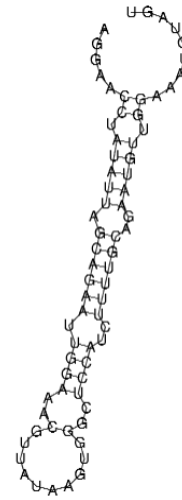
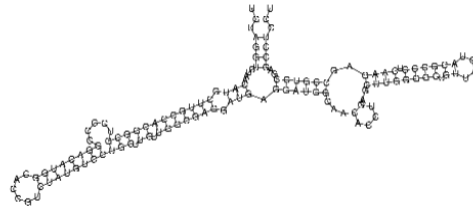


GncR31

GncR32

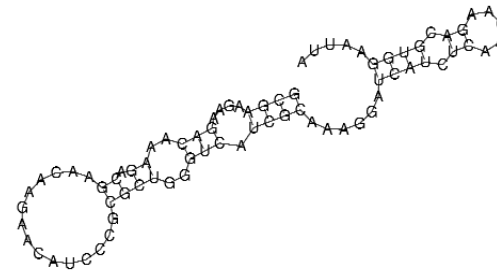


GncR33

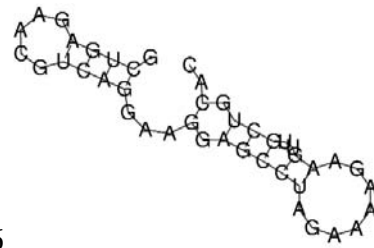


GncR34

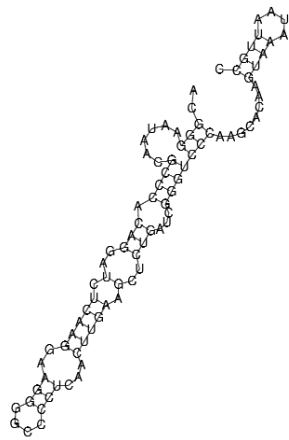
GncR37

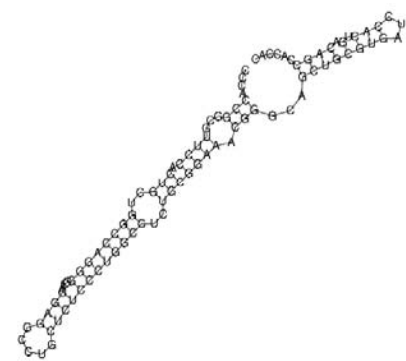


GncR36

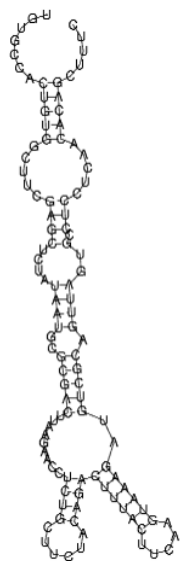


GncR35

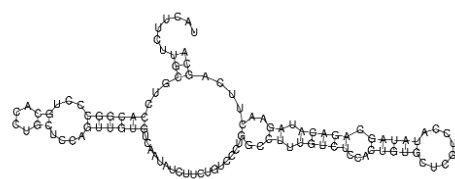




GncR38



GncR39



GncR40

5. Multiple Structural alignments of 6 groups of GncR RNAs which have overall similarities above 40%

Group1: GncR2, 6, 17

```
GncR17 GAGGTAATAGACCAGGCTGCCAGCCCGG--CGAAGGTC'TGCAAGTGTGACGGAGACAATG-GCT-ACACGCTCCAGGGCGACGCGTGCAC
GncR17 .....(((.....(((.....))--.....)).....((((.....(.....-.))-.)..)).....(((.....(((.....
GncR2 TCCCTGGGCGTCGGGCAGAAAGTGCCGGTCCCTCTGGATTCCGGGGAGTGTCTGGTGCCGATCGGACACTCCCTAGCCGCCACACTGACAG
GncR2 .....(((.....(((.....)).....)).....((((.....(((.....(.....(.....(.....(.....(.....
GncR6 CCGA-TC-----GAAGACCAAG--CGGT--GCTAGGTTTC--AAG--CCAGGT-CCA---AG-ACCCGG-----GCAGTCTGTGCT-
GncR6 .....-(.....-(((.....(.....-..))--..(.....-..))--((.....(.....(.....(.....(.....(.....

GncR17 CAAGGCGGCTCCTGACAACGCGTGCCAGAC-CCTGGGAACCGCCGGGTG---TGCC--AC
GncR17 (.(((.....)))..)))).(((.....-..))).....(((.....))---..))--..
GncR2 TTATGGTTGCAGGACAAGCTTAGCGAGTCCGAACCTCGACAGGGATACTCTACAGCGTTCC
GncR2 .....(.....(.....(.....(.....(.....(.....(.....(.....(.....(.....(.....(.....(.....
GncR6 G-TGGGGC-GCCGC-T-GTAGACGT-CTTCCGAACACAC--CTG-CG--A-----TAAAC
GncR6 (-(((.....-..)))--)))))--)))))((.....(.....-..))--)--.....
```

Group2: GncR3, 4, 6, 14, 37

```
GncR14 GCTC-----CTAGAGGAAGAG-GCAGGC-ATGCAGGATATTTTTT-GGATGGACAGC---CCTCATAAGGGCAGC--AGAGAGTGGCCA-CGCAGGC---
TGCAC
GncR14 .....-----((.....-(((.....-..))).....-((.....))((---(((.....-..)))..((.....-..)))---
.....
GncR3 CTTCAA-----CTCAGCCGGACAGCCGGAGGCCGGAGACGGAGCAC-GGTCAGGCGGG---CGGGGTGCAGTGCCAG-CCCCAGCCGCAG-AGCGGCT--
TCCTTA
GncR3 .....-----((.....-(((.....-..))).....-((.....))((---(((.....-..)))..((.....-..)))--
.....
GncR37 GCGAAGAAGACAAAGACGAACAAGAATC----CCGCG-----C-----TGGGTC--AT-CGCAAAGGAT-CATCTCA--AG-AAGACGTGG-----
AATTA
GncR37 .....-----(((.....-(((.....-..))).....-((.....))((---(((.....-..)))..((.....-..)))---
.....
```


GncR9 .((.....)).....

Group5: GncR18, 23, 36

GncR18 TACCACTCTGACCGTG-AGGCGCATGCC--TAGGG-CATGG--AGAAGAGCAGA--CTTGAG
 GncR18 ..((..(.....((.....-(((.....))--..)))-....))--.....((.....--..))...
 GncR23 CCCGATGACGAATAGCTGTCCTGGCGGAGGCGGTCATGACGACGAAGCCATCACGTAGGATC
 GncR23 ..((..(.....((.....((.....))..)))-....)).....((.....))...
 GncR36 --GC-----T-GAGAACGTCAGGAAGG-AGC--CTAGAAAAG--AAGTT-GC-T--GC--AC
 GncR36 --((-----(-(.....)))...((-((--(.....))--..))-.)-)--..

Group6: GncR19, 31, 34

GncR19 CAGAAGAATGCCAGCAAGTCATGCAATGCCTGTGGATCCGTCC'TTCGACCT'TCCTGACAGACATGTGTCTTTTGGCATGCAGCCCTGC
 GncR19 .(.....((.....((.....)))))...(((.....((.....))..))..)...(((.....))..)).....)
 GncR31 GGCGCAGACAACAG-CAAG-AGTCCAGA-TGGAGT---ACCT----GCACTCCGCCAAGG-T-----TTAGCG-----TAG
 GncR31 .((((.....((.....-.....))((.....-..)))---(((-----((.....))..)))-)-----))..))-----)
 GncR34 ---AGGAA----CCT---ATA--TTAG--CAGAATTGG-AA-ACGTTATAAGTGGGC-TCCATCTTTTGC-AGAATGTTGAAATGTAGT
 GncR34 ---((.....-....))---((.....-.....))..-..-)-(((.....((.....-..)))...(((.....))..)))))...)

Appendix-3

1. Sequence and genomic locations of the *Giardia* U-snRNA candidates

U-snRNA candidate	Genomic locations	Sequence
U1	Ctg02_25_94598_94719	AAACAUCAGCGGCAUCGUCAUCACGAAGAUGAGCAAAGCAUAAAGUUCGAGAUCUCAUCGUGUCUG CGAAGAGGAGGUUGACCAGGUUGCCGGCGGCAGAAUUUUGGCGGGUGAUGUCCG
U2	Ctg02_10_171577_171750	ACUUGCGUCGAACCACAGCUGCAUUGAACAAUAGUUUCUGCUCAAUUGAGAGAUCAUAUAAUAUGGC UGAUUAGCGUGCAGCUGCAUGCCUUUCAUAUUCGUUUGUUUGCUUUGUUUUAAACUAACAA CUAGGAUAGUCGCCUUGCAGCGACAAGAAUAUCCUACG
U4	Ctg02_40_7695_7563	AAUAUUGCGAGAAAACCCUCUUAGAAUUGAUAGAAGACAGUCCUGGCGGGAUUCCAAUAGAAACUGUU AAgCUUCUAaCCUUUCAGAuGCUUCGUGGUGUCGAAUUUUUgUGGGAGUUCAUGGAGAUUAUGUCA
U5	Ctg02_42_10253_10357	ACAACCUGCAGAUCAUUCUUCUGCGGUGGAUGUAUCUAUCCUGGUACGAGAUUAUGUUGGGAGAGGA AAUGGCAGACAGUUGCAUUUUUUGGGGUUAUGGGCUG
U6	Ctg02_13_58084_58200	GAAGUGUCCGGGAACAAGUGAGGCCUGCACUUUUCUGCAAACAGAGGAAGUUCAAGCUGUUCGUGCAU UGAGUAUAUUAUCUACAGAGUCGUGGUACUCAGACCCUACAGUGUCCUCU

2. Upstream 100nt sequences of the *Giardia* U-snRNA candidates

U-snRNA candidate	Upstream 100nt sequence
U1	ACAGGATCAGTCCAACAATCGGCAGGATGATGAAGAGAGCAATCAGGCCTCCAATCCCTTTGGGGTCGATCTCGATGG CCATGTAGACCAGGTAGCAGAC
U2	ACAACGATCTTCTCATTATCATCCAGCCTGCTTAGCCCTTTAATGATAGCACGTAAAAGATGAGATTTTCCCGTTCCT GCAGAGCCTGAGAAGAAGAGGG
U4	ACACTCCGGGAGGTCAAGATCTTGAAAAACGCGCAGCATCCTCATATAGTATGTCTCAAAGAGGCTTTCAAGCGCAAG CAACGGCTCTATTTAGTCTTTG
U5	AGTCCTAACTTGGTAAGTCGGCACCGTCGCGTTTGGAACACGCGATAACGTGGGGCCATAATTGATCCTTCATCAGG CAGCACGGCAGCAAGCTGAGAT
U6	AATCTTAATAGTAAACTGTCCGTCTTCTGTTCCCTTATTTCGTGATCACCAAGTCCTTGTAATGGGGTATCCAGGAAT TAGATTGCCAAAAGAGAGGTGA

3. Multiple sequence alignment (ClustalW) of Prp8 protein homologues from unicellular eukaryotes

----- Majority												
	10	20	30	40	50	60	70	80	90	100		
1	-----										Paramecium tetraurelia	
1	-----										Dictyostelium discoideum	
1	MSHNGSFEQSS	EDNKNEGSDVLT	NSTQHL	ENNV	INNYDDANKS	DELNSSHNV	MNDKAS	VENKQDN	MCCNN	INDIFFDKPDN	INNNNNNNEKNNMNDINNIP	Plasmodium falciparum
1	-----										Cryptosporidium parvum	
1	-----										Trichomonas vaginalis	
1	-----										Trypanosoma brucei	
1	-----										Encephalitozoon cuniculi	
1	-----										Giardia intestinalis	

----- Majority																			
	110	120	130	140	150	160	170	180	190	200									
1	-----										Paramecium tetraurelia								
1	-----										Dictyostelium discoideum								
101	QNV	PNGF	INN	IGNIP	YNNMNA	FPPNMP	KLP	TNMP	FLPPNMP	ILPPHL	QHMPNVLPHL	QNPVPPHLAS	FNP	MINL	PNL	PPHMHNL	PPNMHSL	PPHMHNL	Plasmodium falciparum
1	-----										Cryptosporidium parvum								

1 ----- Trichomonas vaginalis
 1 ----- Trypanosoma brucei
 1 ----- Encephalitozoon cuniculi
 1 ----- Giardia intestinalis

----- Majority

-----+
 210 220 230 240 250 260 270 280 290 300
 -----+

1 ----- Paramaecium tetraurelia
 1 ----- Dictyostelium discoideum
 201 PPNMHS LPPNMNYIPPGINNYMPNMMNMPPPYMMKMPNMKMSNKIINNVSNNVADNVRNSNLYNEEGIQPNNIHNNIHNNDHGGQDINSPPYLSQG Plasmodium falciparum
 1 ----- Cryptosporidium parvum
 1 ----- Trichomonas vaginalis
 1 ----- Trypanosoma brucei
 1 ----- Encephalitozoon cuniculi
 1 ----- Giardia intestinalis

-----N--MNESNXSGHLGEKIAKWKQLNKKKYLEKKKFG L Majority

-----+
 310 320 330 340 350 360 370 380 390 400
 -----+

1 -----MQLKPRLLQSVQPTL Paramaecium tetraurelia
 1 -----MDFNNALQNEQASQINLGIKHLNWSKLNARYCRRSKVNG Dictyostelium discoideum
 301 SYLPNNIKMNNNEIDQLEVNGLSLSSPFNEQQKKKDMNNKKNKAKKYHDFEGDEENYNTSERDENS MYDSNAFSIIEKARKWKMLNSKKYSKSKKFGV Plasmodium falciparum
 1 -----MNLIRQSDTIEDLKKWQVQKKKYAEKRFKGF Cryptosporidium parvum
 1 -----MKSGGGISLTEDGFEHQYNGEGVTEKWSSGHRRIAERWNLNTRKRYGYRATYQE Trichomonas vaginalis
 1 -----MDDTNSNINQSNESQHLEEKAKKWIQLNKKYSEKRFKGA Trypanosoma brucei
 1 -----MLPYDPRVNSRIQLSIVQDILGHSGNPIYSLDVSDIPVMLGNLARTL Encephalitozoon cuniculi
 1 -----METGDTGVLQERIAQWKHLRKKHFKLEIKTT Giardia intestinalis

VEG-EKEELPPEHLR KIVKDHGDMSSKKFRADKRVYLGALKYMPHAILKLENNMPMPWEQVRYVKVLYHITGAITFVNEIPTVIEPLYIAQWGTMMWMMR Majority

-----+
 410 420 430 440 450 460 470 480 490 500
 -----+

17 VGSTENSSFP-----SQPRSMARKLAAMGSLKFLVYAIRKALETMPQPWEAVRYVTVAHQKAGALTYILSKSTSSEHDLIRQWTRVCD SIP Paramaecium tetraurelia
 41 NQIVRKSKLPPEHLRKLINIHGDMSSRKFQNEKRVYLGALKYIPHSIFKLEENIPMPWERVKYVDCLYHTTGAITFVNEIPWVIEPIYIAQWSTMWMTMR Dictyostelium discoideum
 401 VE--EKEEMPCEHLR KIVKEHGDMSNKKYRYDKRVYLGALKYIPHAVFKLEENIPMPWEQIKNTKVIYHITGAITFVNETFVVIDPLYIAQWGTMMWIMMR Plasmodium falciparum

34 VEG-QKEPQPPEILRKIFKDHGNGLESKKYRQDKRVYLGALKYMPHAIYKLLNMPMPWEQVTRTVKVLYHITGSITFCYEIPKVIEPVYTAQWGTMWVMMR *Cryptosporidium parvum*
 55 AVA-QKDEVPPPEYLRKLVKDNGLSGKRFAERKLCVALLRYMPLALYKLLNMPMPWEEARYVNVVYHMRGVLTLVEDTPTAPEPLYLAQWGSIWTKMR *Trichomonas vaginalis*
 41 VEI-RKEDMPPEHLRKI IKDHGMSNRRFRDDKRVYLGALKYMPHAILKLLNIPMPWEQVKYVKVLYHLSGAITFVNEIPFVIEPIYIAQWATMWVTMR *Trypanosoma brucei*
 48 HVLSLQKPIPQQHLLRVQSPCSVKQVLTQVDRRQISMFPVRVEDLLVNLPSVWMDVS-IHCCDVPCHR-GLVCISIDVHCSGDNQLFCCCGLCSSPHA *Encephalitozoon cuniculi*
 33 SKGPSAKELPPGHIRQIMTSHGNMSHDKFAGQKRLYIGALKYAPHAVLKFLNMPMPWEEELRKRVRVLYHTAGALTFVNEVPRVVPQYLAQWAETWIAMR *Giardia intestinalis*

REKDRKHFK---RMRFPFDDEEPLDYSNILDVEPLEPIQMELEDEDEDAVIDWFDYDSKPLLYTK--ING--SYRKWKLTLVEMGTLFRLASPLLSI Majority

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
 510 520 530 540 550 560 570 580 590 600
 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

103 KDAAPPS-----YPIFDGTAPYMCYRDNLAFARPLPLFKTN-----AIIRFHRLKKNLCADLSATQKI *Paramecium tetraurelia*
 141 REKDRKHFK---RIRFPFDDEEPPIDYCNILGVEPLDSIQIVLDGEEDISVYDWFYLDKQKQFYFHKKGNYKKHQWHLTFEQLGTYRLSMQILPI *Dictyostelium discoideum*
 499 REKDRKHFK---RMRFPFDDEEPLDYADNILDIEPLECTIQMCLKDKEDKSVIOWFDYDSKPLLYNRNHIPG-TSYKKYKLSLEQMGVLYRLGNQLFSD *Plasmodium falciparum*
 133 REKDRRNFK---RMRFPFDDEEIPLDYGDNILDVEPLEPIQMELEDEREDNAVDFWFDYDQPLRYTK-LLNG-PSYRSWQLTLEVQQNLFRLANQLLSD *Cryptosporidium parvum*
 154 SHKVELQQECGTFRRVISKGNENEPIDFSYIMDREPPPALYDDLDEEADAALDWFYDPPRRLVHPNQRGSRPENGYYFTIDVIEITLFRNAIPILPN *Trichomonas vaginalis*
 140 REKDRTHFR---RMKFPLDDEEPLDYSNILDNEVEDEPIQMELENDSEVIDWLYDSKPLVNTK-FVNG-SSYRKWRLNLPIMSTLFRLASPLLSI *Trypanosoma brucei*
 146 FLLRERYSD-----S---KSLECLLKKSIWVLGVCVPMPPNIRKERMPPEHLRRIVRTSRDMHPSF---MG---AMRYMPHALHNLRSMPMPWESI *Encephalitozoon cuniculi*
 133 REKDRHQIR---RLRFPFDDEEQLDFVTNIEGVEPPEAILMELDEDEDAVYEWFYDYLGLPSQY--ING-LSYRKWKLPLTVMSTLYRLARPLVDQ *Giardia intestinalis*

LLDDNYFYFLDLSFFTAKALNMAIPGGPKFEPLSRDVEDDEDEDWNEFNDINKVIIRDDIR-----TEYKIAFPFLYNSRPRKVAVAPYHYPANVFIK Majority

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
 610 620 630 640 650 660 670 680 690 700
 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

160 LLSDSRSLKSRDKFSFEQHLF---S-VFTNLSFFHGDDN----FLMYQHVVYLYHAYT-----EDQLLTYPALYYPIS-----MH- *Paramecium tetraurelia*
 238 LLENNSYFLFNKDAFFTAKALNIAIPGGPKFEPLNSTSFDEDEDWNEFNDLN RVIFRSITR-----SEYKIAFPHFYNSLPKFVSTSVYHYIVNIFTK *Dictyostelium discoideum*
 595 FQDDNYFYFLNLSFYTAKALNMAIPGGPKFEPLYRDIYEDEDWNEFNDINKIIRQQIR-----TEYKIAFPFLYNSRPRKIAVSKYHSPMCVYIK *Plasmodium falciparum*
 228 IVDHNYFYFLFNFPFYTAKALNMAIPGGPKFEPLYRDIYEDEDWNEFNDINKIIRQQIR-----SEYKIAFPFLYNSRPRKIAVSKYHSPMCVYIK *Cryptosporidium parvum*
 254 LDDRNYYLWDLKSFYAAMHIAIPRAPKFEAPSTIQEEEGE-WTEFNDLRRVHRDDPRKPRFTMLTERQIAFPFLYSDVVDGVTVAPYRPAQIRVE *Trichomonas vaginalis*
 235 LTDSNYFYFLDSDNSFFTSKALNMAIPGGPKFEPLFRDVEDDEDEDWNEFNDINKVIIRNKIR-----TEYKIAFPFLYNSRPRKVKTPYHTPNNCYIK *Trypanosoma brucei*
 230 RYVDVVYHVSGLITFVAEKRRREDAEYKRRWSRVSASFSSQTQKRLVRIILRYPVFDDDDPVVDYG--SIIGLPVPTAVECEGSEECRDVFLDEDERYLF *Encephalitozoon cuniculi*
 227 YEDPNSKYLIDLPSLFTSKALNEVIPPGRFPEPLFTDVPDNPQE-WTEFNDINKIIRTPIS-----TEWKIAYPNLYNRRPKISIAPIHYPLSFCFAK *Giardia intestinalis*

NEDPDLPPFNFGPXLNPIPSY----- Majority

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
 710 720 730 740 750 760 770 780 790 800
 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

227 -----PPANLPIFLSPSPSA----- *Paramecium tetraurelia*
 331 PENPNSPIFEFNEYHPIPSN----- *Dictyostelium discoideum*

688 LEDIDLPPFYFDLIINPIPSYKIRKFNKSSEKKDSELFDDDFYLTYTRKEIYYYDHGDDDDKKKSTSKSRKSHKSHSDADDNRYDKGYRKYRKSSSSYKSF Plasmodium falciparum
 321 QDNPEIPTYNFDPVINPISAY----- Cryptosporidium parvum
 353 NEDPAVPCFSWNPSLNPIKAIQKRHSDPVG----- Trichomonas vaginalis
 328 NDSPDLPGFYFGAALNPIPSY----- Trypanosoma brucei
 328 NCKTFLISRHLGVQLDNGPYVG----- Encephalitozoon cuniculi
 319 YNTIITPVFQLAPNLSSISRP----- Giardia intestinalis

-----TSGNKNEVQIXDEELD----- Majority

	810	820	830	840	850	860	870	880	890	900		
242	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	Paramecium tetraurelia
352	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	Dictyostelium discoideum
788	KRDKRKSTNSSNDKDI	DEEDYNSGVSSID	NNDNSDTYISS	SKYNSNMSSRT	SKNKDETYEID	STVENDSHD	GSGLKKEKN	KKRKNPYND	DNYKGD	DKNK	-----	Plasmodium falciparum
342	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	Cryptosporidium parvum
383	-----S-----	-----	-----SS-----	-----	-----	-----	-----	-----	-----	-----	-----	Trichomonas vaginalis
349	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	Trypanosoma brucei
350	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	Encephalitozoon cuniculi
340	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	Giardia intestinalis

-----IXI-EGFLPLLHETELETERTINGIALLWAPXPFNERSGKTRRA-XDIPLVKSWEYKEH---ISSD Majority

	910	920	930	940	950	960	970	980	990	1000		
242	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	Paramecium tetraurelia
369	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	Dictyostelium discoideum
888	SDDDDNYKGD	NDNNDNNKY	SDNIS	SCKKNKMI	IKHVEYGIL	PLLNHPY	TERTINGIQ	LHYAPY	PFNKKCGY	TRRG-	IDIPLVQ	SWFKEH---ISTK Plasmodium falciparum
359	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	Cryptosporidium parvum
408	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	Trichomonas vaginalis
367	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	Trypanosoma brucei
366	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	Encephalitozoon cuniculi
356	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	Giardia intestinalis

QPVKVRVSYQKLLKNWVLSLHHRKPKSQTKR-----NLLKILKNTKFFQSTEIDWVEAGLQVCRQGYNMLNLLIHRKNLNYLHLDYFNFLKP Majority

	1010	1020	1030	1040	1050	1060	1070	1080	1090	1100		
289	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	Paramecium tetraurelia

429 QPVKIRISQQKLLKNWILNSLHNVAKNCKRR-----NFLKILQNTKFFQSTEMDWVEVGLQVCKQGYNMLNLLIHRKNLTYLHLDYFNFLKP Dictyostelium discoideum
 984 YPVKVRVSYQKLLKCWVLNHLHNSKRPKSMKKK-----YLFRIKSTKFFQCTEMDWVEVGLQVCRQGYNMLNLLIHRKNLNYLHLDYFNFLKP Plasmodium falciparum
 420 YPVKVRVSYQKLLKCWVLNLSLHHRKPKWQNKR-----NLLKAFQATKFFQLTEIDWVEEGLQIARQGYNMLNLLIHRKNLNYLHLDYFNFLKP Cryptosporidium parvum
 475 TRDKILRSYQQLKHHVAKNLRDRQKERPKKEGGNQDEGGQPVRRLDELANLDFHKTIDWLEAGLQVMRQGHNMLVQLINVKSLPYVHINYNFQAKP Trichomonas vaginalis
 428 HPVKVRVSYQKLLKCHVLNKLHHRKPKAQTKR-----NLFKSLKATKFFQSTEIDWVEAGLQVCRQGYNMLNLLIHRKNLNYLHLDYFNFLKP Trypanosoma brucei
 427 ---GGPRVGNLLRNYARNMQKTRKRPT-----HILKELKNTRYFORTEIDWVEAGLQVYQGHRLMSEVLRRKLSYLVLDWVFNFLKP Encephalitozoon cuniculi
 419 QPVKVRVSYQKLLKNVLRNLSHHRKQYVVRRRK-----TITKIFKTPYFQSTTLDWVEAGLQVVQGYNMCNLLIKRHRVFLHLDYFNFLKP Giardia intestinalis

IKTLTTKERKKSFRGNFHLCREILRLTKLLVDSHVYRLGNIDAFQLADGIQYVFSHVGLLTGMRYKYRMLRQIRMCKDLKHLIYYRFNTGSVKGKPG Majority

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
 1110 1120 1130 1140 1150 1160 1170 1180 1190 1200
 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

368 TKVLSTKERRKQSRFGKAFHLVRELTKFLKYMVDTHIAHRLLLIDYPTFLTGIHHLFLNIGTVTSVYRYKYKISNQIRQLKALGMLCGDVP----- Paramecium tetraurelia
 517 IKTLTTKERKKSFRGNFHLCREVLRRLTKLVVDCHVYRLGNIDAFQLADGLQYIFNHVGLTGIYRYKYRIMRQVRMCKDIKHIVYYRFNTGSVKGKPG Dictyostelium discoideum
 1072 VKTLTTKERKKSFRGNFHLCREILRLTKSIVDSHVYRLGNIDAFQLADGIQYIFAHVGLTGMRYKYRMLRQVRMCKDLKHLIYYRFNTGSVKGKPG Plasmodium falciparum
 508 VKTLTTKERKKSFRGNFHLCREILRLMKLACDSHVYRLRNIDAFQLADGLQYVFSHVGLVTGMRYKYRMLRQIRMCKDLKHLIYYRFNTGSPVKGKPG Cryptosporidium parvum
 575 TRTLTTKEIKKSRLGPAFHLIRELLGFMKQLIDMHTMYRLGKNDSIQLADAIQYLFSHLGLRGTGVYRYKLRAMRQIKRSRDLKHLVLYSKFNVGEVLRGPG Trichomonas vaginalis
 516 IKTLTTKERKKSFRGNFHLCREILRLTKLVVDVHVFKREAKGDADAFQLADAIQYLFSHLGLLTGMRYKYRMLRQIRMCKDLKHLIYYRFNTGAVGKPG Trypanosoma brucei
 509 IRQLTTKERKKSFRVGTSYHLTREMLKFIKHLVDIHVLFRRQGHIDCYELMGNVGHVNLNVGLTGIYRYKYKLMKQIKRCKSWKRLSDYARTEG----- Encephalitozoon cuniculi
 508 IKTLTNKERRKSRFGNAYHLMREFFRFTKLLLDCHIYRLGQIDAVLADALQYVFSHAGHLTGMRYKYKLMHQVRTCKDLKHLVLYSRFNTGEVKGKPG Giardia intestinalis

VGFWAPMWRVWVFFLRGIIPLLERWLGNLLARQFEGRDSKGI--KTVTKQRVESHFDVELRAAVMHDILDMPEGVRANK--ARTILQHLSEAWRCWKA Majority

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
 1210 1220 1230 1240 1250 1260 1270 1280 1290 1300
 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

458 --FYHPLQCVMSNLYRGLSPLEAYLSRLLARTAGIIDKDSGNSSRRVTHQRSLANQIVEQRNRYAGRFVSMYPQFTTRSA-MTKLFLAHLAEAWLCWRA Paramecium tetraurelia
 617 VGFWEPSWRIWIFFLRGIIPIILERWIGNLLSRQFTGRQNSNPY--RSISKQRVESHFDLELRASVMHDILDMPEQLRSSK--ARPILQHLSEAWRCWKA Dictyostelium discoideum
 1172 CGLWAPLWRVWVFFLRGVIPLLERWLSNLLARQFEGRVSKGIA--KTVTKQRVESHFDLELRASVMHDIIDMIPEGLKNNKGGARLILQHLSEAWRCWKA Plasmodium falciparum
 608 VGFWTPMWRVWVFFLRGIIPLLERWIGNLLARTFEGRHSGKIS--KTVTKQRVESHFDLELRASVMHDILDMPEGVRANK--AQTLQHLSEAWRCWKA Cryptosporidium parvum
 675 CGFWAPSWRVWVFFLRGIMTPLLQRYLGNLTDVLRGREGAKGHDGKRITRQVETDKDVNIKEAFRRELREMLPPDVRTEV--IRTMDQHMNEAFRHWR Trichomonas vaginalis
 616 CGFWAPMWRVWVFFLRGIVPLLERWLGNLLARQFEGRQTKGMA--KTVTKQRVESHFDYELRAAVMHDILDMPEGKANK--SRIILQHLSEAWRCWKS Trypanosoma brucei
 602 ---WGEQWRTWCFMFRGHIPLLRGRIISGLVTRIAEGRDYNPKP---LSKQSRSESGYDVALKRQIMAEASSILHPTQ-----VRRLLQHFGEAWRCWKA Encephalitozoon cuniculi
 608 VGFWGPMWRVWVFFMRGSIPLMERWLSRVAREYEGFRSKRLP--STVTKQRVESNYDIELRASVLHDITDTMPEGIRNAK--AHTVLAHMSEAWRCWKA Giardia intestinalis

NIPWKVP-----GLPPPVENIILRYVKLKADWWTN-----SAYYNRE----- Majority

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
 1310 1320 1330 1340 1350 1360 1370 1380 1390 1400
 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

555 GMAYDQVYS-----QMSPEVADLVQAYVSERADLYTA----SIACTKK----- Paramaecium tetraurelia
 713 NLPWKVH-----GMPPAVENIILRYVKLKADWWTN----SAYFNRE----- Dictyostelium discoideum
 1270 NIPWKVV-----GLPLPVENIIRYIKLKADWVVN----ATYYNRE----- Plasmodium falciparum
 704 NIPWKVP-----GLPAPIENIILRYVKYKADYYTN----SAYYNRE----- Cryptosporidium parvum
 773 GLRWSVP-----GLAKPLTDLVNKYVKLRAEEYVR----VTQYQRK----- Trichomonas vaginalis
 712 NIPWKVP-----GLPIPIENMILRYVKSADWWTN----IAHYNRE----- Trypanosoma brucei
 689 NVPYHIVLEHIKALEVKRSGSVSLEQRGLRGTAETTTVSLEALKAMTDVARQSSGNAFISSEQLITFSDGFVSKMEEMSRIAGVGHKLKELFELQRIIDK Encephalitozoon cuniculi
 704 NIPWKVP-----GLPPPLEAIILRYVKAKADWWTK----NAHYARE----- Giardia intestinalis

-----RIKRGATVD-----KTVCKKNLGRLTRLYLKAQEQRQHNYLK-----DGPYLSPEEAVAIYTTAVHWLESRGFHIPPPLNYKHDTKLLILA Majority

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
 1410 1420 1430 1440 1450 1460 1470 1480 1490 1500
 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
 594 --RIASKNKWIA-----KSEHYKYCGRAGRQDMRELIVANAAYLCEPIQKDLRVSLGTVYSLAYLCITVAARVCACGSHIPFPSQEFYDGLKLELELA Paramaecium tetraurelia
 750 --RIRRGATVD-----KTVVKKNLGRLTRLYIKAEQERQISYK-----DGPYISSEEAVALYTTAVHWLESRRFIHIPPPLNYKHDTKLLILA Dictyostelium discoideum
 1307 --RIKRGATVD-----KTVCKKNLGRLTRLWLKAEQERQHEYLK-----DGPYVSGEEAVALYTTAIHWLESRKFTHIPPPPLNYKHDTKLLILA Plasmodium falciparum
 741 --RIRRGATVD-----KTVCKKNLGRLTRFLKQEQERQHNFMK-----DGPYLTTEDEAVAIYTTALVRWLESRKFIHIPPPVNYKHDTKLFLLA Cryptosporidium parvum
 810 --RINEGDTV-----KQAFMKNLGRLTRLKLMEEQNRQRSYMEG-----TDTDIITPEQATEIYRMMANWLESDRGFKKISFPKASRPAELRLELELA Trichomonas vaginalis
 749 --RIKRGATID-----KTASKKNLGRLTRLWLKAEQERQHNYLK-----DGPYVSAEEAVALYTTVHWLEKRRFSAIPFPQTSYKHDIKILTLA Trypanosoma brucei
 789 YVRLKSEWYVDSAVGVGKSKKEKKRLGKITRLYMKERMAEQVEYLG-----LPFLRPEEAVAIYRLSAEYFRSKGTGRIPFP--EKNEERFLHIA Encephalitozoon cuniculi
 741 --RIARDGTVD-----KAITRKNTRGLTRLYLKQOSDYQANYLK-----EGPYITPEQGVAMLTMTQNWLEMRQFTPIPPPPMQYKHDTKMLILA Giardia intestinalis

LERLKEAYSVKSRNLNQSOREELGLIEQAYDNPHELTLSRIKRHLLTQRTFKEVGFIEFMDLYTHLVPVYEVDPLEKITDAYLDQYLWYEA-DKRNLFPNWVK Majority

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
 1510 1520 1530 1540 1550 1560 1570 1580 1590 1600
 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
 683 LRDLREDVLSGSILTVADRQLLTLIEKATSLPHEFLIRIKEILLKKRTFDDAVQIEYAEARTCVYPIYLTGLTRVVDVYFTYYLSYQITSSPLHYLLFKR Paramaecium tetraurelia
 833 LERLKEVYSVKSRNLNRSQREELTLIENAYDNPHELTARIKRHLLTQRTFKEVGFIEFMDLYSHLIPVYEIDPLEKITDYLQYLWYEA-DNRKLFPNWVK Dictyostelium discoideum
 1390 LEKLETFFTVKNRNLNQSOREELGFIEQAYDNPYETLSRIKRHLLTQRAFKEISISFLDLYTHLVPVYEVDPLEKITDAYLDQYLWYEG-DLRNLFPNWVK Plasmodium falciparum
 824 LERLKEAYSVKSRNLNQSOREELALIEQAYDNPHEALSrvKRHLLTQRVFKVRLVFMDFYSHLVPVYDVEPLEKITDAYLDQYLWYEA-DKRRLFPNWIK Cryptosporidium parvum
 895 LNRLRDQHNIANRLTQAQREEQARIEEAFNSPHETLSKIVDCLARVRRFKNVEVEYMDTFSSLYPIYNNVVPSEKLVDSFLDQYLWYEAAMDQQLFPNWVK Trichomonas vaginalis
 832 LERLKEAYSVKSRNLNQSOREELSLVEQAYDNPHEALRIKRHLLTQRTFKEVGFIEFMDMYTHLVPYIDVDFPEKITDAYLDQYLWYEA-DKRQLFPNWVK Trypanosoma brucei
 879 VDRLLKKS-----ATAEESFLDKALEESADTIFRIKKSLLTQRSFKEVGVTLKRHGDAIECYHVSHVERLVDAFLCAYLFYES-DRLNIFPEFIK Encephalitozoon cuniculi
 824 LENMRFGHDVSMRMNQTLEELGLIENAHDPHEALIRIKRDFMTARAFREVKFTFLEHYTRVIPNYEIIYALEKMTDAYLDQYLWYEA-DRRHLPFPWVQ Giardia intestinalis

-----PSDNEPPLLVKWCQGINNLDGIWDTSDGECVVLETTQFEK--IYEKIDLTLNRLRLIIVDHNIAIYITSKNNVVITFKDMNYTNSVGV Majority

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
 1610 1620 1630 1640 1650 1660 1670 1680 1690 1700

```

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
783 GFSSIDLSPYSMELPAELTIRYCKHVHSTCSGLPATDGEFCFLVHMHMLNTDSYFRGNLHVIGKVISLLFDPVISSFLITRLSSSFYFKDMTYTAVRGV Paramecium tetraurelia
932 -----PSDNEPPLLVIKLCNGINNLDFGFWKFDSDSVGLLVETQFEQ--IMEKIDLTLLNRLRLITDHNIAADYITSKNNINVTYKDMNLYNSYGI Dictyostelium discoideum
1489 -----PSDNEPPLLVIKMCQGINNLHNIWDTKNNECVVMLOQFSK--IYEKIDLTLLNRLRLIVDHNIAADYITAKNNTNITFKDMNHINSFGI Plasmodium falciparum
923 -----PSDSEPPPLLVIKWCQGINNLHGIWVSDGQCVVLLSEKFEK--VYEKIDQTLNRLRLIVDHNIAADYMTAKNNVVIPIYKDMNHTNVHGV Cryptosporidium parvum
995 -----PSDLEPVPILVIKWCQGINNDSPGIWDFDRDESIVLLHAKLEDD-FYGNIDWNLFRLLELIMDKSLAEYIVSRHDVVVEFKDMAYHCRKGM Trichomonas vaginalis
931 -----PSDNEPPLLVIKWCQGINNLQVWETSQGEVLLLETQFSK--VYEKMDLTLMNRLRLIVDQNIADYMSGKNNVINYKDMNHTNSYGL Trypanosoma brucei
969 -----PGD-EMNRTLMDFCDKISS---FKAADERLVLYEGKYEG--IMRMVDNLLSKLLKLVDPALADYIISRNNCKVVYKDMVYTNHVGFE Encephalitozoon cuniculi
923 -----PSDLIPPPVLVHKWCERINSLVDAWNTEGQTMVLVETSLEK--FYEQIDLTFNLNYMLRLVVDHNLADYMTSKNNVKISFKDMSYLVNGVGL Giardia intestinalis

```

IRGLQFSSFVAQYYGLIVDLLLILGLTRAQDIAGPPN-PNSFLTFS-SVQIEIRHPIRLYCRYVDKIYILFKFTAEEAKDLIQRYSLENDPDP-----NNEN Majority

```

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
1710 1720 1730 1740 1750 1760 1770 1780 1790 1800
-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
883 APSFQFSHLLTLLSILDLTILLHYDEQPFSS----PTIILR---VVTFIIAFAKDYRQSGVPLSQLWLRKACLEKGDLSRLSIFYG-----TYK Paramecium tetraurelia
1021 IRGLQFSSFVCQYYLLIVDLLLGLTRANQIAGSPSPHNEFLKYS-DKKIELSHPIRLYCRFVDKLYIIKLSKQEIKEIVQRYLSENPDFS----NNQN Dictyostelium discoideum
1578 IRGLQFSSFVQYYTIIIDLLLGLTRAYDIAGPYNDVNQFLTFQ-NVQIETRHPIRLYCRYVDKIWILFKFTNEESKDLQKFLTENPDP-----NNEN Plasmodium falciparum
1012 LHGIQFTSFIMQFYGMVLDLLILGLNRAQDLAGPYNNPHEFMTYS-NIQQEIRHPIRLYCRYIDKIFMVFRFTQEEARELILQRYLSENPDP-----NNEN Cryptosporidium parvum
1085 LRGMFSSFLAQYWGVLVIDVLLGLTQRSQEIAGPARRPNPFMSWMRDPLLATSHPIRGYCRYKNEVYVLLKYTKVEADDVRHRYLEETKNDPQKRAENAS Trichomonas vaginalis
1020 IRGLQFASFIFQYYGLVLDLVLGLERASALAGPPNPNPNSFLTFF-SVQTEAHPIRLYSRYVDRIVLVLYKFTADEARKLIQKYMSEHPDP-----NNEN Trypanosoma brucei
1053 IKGLQLSSFYKYSFIVDLCVLG-ED-----VFMDKSRKWFYFRHMDDIYIVFRLQRKEEDSLEDYGREAEER--MEEAMNER Encephalitozoon cuniculi
1012 IHGLQFTSFIAQYMGLLVLDLILGLRRANEMCGPPSMPNSLQFA-SIEDEIRHPIRMYQRYATRILHILYKFNAEQARDLIRDYCDVNSN-----NNDE Giardia intestinalis

```

VVGYNNKCKWRPDCR-----MRLMKHDVNLGRAVFEIQNRLRSLTTLDEWHS-----FVSVYSKDNPNLLFSMAGFEVRILPKIR-- Majority

```

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
1810 1820 1830 1840 1850 1860 1870 1880 1890 1900
-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
967 LLSYVRQGESLYLVLQ-----EDRQANIESSKTTIMEQVMKSIPOSIACYISRQ-----SSFCDNQTFYFVSLVNQPIRFEFVG--- Paramecium tetraurelia
1116 LIGYNNKCKWPRECR-----MRLVKNQDVIIGKSVYWELSNRLLPKSITTELEWERS-----FVSVYSKSNPNLLFSLAGFSVRILPCTCRIG Dictyostelium discoideum
1672 IVGYNNKTCWRPDCR-----MRMKHDVNLGRATFWEIQNRIIPRSLTSLDWDHYN-----TFVSVYSKDNPNLLFSIAGFEVRILPKIRQL Plasmodium falciparum
1106 IVGYNNKCKWPKDCR-----MRLMKHDVNLGRAVFNWIKNRLRCLTTLAWEHS-----FVSVYSKDNPNLFLNMGCFEVRILPKIR-- Cryptosporidium parvum
1185 VYGFKNFKQWPRDAR-----MRLFLNDVNLARAVIWEFRGRLPPGIADINESNA-----LASVYSKDNPNLLFDMGGFVSRILPVVR-- Trichomonas vaginalis
1114 VVGYNNKCKWRPDCR-----MRLMKHDVNLGRAVFWQIKNRLRSLTTLIDWEDS-----FVSVYSKDNPNLLMNMAGFDIRILPKCR-- Trypanosoma brucei
1130 RRYNYFDG-AWRDACDEESGPLFSNYGRREEGKGEILRRCIHAEVATRMLPSLGRIRFRG-----CSIF---PFVKFSMVGVDVLISKK--- Encephalitozoon cuniculi
1105 MLGYNNKTCWPKDAR-----MRLIKHDVNLGRAVFDLQNLRLRSLCEVNWNSSENSASGFHQSFASVYSKDNPNLLFYMCGFEVRILPKIR-- Giardia intestinalis

```

-----GTEDEILEK----- Majority

	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000		
1041	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Paramecium tetraurelia
1195	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Dictyostelium discoideum
1753	SYGYNGIMYTSYMNEYPRGVGTKDETSKKNGLLHDDKESKKVGLKDEVTKGKSHVDKNEENSDDNNKNDKNDSTHANTHDMVGDNNYDGGVKNFYNSS										-----	Plasmodium falciparum
1183	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Cryptosporidium parvum
1262	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Trichomonas vaginalis
1191	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Trypanosoma brucei
1210	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Encephalitozoon cuniculi
1192	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Giardia intestinalis

-----	ESTWKLQNE	-TKEITAXAFLRVSEESIENFENRVRQILMSSGSTTFTKIANKNWNTALIGLVTYTYREAVVDTEELLDLLVRCEENKIQT	Majority
-------	-----------	---	----------

	2010	2020	2030	2040	2050	2060	2070	2080	2090	2100		
1047	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Paramecium tetraurelia
1217	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Dictyostelium discoideum
1853	GGEKNVVVSSSVKEGTWKLNEMTKEITAEAYLKVSDNSMKRFENRVRQILMSSGSTTFTKIANKNWNTCLIGLMTYFREAIVYTEKLLDLDLLVRCEENKIQT										-----	Plasmodium falciparum
1192	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Cryptosporidium parvum
1270	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Trichomonas vaginalis
1200	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Trypanosoma brucei
1214	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Encephalitozoon cuniculi
1201	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Giardia intestinalis

RIKIGLNSKMPNRFPPVVFYTPKELGGLGMLSMGHILIPQSDLRYSKQTDG-	ITHFRSGMSHDEDQLIPNLRYIQTWESEFI	-----	Majority
--	--------------------------------	-------	----------

	2110	2120	2130	2140	2150	2160	2170	2180	2190	2200		
1127	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Paramecium tetraurelia
1304	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Dictyostelium discoideum
1953	RIKIGLNSKMPNRFPPVVFYTPKELGGLGMLSMGHILIPESDLRYMKQTDNGRITHFRSGLSHEEDQLIPNLRYIQTWESEFI										-----	Plasmodium falciparum
1279	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Cryptosporidium parvum
1357	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Trichomonas vaginalis
1287	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Trypanosoma brucei
1294	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Encephalitozoon cuniculi
1288	-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										-----	Giardia intestinalis

-----	DSQRVWAEYALKRQEAQAQNKRL	-----	TLED	---	LEDSWDKG	-IPRINTLFQ	--	KDRHTLAYDKG	Majority
-------	-------------------------	-------	------	-----	----------	------------	----	-------------	----------

```

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
      2210      2220      2230      2240      2250      2260      2270      2280      2290      2300
-----+-----+-----+-----+-----+-----+-----+-----+-----+
1225 YNGRSPLFGAQTGFVYVEYDACGAEQQYSFKEISECLSSQWTGFAVSTFREAVFQALSQGSNSIEQQSPLEIAWANGCIPRLTTLIHYAKDLYCLLYRNP Paramecium tetraurelia
1388 -----ESQRVWLEYSLKRQQAQLQNKRL-----TLED---IEDSWDKG-IPRINTLFQ--KDRHTLAYDKG Dictyostelium discoideum
2037 -----ESQRVWCEYALKRNECHNQNKKI-----TLED---LEDSWDKG-IPRINTLFQ--KDRHTLAYDKG Plasmodium falciparum
1362 -----DSQRVWAEYALKRQEAQVQNRRL-----TLDD---LEDSWDHG-IPRINTLFQ--KDRHTLAYDKG Cryptosporidium parvum
1440 -----ESVKAWTEFNMRDREAKAAGTRL-----SIDD---IEHIINKG-VPRIRVLFS--RHAKLFQFDKG Trichomonas vaginalis
1370 -----DSQRVWAEYAIKYEAKSQNKNL-----TLED---LEDSWDRG-IPRINTLFQ--KSRHTLAYDKG Trypanosoma brucei
1363 -----ESNRVWKEYGRSGK---MEP-----DKG-IPRMSTLLQ--RSR-WLLYDRG Encephalitozoon cuniculi
1371 -----DSKRVDQHYTNMRKEAGALNKKI-----TIED---LDLWDRG-IPRINVLQF--RDRHTLAYDKG Giardia intestinalis

```

WRVVRQDFKQYQGLKNNPFWWTHQKHDGKGLWN--LNNYRTDMIQALGGVEGILEHTLFKGTYPFTWEGLFWEKASGFEE-SMKYKLLTNAQRSGLNQIPNR Majority

```

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
      2310      2320      2330      2340      2350      2360      2370      2380      2390      2400
-----+-----+-----+-----+-----+-----+-----+-----+-----+
1325 FLHHLAKGISIGNALTLKSWYNKLLGSLYD--LQGYKKIITAIFGGVEEILHHTLYPATDFSDYKSVVWSTATEHETGLAKRTNLTRARRQGLSQIPNR Paramecium tetraurelia
1443 WRIRQIFRQFQILRNNPFWWTHQKHDGKGLWN--LSNYRTDMIQALGGVESILEHTLFKGTYPFTWEGLFWEKSSGFEE-SMKYKLLTNAQRSGLNQIPNR Dictyostelium discoideum
2092 WRIRQLFKQYQI IKSNNPFWWTHQKHDGKGLWN--LNNYRTDMIQALGGVEGILEHTLFKGTYPFTWEGLFWEKASGFEE-SMKYKLLTNAQRSGLNQIPNR Plasmodium falciparum
1417 WRVVRQDFKQFQMLKQNPFWWTHQKHDGKGLWN--LNNYRTDMIQALGGVEGILEHTLFKGTYPFTWEGLFWEKASGFEE-SMRFKLLTHAQRSGLNQIPNR Cryptosporidium parvum
1495 FRCRMEFQRYLAGKYLKNWFFHQEHDGNICGVLERYVDTNIALGGVEAILEHSLFRGTGFPSEWEGIEFNRAGGFEN-SKKDKLAKQQRAGLANVQIPNR Trichomonas vaginalis
1425 WRVRTDWKQYQVLKNNPFWWTHQKHDGKGLWN--LNNYRTDI I QALGGVEGILEHTLFKGTYPFTWEGLFWEKASGFEE-SMKYKLLTNAQRSGLNQIPNR Trypanosoma brucei
1402 FRMISMFRYISG-KPDWFWFTDAKHDGKGLWS--MERFTLDTLEALGGVGGIADHTLFGATYFRSFKVFWED----MV-VEKYRKLTHAQRGMSQIPNR Encephalitozoon cuniculi
1426 WRTRLYFKKYSLFKTNPYAWTHHHHDGKGLWN--LKDYRADVIQALGGVEGILSHSIFKATGYKHWEGLFWDNTTGFEE-ALKYRKLTHAQRGMSQIPNR Giardia intestinalis

```

RFTLWWSPTINRANVYVGFQVQLDLTGIFMHGKLP TLKISLIQIFRAHLWQKIHESVVM DLCQVLDDELXDLDIETVQKETIHPKSYKMNSSCADILLF Majority

```

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
      2410      2420      2430      2440      2450      2460      2470      2480      2490      2500
-----+-----+-----+-----+-----+-----+-----+-----+-----+
1423 RFALWWSPTINRANVYVGFQVQLDLTGIFMHGKLP TLKTAIVSLFRGHAWPMIHS SLVKTL LAILQDAFRGLPLDIFKAESVHPKRSYHYHTSCADISVT Paramecium tetraurelia
1540 RFTLWWSPTINRANVYVGFQVQLDLTGIFMHGKLP TLKISLIQIFRAHLWQKIHESIVM D ICQVLDNEVDALGIEMVQKEAIHPKRSYKMNSSCADILL Dictyostelium discoideum
2189 RFTLWWSPTINRANVYVGFQVQLDLTGIFMHGKLP TLKISLIQIFRAHLWQKIHESIVM D ICQVFDLNCDDLDIETVQKETIHPKRSYKMNSSCADILL Plasmodium falciparum
1514 RFTLWWSPTINRANVYVGFQVQLDLTGIFMHGKLP TLKISLIQIFRAHLWQKIHESIVM D ICQVDFMEMETLEIETVQKETIHPKRSYKMNSSCADILL Cryptosporidium parvum
1594 RFALWWSPTINRANVYVGFQVQLDLTGIFMHGKLP TLKISLIQIFRAHLWQKIHESIVM D ICQVDFQELDNLEISVFNKEAIHPKRSYKMNSSCADILL Trichomonas vaginalis
1522 RFTLWWSPTINRANVYVGFQVQLDLTGIFMHGKLP TLKISLIQIFRAHLWQKIHESIVM D ICQVDFQELDNLEISVFNKEAIHPKRSYKMNSSCADILL Trypanosoma brucei
1494 RFTLWWSPTINRANVYVGFQVQLDLTGIFMHGKLP TLKISLIQIFRAHLWQKIHESIVM D ICQVDFQELDNLEISVFNKEAIHPKRSYKMNSSCADILL Encephalitozoon cuniculi
1523 RYTLWWSPTINRANVYVGFQVQLDLTGIFMHGKLP TLKISLIQIFRAHLWQKIHESIVM D ICQVDFQELDNLEISVFNKEAIHPKRSYKMNSSCADILL Giardia intestinalis

```


ATFKWPVS-KPSLLNDTK-----DTYGATTT-----KFWIDVQLRWGDYDSDH-IERY Majority										
	2510	2520	2530	2540	2550	2560	2570	2580	2590	2600
1523	WTRSLTVQ	QDYSIQIQ	KSSEPHHNS	QEESGAHASS	GEQ-----	VDSCTHL	WIDLHLT	WGNVD	TCTSLAKY	Paramecium tetraurelia
1640	SSYKQVAT	NPSLLLDK	-----	DDISSNSLIN	-----	TNKF	WIDIQL	RWGDYD	SDH-IERY	Dictyostelium discoideum
2289	ANYKKGIS	-KPSLLTDED	-----	HIFTNNTL	LGSTSGT	NNNIML	NSNMINS	GSNNSS	NNMNSV	Plasmodium falciparum
1614	AAFKWPIS	-KPSLIHDTK	-----	DTYDGT	TTT-----	KYWL	DVQL	RWGDYD	SDH-IERY	Cryptosporidium parvum
1694	STSRWPV	TSKPTVLS	DETG-----	DEYRAHT	TTT-----	KYWI	DVQL	RWGNYS	SHN-IAEY	Trichomonas vaginalis
1622	ATHKWQVS	-RPSLLNDR	-----	DTYDNT	TTT-----	QYWL	DVQL	KWGFDS	HD-IERY	Trypanosoma brucei
1588	GDFCVDSP	-ISILEERD	-----	GGSVRCS	-----	ELWI	DVQL	RWGDYD	KRN-PHKY	Encephalitozoon cuniculi
1623	SQNKWPST	-EPCFVNETK	-----	TFHGEFL	TT-----	RFWV	DIQL	RWGDYD	MHD-IERY	Giardia intestinalis

SRAKFLDYTSD-SQSIYPSPTGILIGVDLAYNLYSAYGNWVPG--LKPLIQKAMAKILK---SNPALYVLRERIRKGLQLYSSEPTPEP-YLNSQNYGEL Majority										
	2610	2620	2630	2640	2650	2660	2670	2680	2690	2700
1591	SKDRHKY	YTSDRSR	GIYRSPH	GIICIDLL	YREIAAY	GS-VPT--	IAIPAIN	KAISEL	LESLHS	NTMMN
1690	CRAKFLDY	TSD-AMSI	YPSPTG	VLIADV	LAYNLY	SAYGNW	IPG---	LKELIQ	KAMAK	IMK---
2377	SRAKFLDY	TTD-NLSI	YPCLTG	VLIIGV	DLAYNLY	SAYGNW	FNN---	LKPLMQ	KALQK	IVQ---
1661	ARAKFLDY	TTD-NISI	YPSPTG	MVAIDL	LAYNLH	SGYG--	YMGSP	NSKNL	IHQARN	KIMK---
1743	TRSRFYE	YSS---	AKMYP	FPAGI	VVAIDL	AYNCHS	AFGYW	VPR---	LKPLMM	KLMTA
1668	SRAKFLDY	TTD-SMS	LPSPTG	CLIGL	DLAYNI	YSSFG	NWFLG	---VK	PLVQ	KAMAK
1632	ARTRFVE	CTAD-PQ	ALYPK	NGFVV	LDLCY	NTWSG	YGNLNEE	---L	KTVL	KSSME
1670	TRSLFYA	YATSG-TQ	SMYPS	STGII	IGVDL	CYNEWT	AFGTW	IPG---	LQEL	IDKAM

FSN-QITWFVDDSNVYRVTIHKTFEGNLTTKPINGAIFIFNPATGQQLFLKIIHTSVWAGQKRLGQLAKWKTAEVVAALIRSLPVEEQPKQVIVTRKGMLD Majority										
	2710	2720	2730	2740	2750	2760	2770	2780	2790	2800
1688	FTG--KVI	IVDDSL	AYNFRML	NRDDTR	ASRVLI	INGFIS	IFNPQT	GRVLV	SVVHAD	TYAGQ
1782	FSN-QTWF	VDDTNV	YRVSIH	KTFEGN	LTTKPV	NGCIL	LILNPN	CNGKLF	FMKVI	HTSVW
2469	FSS-QTWF	VDDTNV	YRVTIH	KTFEGN	LTTKPI	NGAIF	ILNPK	TGQLFL	KIIHTS	VWIGQ
1754	FSN-QI	IWFVDD	TNVYRV	TIHKTF	EGNLTT	KPNNGA	IIIFNP	KTGQL	FLKVI	HTSVW
1833	FSEGMRT	WIVDD	SATVY	TSEQPT	AEGGRK	FRSENG	AVLIFE	PATGNL	KLSIV	HKS
1760	FSN-KIM	WFVDD	SNVYRV	TIHKTF	EGNLTT	KPINGA	IFIFNP	RTGQL	FLKII	HTDVW
1720	FTS-G-LL	VDKAL	LRK--EK	-----	TLFV	LDPA	SAGNLY	FKYS	----	GESK
1762	FGN-KIT	WIVED	KHVYR	VKIQK	TFEGN	YTTSP	VNGGV	FIMNP	ATGQL	FLKII

2007 AAPSEERQQMAKEVEEEE-----DKAALKTVTTATRDADGNQHIIQTFSQYEQQFKSKSDWRSRALTSRGLVMRANTLMIPPP-----VVKPK--- Giardia intestinalis

-----FTYVVPKNNLLKKFIEISDLKTQIGGFLFGISPPDNP-----VKEIRCIVMPPQIG Majority											
	3110	3120	3130	3140	3150	3160	3170	3180	3190	3200	
2067	-----FCRSIIDQAVGATTDNSSRLCVIIGQVTKSQP-----ILPCSGLOAPVALTFGLQHS										Paramecium tetraurelia
2123	-----IIPYIIPKNNLLKTFIEISDLRTQIGAFMYGKKIVESIGKGCYSENCEMETVIEIRCLVLAPQHG										Dictyostelium discoideum
2805	TISASASSHNILNKNGTNSDNQNSHYHTSINSINDYTYVIAKNLLEKFCISDLKIQVGGFLFGSSPEDNS-----YVKEIKCILIPPQIG										Plasmodium falciparum
2084	-----FTYVLPKNILKKFISADLKTQIAAYLYGISPPDNL-----QVKEIRAIVMIPQIG										Cryptosporidium parvum
2164	-----DQLIFPQELLKILFPFCFDVQAQFCAYLFGQTLDPDSP-----NVKEVLCIMVPPQKS										Trichomonas vaginalis
2091	-----GFTYVFPKNILKKFITIADLRTQIMGYCYGISPPDNP-----SVKEIRCIVMPPQWG										Trypanosoma brucei
2003	-----RIPLNLIIEGFMRLVDPHVLTFGLVIG-----G-----DILSFGMVPQFS										Encephalitozoon cuniculi
2090	-----LELIIPENIYRRFVEISDPYMQICGFLFGVKMNDTL-----QVISIVIPPQNG										Giardia intestinalis

SRVSVTMP SILPXS-----YLXGLEPLGWIHTQVTELS-----LSPR-DVKVHGRLLENK----- Majority											
	3210	3220	3230	3240	3250	3260	3270	3280	3290	3300	
2119	VTMNGKPRTLIELRHIRAYLGGWDFFSYAAIPITELLQDVKFTEEKTMKFYGVLVDFSR-----										Paramecium tetraurelia
2189	NQNSINMTDILPKNP---EISDLEFIGLIKTKVQEEFS---IAIS-DIDYLWRISQNNL-----										Dictyostelium discoideum
2891	NYQSVTLSSYMPSSK---YLQNLLELLGWIHTQTTNCSNTNNHLTAY-DMVAHFNFQECKRQMSKGGKVADASHNDDDDVDDYDDDDYNNNEDDYNNNED										Plasmodium falciparum
2135	SRDNVTMPHQMPDSE---YLRNLEPLGWLHTQSTETM---HLSTY-DITLHARLIQENQ-----										Cryptosporidium parvum
2215	SAVEYTPSCIPHDHPTLTENHLSLLGVLRCSGGEPS-----IHSR-DVAIHGRLLACNEG-----										Trichomonas vaginalis
2143	TPVHVTVPNQLPEHE---YKDLLEPLGWIHTQPTTELP---QLSPQ-DVITHSKIMSDNK-----										Trypanosoma brucei
2042	SLSGIHSSFLVPPGD-----IVGVVNG-----D---DLEVAGTLCERYK-----										Encephalitozoon cuniculi
2138	DRDEIDFKQILPNHD---FLDGASPIGF IHTRVGENS---SLEPR-DAKVLASLCKKNP-----										Giardia intestinalis

-----WDVDNTAIVVVSFTPGSCTITAYKLTHSGFSWAKN---NKDL--LNVKPFSTNHYEKVQIILLSDEFGLFFLVPDDGIWNYN Majority											
	3310	3320	3330	3340	3350	3360	3370	3380	3390	3400	
2179	-----AKLFPDKLLESSTLKKSSVDIISLVGRRTSVCLLTFYFGADRIEQSNFCTYYLQKQNTDKTVQKQFVSRVVRVHVASNYS										Paramecium tetraurelia
2241	-----DVQVDNIAMVSCSFTPGSTTLSSAIKLDHDCIDWYKNNIENKDLNLYETFKTCSDDYTEKIKLILSETYNGFFLIPEDGVWSYN										Dictyostelium discoideum
2987	DNINNNSEGGTKRDETYKMDKNKTIILTCSTFTPGSCTINAYKLTSDGYSFAKS---KKNSSDLYVFPNPNLYEPVQIILSNVFGYFLIPDDHIWNYN										Plasmodium falciparum
2187	-----SWDAERCIVQTVSFTPGSCTITAYKLTHQGFWEWGN---NKDL--NAVHPSSTQHFEKVQIILSDKFRGFFMVPDNDHMWNYN										Cryptosporidium parvum
2270	-----LQTEGLTIVVGVSVQDGIIRCYTTREGISWALEEYSHALQ--REPTVPLHVIPARVTLSTELQGFLLVPTDNGWNHT										Trichomonas vaginalis
2195	-----SWDGEKTVIISVSVAWP-CTLTAYHLTPSGFEWGN---NKDS--LNYQGYQPQFYEVQMLLSDRFLGFYMPDRGWSWNYN										Trypanosoma brucei

2079 -----IVD-PLAVLISDRIR-----VVKKSGCNWN-----EVH-----A-VLGKDLGVFVVP--EMWNYN Encephalitozoon cuniculi
 2190 -----KIDSDNFANVVISFPVGGCIMAATLSREGFEWAET---NIG-MDNPKDFDDNFAKVLGISITNEINGWMMAPENGIWNYS Giardia intestinalis

FMGVKLDXNTK--YGLIVDNPXKFYDEVHRPQHFLSFARLE----DEEDEADVENLFI					Majority
3410	3420	3430	3440	3450	
2259 LEFYCIDPITQCEKPGGIDQSVSYQSCIQFSKLFASQTQIEW-----EELECHVDRI					Paramecium tetraurelia
2323 SMVVKYNYSSK--CSYIVDKPNAFYDEVHRPQHFLQFAYLENIDDFDECGLLEIDEIFE					Dictyostelium discoideum
3084 LMGIKFNNNQK--YAPHLDIPQPFYADIHRPNHFLQFSLLD---QRDADEADVETSFI					Plasmodium falciparum
2264 FIGLGLVQQMK--YGLILSNPKDFYHEVHRSSHFIFIRNE--DKDQVDEADNEDFLS					Cryptosporidium parvum
2349 FRGATWREDDT--FDVRVDTPQFFFATHRPDHLNLFARLT--EEEATIDMADLENLMA					Trichomonas vaginalis
2271 FMGVKHSTNMT--YGLKLDYPKNFYDESHRPAHFQNWQMAPSANDDEENQPENENLFE					Trypanosoma brucei
2125 FARPFYDDRLE--YTWKIGMPHGFYDGFRCRVGHFSRFYQDR---AGGEEWQED					Encephalitozoon cuniculi
2267 FNSLRLQSVDPN-YPI SVQNPKTFFDMYHRVQHFTSFKREMNI---GEELSIDVDNNFI					Giardia intestinalis

4. Functional domains within the *S. cerevisiae* Prp8 protein:

>3_splice_site

```

                LGLNSKMP TRFPPAVFYT PKELGGLGMI
SASHILIPAS DLSWSKQTDI GITHFRAGMT HEDEKLIPTI FRYITTWENE
FLDSQRVWAE YATKRQEAIQ QNRRLAFEEL EGSWDRGIPR ISTLFQRDRH
TLAYDRGHRI RREFKQYSLE RNSPFWWTNS HHGKGLWNLN AYRTDVIQAL
GGIETILEHT LFKGTGFNSW EGLFWEKASG FEDSMQFKKL THAQRTGLSQ
IPNRRFTLWW SPTINRANVY VGFLVQLDLT GIFLHGKIPT LKISLIQIFR
AHLWQKIHES

```

>MPN

```

                EQ NVYVLPKNLL KKFIEISDVK
IQVAAFIYGM SAKDHPKVKE IKTVVLVPQL GHVGSVQISN IPDIGDLPDT
EGLELLGWIH TQTEELKFMA ASEVATHSKL FADKKRDCID ISIFSTPGSV
SLSAYNLTDE

```

>RRM

```

        EK IDFTLLNRLR RLIVDPNIAD YITAKNNVVI NFKDMSHVNK
YGLIRGLKFA SFIFQYYGLV IDLLLLGQER ATDLAGPANN PNEFMQFKSK
E

```

>U5_binding_1

```

                M PESIRQKKAR TILQHLSEAW RCWKANIPWD
VPGMPAPIKK IIERYIKSKA DAWVSAAHYN RERIKRGAVH EKTVMVKKNLG
RLTRLWIKNE QERQRQIQKN G

```

>U5_binding_2


```

: .*.*...:.*.*: .* .: :. **:*:* *****:*****:.*
3_splice_site      VGFLVQLDLTGIFLHGKIPTLKISLIQIFRAHLWQKIHES----- 288
Prp8_Giardia_2309  IGYRSQIDLGTGVYMCCKLATLKTAYVSLFRGHAWPMIHSS

MPN                -----EQNVY 5
Prp8_Giardia_2309  RGTSLHWQFQNIIRSMVICAEGQEGHIRGTVNEVIQLQQTSSISLSIDLYAQFRGTDEQTGL 1860
                                                                **.

MPN                VLPKNLLKKFIEIS---DVKIQVAAFIYGMSAKDHPKVKEIK---TVVLVPQLGHVGSVQ 59
Prp8_Giardia_2309  ILEGTPIQHFIAVLLMLQLANLSPLTIYKIIISQSNLELKKSLPEAELAYRPTYASLTTKS 1920
:* . :*** :      :      . ** : :. : :*:      :. * . : : .

MPN                ISNIPDIG-----DLPDTEGLELLGWIHTQTEELKFMAASEVATHSKLFADKKRD--- 109
Prp8_Giardia_2309  MERLSTIGNLTDVFLHLPRAPYQQWCPVISAMTERVINESAKKLGVRSDCLSPAEEKDLV 1980
:..:.* *      .** :      * : **.: :*.:.:.*. : : :.

MPN                -CIDISIFSTP-----GSVLSAYNLTDE----- 132
Prp8_Giardia_2309  LGAELVISNAPERRTSIFSRVLVLSANTLGERMLSSTQPKEVTDVRSQSRLSNLHFGEQWL 2040
: : * .:*      . : *** .* :.

RRM                -----EKI 3
Prp8_Giardia_2309  KRGFSSIDLNSPYSMELPAELTIRYCKHVHSTCSGLPATDGEFCFLVHMLNTDSYFRGF 840
                                                                . :

RRM                DFTLLNRLRLIVDPNIADYITAKNNVVINFKDMSHVNKYGLIRGLKFASFIFQ----- 57

```

```

Prp8_Giardia_2309      NLHVIGKVISLLFDPVISSFLITRLSSSFYFKDMTYTAVRGVAPSFQFSHFLLTLLLSIL 900
:: ::.:::: *:.** *:.::: :: . : *****:. *: .:.*: *::

RRM
Prp8_Giardia_2309      -----
DLTILLHYDEQPFSPPTIILRVVTFIIAFAKDYRQSGVPLSQLWLRKACLEKGIDLSRLSI 960

RRM
Prp8_Giardia_2309      YYG-----LVID----LLLLGQERATDLAGPANN-----PNEFMQFKSKE----- 93
FYGTYKLLSYVRQGESLYLVLQEDRQANIESSKTTIMEQVMKSIPQSIACYISRQSSFCD 1020
:**          * :          **:*:* ::: .. ..          *:.: : *::

U5_binding_1
Prp8_Giardia_2309      -----MPESIRQKK-----ART 12
LSRLLARTAGIIDKDSGNSSRRVTHQRSLANQIVEQRNRYAGRFVSMYPQFTTRSAMTKL 540
:..* :::                          ::

U5_binding_1
Prp8_Giardia_2309      ILQHLSEAWRCWKANIPWDVPG--MPAPIKKIIERYIKSKADAWVSAAHYNRERIKRGAH 70
FLAHLAEAWLCWRAGMAYDQVYSQMSPEVADLVQAYVSERADLYTASIACTKKRIASNKW 600
:* **:*:* *:*.:.*          *.. : .:..: ** :.:. .:.* .

U5_binding_1
Prp8_Giardia_2309      VEKTMVKKNLGR LTR----LWIKNEQERQRQIQKNG----- 102
IAKSEHYKYCGRAGRQDMRELIVANAAYLCEPIQKDLRVSLGTVYSLAYLCITVAARVCA 660
: * : * ** *          * : *          . ***:

U5_binding_2
Prp8_Giardia_2309      -----RQ----- 2
DLTILLHYDEQPFSPPTIILRVVTFIIAFAKDYRQSGVPLSQLWLRKACLEKGIDLSRLSI 960

```


**

```

U5_binding_2      -----RMEEVVSNDDEGVWDLVDER-----TKQRTAKAYLKVSEEEIKKFDSR----- 44
Prp8_Giardia_2309 FYGTYKLLSYVRQGESLYLVLQEDRQANIESSKTTIMEQVMKSIPQSIACYISRQSSFCD 1020
                   :: . * :.*.:: ::*          :*      :  :*      :.*  : **

U5_binding_2      -----IRGILMAS----- 52
Prp8_Giardia_2309 NQTFYFVSLVNQPIRFEFMGYIILAKSIKDLGMSTNILHSVSATFIANFIYHANQLISSA 1080
                   *:.: **:

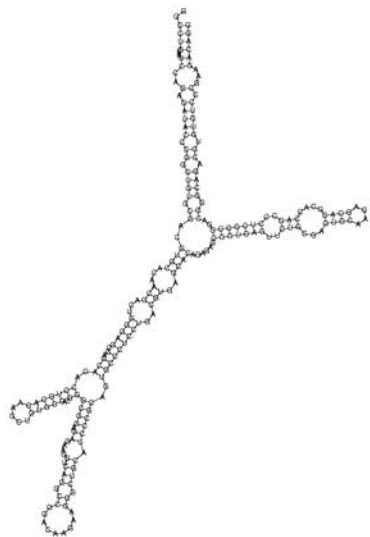
U5_binding_2      GSTTFTKVAAKWNTSLISLFTYFREAIVATEPLLDILVKGETRIQNRVKLGLNSKMPTRF 112
Prp8_Giardia_2309 VSTSFSKIIAKWNSLLLNLCVIYYREALQSPRFLRILMAYEEKVCNKIKQGLNSKMPNRF 1140
                   **:***: ****: *:. . *:*:*: :  :* **:  *  : : *:* *****.**

U5_binding_2      PPAVFYTPKELGGLGMISAS-----HILIP----- 137
Prp8_Giardia_2309 PNVIFYSPRELGGLGMLSVGSAGVYPSSEELNPKYPVAERSRRWDQKHQEVLLPSVIHFI 1200
                   * .:***:*****:*..          .:***

U5_binding_2      ---ASDLS----- 142
Prp8_Giardia_2309 SPWADELNRSFLGYQRLLSIFCEFYNGRSPLFGAQTGFYVYEYDACGAEQQYSFKEISEC 1260
                   *.:*.
    
```

Appendix-4

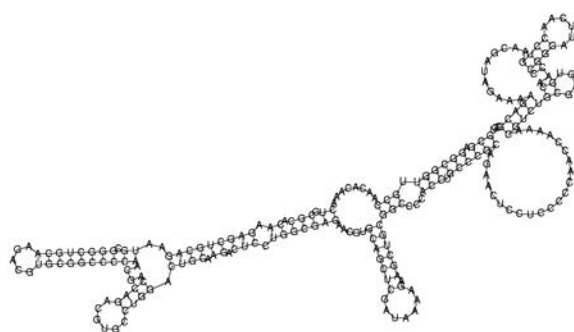
1. Structures of the sense and antisense Girep RNAs



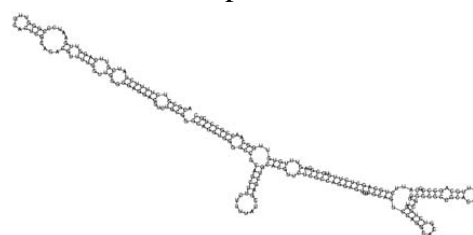
Girep1-sense



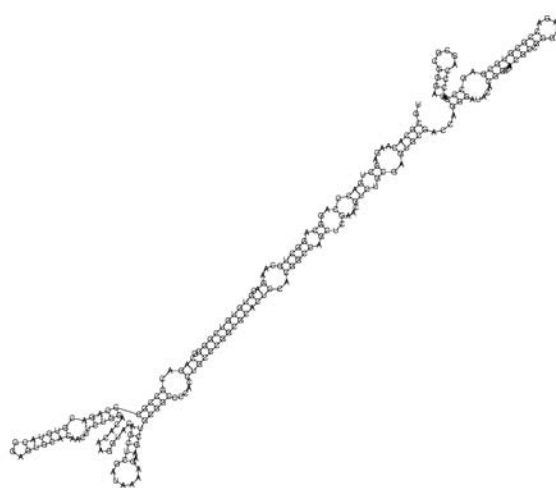
Girep1-antisense



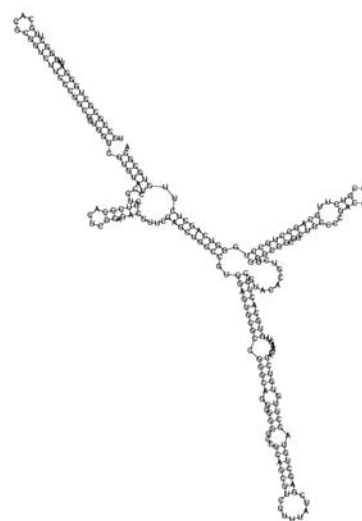
Girep2-sense



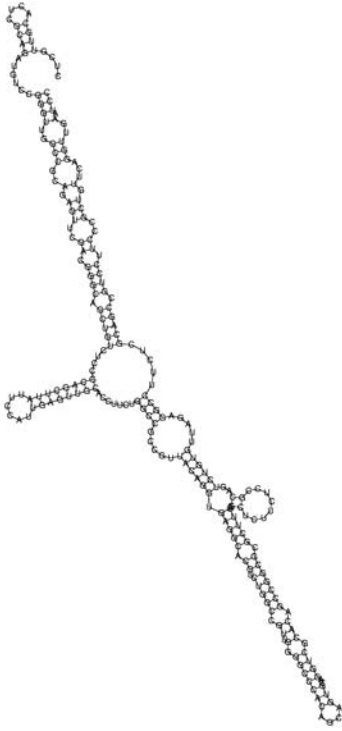
Girep2-antisense



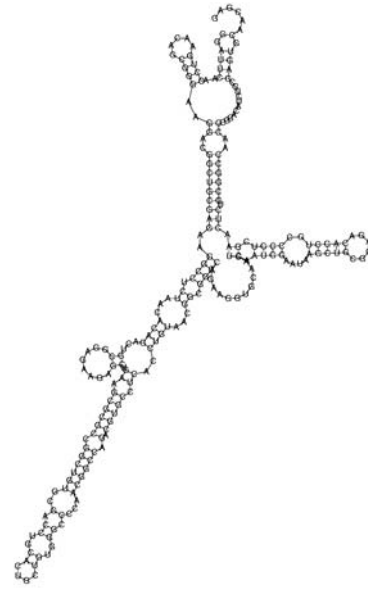
Girep3-sense



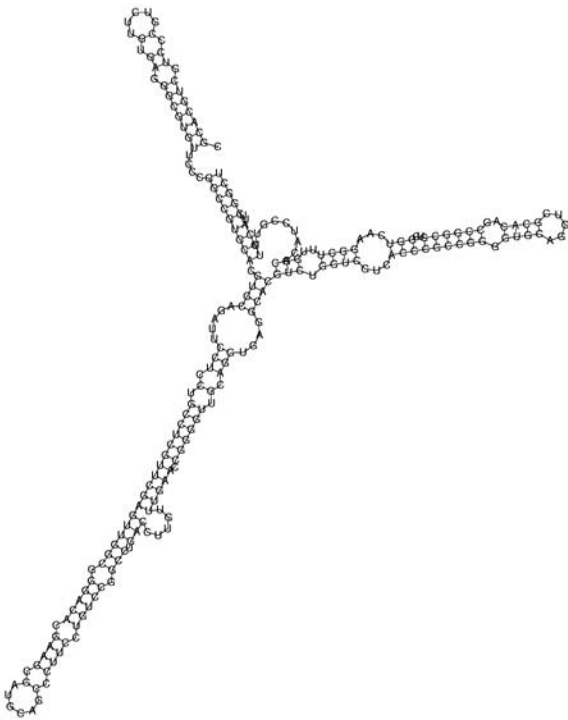
Girep3-antisense



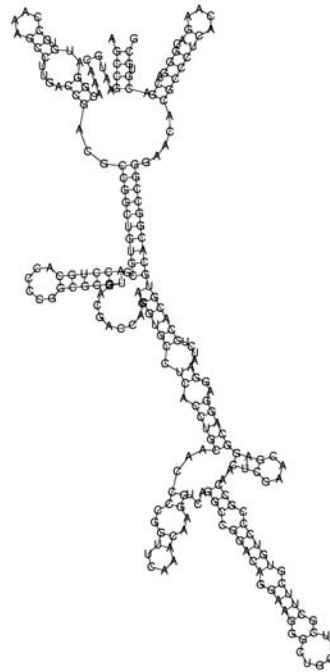
Girep4-sense



Girep4-antisense



Girep5-sense



Girep5-antisense

2. 3'-RACE results of the truncated *Giardia* Dicer transcript

DNase-treated 5ul total original RNA was cleaned by phenol:chloroform extraction and resuspend in 50ul H₂O. The concentration is about 40ng/ul.

One-step RT-PCR was done using 3_RACE_R and GiDicer_cand_1_L primers (C. therm polymerase One-Step RT-PCR system, Roche).

Primers:

3_RACE_R

5' GCT GGACTCT CTA GCG GCC GCT-oligo dT20 3'

GiDicer_cand_1_L

5' C C A T C C C A C T G G T T G C T A C T -3'

50ul reaction:

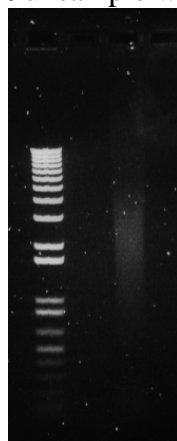
10ul total RNA + 1.5ul 3_RACE_R
85C for 2min – 2min RT – on ice

Enzyme mixture was added to the RNA/primer mix:

2ul 10mM dNTP mix
2.5ul DMSO
2.5ul 100mM DTT
0.5ul 40u/ul RNaseOUT
1.5ul GiDicer_cand_1_L
4.5ul H₂O
10ul 5*buffer
2ul RT-PCR enzyme mix
13ul H₂O
60C 30min – The reaction was then switched to PCR automatically

1) 94C	2min	
2) 94C	30sec	
3) 58C	1min	
4) 72C	3min	Go to 2) 30 cycles
5) 72C	10min	
6) 10C	pause	

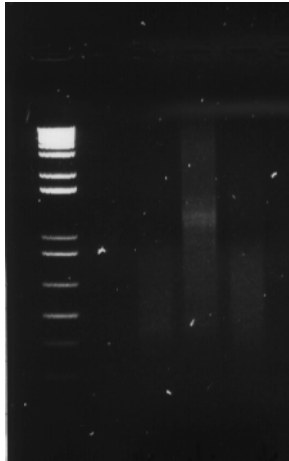
9ul sample was run on 1% agarose gel:



The smear between 2kb and 3kb was cut out and soaked in 0.3M NaOAc (pH 7.5) overnight at 40C and cleaned by phenol:chloroform:isoamylalcohol, and EtOH precipitated.

PCR was done using the resuspended DNA as template using a nested primer GiDicer_3RACE_L and the 3_RACE_R. (Red Hot polymerase kit, ABgene)
GiDicer_3RACE_L: TGGGCC TTTTGG TGTAAGTC

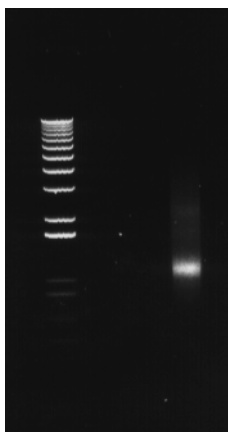
3ul sample was run on a 1% agarose gel



The light bands and smearing between 1kb and 1.4kb were cut out and soaked in 30ul water, incubated at 50C for 1h and 2ul sample was taken out for the 2nd PCR.

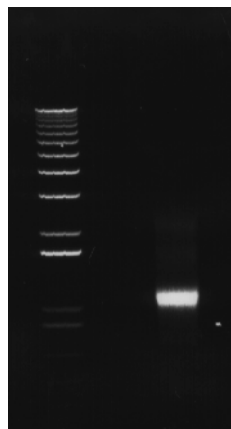
The 2nd PCR was done using the same pair of primers as in the 1st PCR, under the same reaction condition.

3ul sample was run on a 1% agarose gel



The light band around 1.1 or 1.2kb was cut out and soaked as described and a 3rd PCR was done under the same condition.

3ul was run on a 1% agarose gel



The band shown above was sequenced using the M13forward primer.
M13F: CGCCAGGGTTTTCCCAGTCACGAC

The sequence is shown the GiDicer_3_RACE_L primer and the poly-A tail highlighted.

```

ASWGC GTGATGTATACGACTCACTATAGGGCGAATTGGGCCCGACGTCGCATGCTCCCGGCCG
CCATGGCGGCCCGCGGAATTCGATTTGGGCCTTTTGGTGTAAGTACACCCGATGTTTTCCAGC
GACTCGAATTGCTAGGAGATGCTGTGTTAGGCTTTATCGTGA CTGCCCGCCTCCTTTGCCTTT
TTCCAGATGCGTCTGTGGGAACACTTGTTGAGCTAAAGATGGAGCTTGTTCGCAATGAGGCTC
TAAACTATCTTGTACAAACGCTTGGACTTCCTCAGTTGGCGGAGTTTTCCAACAACCTTGTGG
CGAAGAGCAAAACATGGGCAGATATGTATGAGGAGATCGTTGGATCAATCTTTACGGGACCTA
ATGGAATCTATGGCTGTGAGGAATTTCTTGC GAAGACGCTTATGAGTCCCGAACACTCCAAGA
CAGTAGGATCTGCCTGTCCAGATGCAGTCACCAAGGCATCAAAGCGTGTTCATGGGAGAAG
CGGGGGCGCATGAATTCAGAAGCCTTGTGGACTATGCTTGTGAGCAAGGCATCAGTGTCTTCT
GTTCTTCGCGGGTGTCAACTATGTTTCTCGAGCGTCTCAGAGACATTCCAGCAGAGGACATGC
TAGATTGGTACCGACTTGGTATCCAGTTTTTCGCATCGTT CAGGCCTATCAGGACCTGGCGGCG
TCGTATCAGTTATAGACATAATGACACATTTGGCTCGAGGCCTATGGCTGGGCTCTCCAGGCT
TCTATGTTGAACAGCAAAC TGATAAGAATGAGTCGGCTTGTCCGCCC ACTATACTGTTTTAT
ATATCTATCATCGTCTGTG CAGTGTCTGTCTTATATGGGTCGCTCACAGAAACCCCTAC
AGGGCCCGTCGCTTCTAATGTTCTCGCTCTCTATGAGAAGATTCTGGCATATGAGTCATCAGA
GGTAGTAAGCATATAGCAGCTCAGACAGTTAGCAGATCTCTGGCCGTACCCATTCTAGTGGC
ACTATCCCCTTCTGATTCCGTTATTGCAAATAGCACTAACTCCTCACGTGTACAAA ACTTGA
GCAAAAAGAAAAAAAAAAAAAGCGGSCGCTARARATCCMGCAAYMYTATTGAATTCSCGSGCS
SCKGSAGGYCGMACATATGGGARGCTCCCMACGCGGGGGRTGCAAGCTGGAATTCTWTTTGCC
TATAGCTTGCAACTGTCAYGACCGGTGTYGYYSGMTCAKYACASCAAGAAAAGACAAAGAYG

```

Bibliography

- Adam, R. D. (2000). "The Giardia lamblia genome." Int J Parasitol **30**(4): 475-84.
- Adam, R. D. (2001). "Biology of Giardia lamblia." Clin Microbiol Rev **14**(3): 447-75.
- Adam, R. D., T. E. Nash, et al. (1988). "The Giardia lamblia trophozoite contains sets of closely related chromosomes." Nucleic Acids Res **16**(10): 4555-67.
- Adams, P. L., M. R. Stahley, et al. (2004). "Crystal structure of a self-splicing group I intron with both exons." Nature **430**(6995): 45-50.
- Alberts, B., D. Bray, et al. (2002). Molecular Biology of the Cell, Garland Science Publishing.
- Alekseyenko, A. V., N. Kim, et al. (2007). "Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes." Rna **13**(5): 661-70.
- Ambrosio, D. L., M. T. Silva, et al. (2007). "Cloning and molecular characterization of Trypanosoma cruzi U2, U4, U5, and U6 small nuclear RNAs." Mem Inst Oswaldo Cruz **102**(1): 97-105.
- Aparicio, S., J. Chapman, et al. (2002). "Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes." Science **297**(5585): 1301-10.
- Aravin, A., D. Gaidatzis, et al. (2006). "A novel class of small RNAs bind to MILI protein in mouse testes." Nature **442**(7099): 203-7.
- Arisue, N., L. B. Sanchez, et al. (2002). "Mitochondrial-type hsp70 genes of the amitochondriate protists, Giardia intestinalis, Entamoeba histolytica and two microsporidians." Parasitol Int **51**(1): 9-16.
- Arnaiz, O., S. Cain, et al. (2007). "ParameciumDB: a community resource that integrates the Paramecium tetraurelia genome sequence with genetic data." Nucleic Acids Res **35**(Database issue): D439-44.
- Aspegren, A., A. Hinas, et al. (2004). "Novel non-coding RNAs in Dictyostelium discoideum and their expression during development." Nucleic Acids Res **32**(15): 4646-56.
- Bachellerie, J. P., J. Cavaille, et al. (2002). "The expanding snoRNA world." Biochimie **84**(8): 775-90.
- Baer, M., T. W. Nilsen, et al. (1990). "Structure and transcription of a human gene for H1 RNA, the RNA component of human RNase P." Nucleic Acids Res **18**(1): 97-103.
- Balakin, A. G., L. Smith, et al. (1996). "The RNA world of the nucleolus: two major families of small RNAs defined by different box elements with related functions." Cell **86**(5): 823-34.
- Baldauf, S. L. and W. F. Doolittle (1997). "Origin and evolution of the slime molds (Mycetozoa)." Proc Natl Acad Sci U S A **94**(22): 12007-12.

- Bardwell, V. J. and M. Wickens (1990). "Purification of RNA and RNA-protein complexes by an R17 coat protein affinity method." Nucleic Acids Res **18**(22): 6587-94.
- Becak, M. L. and L. S. Kobashi (2004). "Evolution by polyploidy and gene regulation in Anura." Genet Mol Res **3**(2): 195-212.
- Beebe, J. A., J. C. Kurz, et al. (1996). "Magnesium ions are required by *Bacillus subtilis* ribonuclease P RNA for both binding and cleaving precursor tRNA^{Asp}." Biochemistry **35**(32): 10493-505.
- Bejerano, G., C. B. Lowe, et al. (2006). "A distal enhancer and an ultraconserved exon are derived from a novel retroposon." Nature **441**(7089): 87-90.
- Belfort, M. and R. J. Roberts (1997). "Homing endonucleases: keeping the house in order." Nucleic Acids Res **25**(17): 3379-88.
- Bennett, M. D. and I. J. Leitch (2005). Plant DNA C-values Database, Royal Botanic Gardens, Kew.
- Bennett, S. (2004). "Solexa Ltd." Pharmacogenomics **5**(4): 433-8.
- Bentwich, I., A. Avniel, et al. (2005). "Identification of hundreds of conserved and nonconserved human microRNAs." Nat Genet **37**(7): 766-70.
- Berg, J. M. (1988). "Proposed structure for the zinc-binding domains from transcription factor IIIA and related proteins." Proc Natl Acad Sci U S A **85**(1): 99-102.
- Bernstein, D. S., N. Buter, et al. (2002). "Analyzing mRNA-protein complexes using a yeast three-hybrid system." Methods **26**(2): 123-41.
- Bernstein, E. and C. D. Allis (2005). "RNA meets chromatin." Genes Dev **19**(14): 1635-55.
- Bernstein, E., A. A. Caudy, et al. (2001). "Role for a bidentate ribonuclease in the initiation step of RNA interference." Nature **409**(6818): 363-6.
- Best, A. A., H. G. Morrison, et al. (2004). "Evolution of eukaryotic transcription: insights from the genome of *Giardia lamblia*." Genome Res **14**(8): 1537-47.
- Bird, A. P. (1995). "Gene number, noise reduction and biological complexity." Trends Genet **11**(3): 94-100.
- Blount, K. F. and O. C. Uhlenbeck (2002). "The hammerhead ribozyme." Biochem Soc Trans **30**(Pt 6): 1119-22.
- Bohnsack, M. T., K. Czaplinski, et al. (2004). "Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs." Rna **10**(2): 185-91.
- Bonen, L. and J. Vogel (2001). "The ins and outs of group II introns." Trends Genet **17**(6): 322-31.
- Borsuk, P., A. Przykorska, et al. (2007). "L-arginine influences the structure and function of arginase mRNA in *Aspergillus nidulans*." Biol Chem **388**(2): 135-44.

- Bowman, C. M., J. E. Dahlberg, et al. (1971). "Specific inactivation of 16S ribosomal RNA induced by colicin E3 in vivo." Proc Natl Acad Sci U S A **68**(5): 964-8.
- Braidotti, G., T. Baubec, et al. (2004). "The Air noncoding RNA: an imprinted cis-silencing transcript." Cold Spring Harb Symp Quant Biol **69**: 55-66.
- Brannvall, M. and L. A. Kirsebom (2001). "Metal ion cooperativity in ribozyme cleavage of RNA." Proc Natl Acad Sci U S A **98**(23): 12943-7.
- Brennecke, J., A. A. Aravin, et al. (2007). "Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila." Cell **128**(6): 1089-103.
- Brosius, J. (1991). "Retroposons--seeds of evolution." Science **251**(4995): 753.
- Brosius, J. (2005). "Echoes from the past--are we still in an RNP world?" Cytogenet Genome Res **110**(1-4): 8-24.
- Brown, J. W. (1998). "The ribonuclease P database." Nucleic Acids Res **26**(1): 351-2.
- Brown, J. W., G. P. Clark, et al. (2001). "Multiple snoRNA gene clusters from Arabidopsis." Rna **7**(12): 1817-32.
- Brown, J. W., M. Echeverria, et al. (2003). "Plant snoRNAs: functional evolution and new modes of gene expression." Trends Plant Sci **8**(1): 42-9.
- Brown, J. W. and R. Waugh (1989). "Maize U2 snRNAs: gene sequence and expression." Nucleic Acids Res **17**(22): 8991-9001.
- Burke, D. H. and S. T. Greathouse (2005). "Low-magnesium, trans-cleavage activity by type III, tertiary stabilized hammerhead ribozymes with stem 1 discontinuities." BMC Biochem **6**: 14.
- Calin-Jageman, I. and A. W. Nicholson (2003). "RNA structure-dependent uncoupling of substrate recognition and cleavage by Escherichia coli ribonuclease III." Nucleic Acids Res **31**(9): 2381-92.
- Carmell, M. A., Z. Xuan, et al. (2002). "The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis." Genes Dev **16**(21): 2733-42.
- Carninci, P., T. Kasukawa, et al. (2005). "The transcriptional landscape of the mammalian genome." Science **309**(5740): 1559-63.
- Cavaille, J., K. Buiting, et al. (2000). "Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization." Proc Natl Acad Sci U S A **97**(26): 14311-6.
- Cavaille, J., M. Nicoloso, et al. (1996). "Targeted ribose methylation of RNA in vivo directed by tailored antisense RNA guides." Nature **383**(6602): 732-5.
- Cavalier-Smith, T. (1985). "Selfish DNA and the origin of introns." Nature **315**(6017): 283-4.
- Cavalier-Smith, T. (1991). "Intron phylogeny: a new hypothesis." Trends Genet **7**(5): 145-8.

- Cerutti, L., N. Mian, et al. (2000). "Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain." *Trends Biochem Sci* **25**(10): 481-2.
- Chamberlain, J. R., Y. Lee, et al. (1998). "Purification and characterization of the nuclear RNase P holoenzyme complex reveals extensive subunit overlap with RNase MRP." *Genes Dev* **12**(11): 1678-90.
- Chan, S. W., I. R. Henderson, et al. (2006). "RNAi, DRD1, and histone methylation actively target developmentally important non-CG DNA methylation in arabidopsis." *PLoS Genet* **2**(6): e83.
- Chaussivert, N., C. Conesa, et al. (1995). "Complex interactions between yeast TFIIB and TFIIC." *J Biol Chem* **270**(25): 15353-8.
- Cheah, M. T., A. Wachter, et al. (2007). "Control of alternative RNA splicing and gene expression by eukaryotic riboswitches." *Nature* **447**(7143): 497-500.
- Chen, C. H., D. I. Kao, et al. (2006). "Functional links between the Prp19-associated complex, U4/U6 biogenesis, and spliceosome recycling." *Rna* **12**(5): 765-74.
- Chen, X. S., T. S. Rozhdestvensky, et al. (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.
- Collins, C. A. and C. Guthrie (1999). "Allele-specific genetic interactions between Prp8 and RNA active site residues suggest a function for Prp8 at the catalytic core of the spliceosome." *Genes Dev* **13**(15): 1970-82.
- Collins, C. A. and C. Guthrie (2000). "The question remains: is the spliceosome a ribozyme?" *Nat Struct Biol* **7**(10): 850-4.
- Collins, L. and D. Penny (2005). "Complex spliceosomal organization ancestral to extant eukaryotes." *Mol Biol Evol* **22**(4): 1053-66.
- Collins, L. J., T. J. Macke, et al. (2003). "Searching for ncRNAs in eukaryotic genomes: Maximizing biological input with RNAmotif." *Journal of Integrative Bioinformatics* **0001**.
- Collins, R. E. and X. Cheng (2006). "Structural and biochemical advances in mammalian RNAi." *J Cell Biochem* **99**(5): 1251-66.
- Cousineau, B., S. Lawrence, et al. (2000). "Retrotransposition of a bacterial group II intron." *Nature* **404**(6781): 1018-21.
- Crick, F. (1970). "Central dogma of molecular biology." *Nature* **227**(5258): 561-3.
- Dacks, J. B., L. A. Davis, et al. (2003). "Evidence for Golgi bodies in proposed 'Golgi-lacking' lineages." *Proc Biol Sci* **270** Suppl 2: S168-71.
- Dacks, J. B., A. Marinets, et al. (2002). "Analyses of RNA Polymerase II genes from free-living protists: phylogeny, long branch attraction, and the eukaryotic big bang." *Mol Biol Evol* **19**(6): 830-40.
- Das, R., Z. Zhou, et al. (2000). "Functional association of U2 snRNP with the ATP-independent spliceosomal complex E." *Mol Cell* **5**(5): 779-87.

- Datta, B. and A. M. Weiner (1991). "Genetic evidence for base pairing between U2 and U6 snRNA in mammalian mRNA splicing." Nature **352**(6338): 821-4.
- David, L., W. Huber, et al. (2006). "A high-resolution map of transcription in the yeast genome." Proc Natl Acad Sci U S A **103**(14): 5320-5.
- de Duve, C. (2007). "The origin of eukaryotes: a reappraisal." Nat Rev Genet **8**(5): 395-403.
- de la Cruz, J., D. Kressler, et al. (1999). "Unwinding RNA in *Saccharomyces cerevisiae*: DEAD-box proteins and related families." Trends Biochem Sci **24**(5): 192-8.
- De la Pena, M., S. Gago, et al. (2003). "Peripheral regions of natural hammerhead ribozymes greatly increase their self-cleavage activity." Embo J **22**(20): 5561-70.
- De Rijk, P., J. Wuyts, et al. (2003). "RnaViz 2: an improved representation of RNA secondary structure." Bioinformatics **19**(2): 299-300.
- de Souza, S. J. (2003). "The emergence of a synthetic theory of intron evolution." Genetica **118**(2-3): 117-21.
- Deng, X. and V. H. Meller (2006). "Non-coding RNA in fly dosage compensation." Trends Biochem Sci **31**(9): 526-32.
- Dennis, P. P., A. Omer, et al. (2001). "A guided tour: small RNA function in Archaea." Mol Microbiol **40**(3): 509-19.
- Dermitzakis, E. T., A. Reymond, et al. (2003). "Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs)." Science **302**(5647): 1033-5.
- Di Giulio, M. (1999). "The non-monophyletic origin of the tRNA molecule." J Theor Biol **197**(3): 403-14.
- Diener, T. O. (1989). "Circular RNAs: relics of precellular evolution?" Proc Natl Acad Sci U S A **86**(23): 9370-4.
- Dix, I., C. S. Russell, et al. (1998). "Protein-RNA interactions in the U5 snRNP of *Saccharomyces cerevisiae*." Rna **4**(12): 1675-86.
- Djikeng, A., H. Shi, et al. (2003). "An siRNA ribonucleoprotein is found associated with polyribosomes in *Trypanosoma brucei*." Rna **9**(7): 802-8.
- Djikeng, A., H. Shi, et al. (2001). "RNA interference in *Trypanosoma brucei*: cloning of small interfering RNAs provides evidence for retroposon-derived 24-26-nucleotide RNAs." Rna **7**(11): 1522-30.
- Doherty, E. A. and J. A. Doudna (2000). "Ribozyme structures and mechanisms." Annu Rev Biochem **69**: 597-615.
- Doolittle, W. F. (1999). "Phylogenetic classification and the universal tree." Science **284**(5423): 2124-9.
- Doudna, J. A., B. P. Cormack, et al. (1989). "RNA structure, not sequence, determines the 5' splice-site specificity of a group I intron." Proc Natl Acad Sci U S A **86**(19): 7402-6.

- Doudna, J. A. and J. R. Lorsch (2005). "Ribozyme catalysis: not different, just worse." Nat Struct Mol Biol **12**(5): 395-402.
- Dowell, R. D. and S. R. Eddy (2006). "Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints." BMC Bioinformatics **7**: 400.
- Dumas, C., C. Chow, et al. (2006). "A novel class of developmentally regulated noncoding RNAs in Leishmania." Eukaryot Cell **5**(12): 2033-46.
- Dunbar, D. A., A. A. Chen, et al. (2000). "The genes for small nucleolar RNAs in Trypanosoma brucei are organized in clusters and are transcribed as a polycistronic RNA." Nucleic Acids Res **28**(15): 2855-61.
- Dunker, A. K., J. D. Lawson, et al. (2001). "Intrinsically disordered protein." J Mol Graph Model **19**(1): 26-59.
- Dyall, S. D., M. T. Brown, et al. (2004). "Ancient invasions: from endosymbionts to organelles." Science **304**(5668): 253-7.
- Earnshaw, D. J., B. Masquida, et al. (1997). "Inter-domain cross-linking and molecular modelling of the hairpin ribozyme." J Mol Biol **274**(2): 197-212.
- Eddy, S. R. (2002). "A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure." BMC Bioinformatics **3**: 18.
- Eddy, S. R. (2006). "Computational analysis of RNAs." Cold Spring Harb Symp Quant Biol **71**: 117-28.
- Edlind, T. D. and P. R. Chakraborty (1987). "Unusual ribosomal RNA of the intestinal parasite Giardia lamblia." Nucleic Acids Res **15**(19): 7889-901.
- Ehresmann, C., P. Stiegler, et al. (1977). "Recent progress in the determination of the primary sequence of the 16 S RNA of Escherichia coli." FEBS Lett **84**(2): 337-41.
- Eigen, M. (1993). "The origin of genetic information: viruses as models." Gene **135**(1-2): 37-47.
- Eigen, M. and P. Schuster (1977). "The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle." Naturwissenschaften **64**(11): 541-65.
- Eigen, M. and R. Winkler-Oswatitsch (1981). "Transfer-RNA, an early gene?" Naturwissenschaften **68**(6): 282-92.
- Eigen, M. and R. Winkler-Oswatitsch (1981). "Transfer-RNA: the early adaptor." Naturwissenschaften **68**(5): 217-28.
- Eisen, J. A., R. S. Coyne, et al. (2006). "Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote." PLoS Biol **4**(9): e286.
- Elmendorf, H. G., S. M. Singer, et al. (2001). "The abundance of sterile transcripts in Giardia lamblia." Nucleic Acids Res **29**(22): 4674-83.

- Elmendorf, H. G., S. M. Singer, et al. (2001). "Initiator and upstream elements in the alpha2-tubulin promoter of *Giardia lamblia*." *Mol Biochem Parasitol* **113**(1): 157-69.
- Embley, T. M. and W. Martin (2006). "Eukaryotic evolution, changes and challenges." *Nature* **440**(7084): 623-30.
- Embley, T. M., M. van der Giezen, et al. (2003). "Mitochondria and hydrogenosomes are two forms of the same fundamental organelle." *Philos Trans R Soc Lond B Biol Sci* **358**(1429): 191-201; discussion 201-2.
- Fast, N. M. and W. F. Doolittle (1999). "Trichomonas vaginalis possesses a gene encoding the essential spliceosomal component, PRP8." *Mol Biochem Parasitol* **99**(2): 275-8.
- Fedor, M. J. and J. R. Williamson (2005). "The catalytic diversity of RNAs." *Nat Rev Mol Cell Biol* **6**(5): 399-412.
- Fedorova, L. and A. Fedorov (2003). "Introns in gene evolution." *Genetica* **118**(2-3): 123-31.
- Fedorova, O., T. Mitros, et al. (2003). "Domains 2 and 3 interact to form critical elements of the group II intron active site." *J Mol Biol* **330**(2): 197-209.
- Feng, J., C. Bi, et al. (2006). "The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator." *Genes Dev* **20**(11): 1470-84.
- Ferrari, R., C. Rivetti, et al. (2004). "Distinct roles of transcription factors TFIIB and TFIIC in RNA polymerase III transcription reinitiation." *Proc Natl Acad Sci U S A* **101**(37): 13442-7.
- Ferre-D'Amare, A. R., K. Zhou, et al. (1998). "Crystal structure of a hepatitis delta virus ribozyme." *Nature* **395**(6702): 567-74.
- Finn, R. D., J. Mistry, et al. (2006). "Pfam: clans, web tools and services." *Nucleic Acids Res* **34**(Database issue): D247-51.
- Fire, A., S. Xu, et al. (1998). "Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*." *Nature* **391**(6669): 806-11.
- Forrest, E. C., C. Cogoni, et al. (2004). "The RNA-dependent RNA polymerase, QDE-1, is a rate-limiting factor in post-transcriptional gene silencing in *Neurospora crassa*." *Nucleic Acids Res* **32**(7): 2123-8.
- Fortner, D. M., R. G. Troy, et al. (1994). "A stem/loop in U6 RNA defines a conformational switch required for pre-mRNA splicing." *Genes Dev* **8**(2): 221-33.
- Frank, D. N. and N. R. Pace (1998). "Ribonuclease P: unity and diversity in a tRNA processing ribozyme." *Annu Rev Biochem* **67**: 153-80.
- Fredrick, K. and H. F. Noller (2003). "Catalysis of ribosomal translocation by sparsomycin." *Science* **300**(5622): 1159-62.
- Furuno, M., K. C. Pang, et al. (2006). "Clusters of internally primed transcripts reveal novel long noncoding RNAs." *PLoS Genet* **2**(4): e37.

- Galli, G., H. Hofstetter, et al. (1981). "Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements." Nature **294**(5842): 626-31.
- Garcia, I. and K. M. Weeks (2003). "Small structural costs for evolution from RNA to RNP-based catalysis." J Mol Biol **331**(1): 57-73.
- Gazzani, S., T. Lawrenson, et al. (2004). "A link between mRNA turnover and RNA interference in Arabidopsis." Science **306**(5698): 1046-8.
- Geiduschek, E. P. and G. A. Kassavetis (2001). "The RNA polymerase III transcription apparatus." J Mol Biol **310**(1): 1-26.
- Gerbi, S. A. (1995). "Small nucleolar RNA." Biochem Cell Biol **73**(11-12): 845-58.
- Gesteland, R. F., T. R. Cech, et al. (2006). The RNA World. New York, Cold Spring Harbor Laboratory.
- Gilbert, W. (1978). "Why genes in pieces?" Nature **271**(5645): 501.
- Gilbert, W. (1986). "The RNA World." Nature **319**: 618.
- Gillin, F. D., D. S. Reiner, et al. (1996). "Cell biology of the primitive eukaryote *Giardia lamblia*." Annu Rev Microbiol **50**: 679-705.
- Ginger, M. R., A. N. Shore, et al. (2006). "A noncoding RNA is a potential marker of cell fate during mammary gland development." Proc Natl Acad Sci U S A **103**(15): 5781-6.
- Goddard, M. R. and A. Burt (1999). "Recurrent invasion and extinction of a selfish gene." Proc Natl Acad Sci U S A **96**(24): 13880-5.
- Golden, B. L., H. Kim, et al. (2005). "Crystal structure of a phage T7 group I ribozyme-product complex." Nat Struct Mol Biol **12**(1): 82-9.
- Goodstadt, L. and C. P. Ponting (2006). "Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human." PLoS Comput Biol **2**(9): e133.
- Gordon, P. M., R. Fong, et al. (2007). "A second divalent metal ion in the group II intron reaction center." Chem Biol **14**(6): 607-12.
- Gottesman, S. (2005). "Micros for microbes: non-coding regulatory RNAs in bacteria." Trends Genet **21**(7): 399-404.
- Grainger, R. J. and J. D. Beggs (2005). "Prp8 protein: at the heart of the spliceosome." Rna **11**(5): 533-57.
- Gregory, T. R. (2005). Animal genome size database.
- Grewal, S. I. and S. C. Elgin (2007). "Transcription and RNA interference in the formation of heterochromatin." Nature **447**(7143): 399-406.
- Griffiths-Jones, S., A. Bateman, et al. (2003). "Rfam: an RNA family database." Nucleic Acids Res **31**(1): 439-41.
- Griffiths-Jones, S., S. Moxon, et al. (2005). "Rfam: annotating non-coding RNAs in complete genomes." Nucleic Acids Res **33**(Database issue): D121-4.
- Gunasekera, A. M., S. Patankar, et al. (2004). "Widespread distribution of antisense transcripts in the *Plasmodium falciparum* genome." Mol Biochem Parasitol **136**(1): 35-42.

- Guo, F., A. R. Gooding, et al. (2004). "Structure of the Tetrahymena ribozyme: base triple sandwich and metal ion at the active site." *Mol Cell* **16**(3): 351-62.
- Gupta, R. C. and K. Randerath (1977). "Use of specific endonuclease cleavage in RNA sequencing." *Nucleic Acids Res* **4**(6): 1957-78.
- Gupta, R. C. and K. Randerath (1977). "Use of specific endonuclease cleavage in RNA sequencing-an enzymic method for distinguishing between cytidine and uridine residues." *Nucleic Acids Res* **4**(10): 3441-54.
- Gutell, R. R., N. Larsen, et al. (1994). "Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective." *Microbiol Rev* **58**(1): 10-26.
- Haas, B. J., J. R. Wortman, et al. (2005). "Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release." *BMC Biol* **3**: 7.
- Hahn, M. W. and G. A. Wray (2002). "The g-value paradox." *Evol Dev* **4**(2): 73-5.
- Hammond, S. M., S. Boettcher, et al. (2001). "Argonaute2, a link between genetic and biochemical analyses of RNAi." *Science* **293**(5532): 1146-50.
- Hammond, S. M., A. A. Caudy, et al. (2001). "Post-transcriptional gene silencing by double-stranded RNA." *Nat Rev Genet* **2**(2): 110-9.
- Han, J., Y. Lee, et al. (2006). "Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex." *Cell* **125**(5): 887-901.
- Hariharan, N. and R. P. Perry (1990). "Functional dissection of a mouse ribosomal protein promoter: significance of the polypyrimidine initiator and an element in the TATA-box region." *Proc Natl Acad Sci U S A* **87**(4): 1526-30.
- Hartig, J. V., Y. Tomari, et al. (2007). "piRNAs--the ancient hunters of genome invaders." *Genes Dev* **21**(14): 1707-13.
- Hartmann, E. and R. K. Hartmann (2003). "The enigma of ribonuclease P evolution." *Trends Genet* **19**(10): 561-9.
- Haugen, P., D. M. Simon, et al. (2005). "The natural history of group I introns." *Trends Genet* **21**(2): 111-9.
- Hausner, T. P., L. M. Giglio, et al. (1990). "Evidence for base-pairing between mammalian U2 and U6 small nuclear ribonucleoprotein particles." *Genes Dev* **4**(12A): 2146-56.
- Havgaard, J. H., R. B. Lyngso, et al. (2005). "The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search." *Nucleic Acids Res* **33**(Web Server issue): W650-3.
- He, H., L. Cai, et al. (2006). "Profiling Caenorhabditis elegans non-coding RNA expression with a combined microarray." *Nucleic Acids Res* **34**(10): 2976-83.
- Helser, T. L., J. E. Davies, et al. (1972). "Mechanism of kasugamycin resistance in Escherichia coli." *Nat New Biol* **235**(53): 6-9.

- Henry, R. W., V. Mittal, et al. (1998). "SNAP19 mediates the assembly of a functional core promoter complex (SNAPc) shared by RNA polymerases II and III." *Genes Dev* **12**(17): 2664-72.
- Herr, W., R. A. Sturm, et al. (1988). "The POU domain: a large conserved region in the mammalian pit-1, oct-1, oct-2, and *Caenorhabditis elegans* unc-86 gene products." *Genes Dev* **2**(12A): 1513-6.
- Hiley, S. L., T. Babak, et al. (2005). "Global analysis of yeast RNA processing identifies new targets of RNase III and uncovers a link between tRNA 5' end processing and tRNA splicing." *Nucleic Acids Res* **33**(9): 3048-56.
- Hinas, A., P. Larsson, et al. (2006). "Identification of the major spliceosomal RNAs in *Dictyostelium discoideum* reveals developmentally regulated U2 variants and polyadenylated snRNAs." *Eukaryot Cell* **5**(6): 924-34.
- Hirt, R. P., J. M. Logsdon, Jr., et al. (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): 580-5.
- Hodges, P. E., S. P. Jackson, et al. (1995). "Extraordinary sequence conservation of the PRP8 splicing factor." *Yeast* **11**(4): 337-42.
- Hofacker, I. L. (2003). "Vienna RNA secondary structure server." *Nucleic Acids Res* **31**(13): 3429-31.
- Hofmann, C. J., C. Marshallsay, et al. (1992). "Characterization of the genes encoding U4 small nuclear RNAs in *Arabidopsis thaliana*." *Mol Biol Rep* **17**(1): 21-8.
- Holberton, D. V. and J. Marshall (1995). "Analysis of consensus sequence patterns in *Giardia* cytoskeleton gene promoters." *Nucleic Acids Res* **23**(15): 2945-53.
- Holmes, I. (2005). "Accelerated probabilistic inference of RNA structure evolution." *BMC Bioinformatics* **6**: 73.
- Hoogstraten, C. G. and M. Sumita (2007). "Structure-function relationships in RNA and RNP enzymes: Recent advances." *Biopolymers* **87**(5-6): 317-28.
- Hook, B., D. Bernstein, et al. (2005). "RNA-protein interactions in the yeast three-hybrid system: affinity, sensitivity, and enhanced library screening." *Rna* **11**(2): 227-33.
- Horton, T. E., D. R. Clardy, et al. (1998). "Electron paramagnetic resonance spectroscopic measurement of Mn²⁺ binding affinities to the hammerhead ribozyme and correlation with cleavage activity." *Biochemistry* **37**(51): 18094-101.
- Houglund, J. L., A. V. Kravchuk, et al. (2005). "Functional identification of catalytic metal ion binding sites within RNA." *PLoS Biol* **3**(9): e277.
- Houwing, S., L. M. Kamminga, et al. (2007). "A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish." *Cell* **129**(1): 69-82.

- Huang, F., C. W. Bugg, et al. (2000). "RNA-Catalyzed CoA, NAD, and FAD synthesis from phosphopantetheine, NMN, and FMN." Biochemistry **39**(50): 15548-55.
- Huang, Z. P., H. Zhou, et al. (2004). "Different expression strategy: multiple intronic gene clusters of box H/ACA snoRNA in *Drosophila melanogaster*." J Mol Biol **341**(3): 669-83.
- Huppler, A., L. J. Nikstad, et al. (2002). "Metal binding and base ionization in the U6 RNA intramolecular stem-loop structure." Nat Struct Biol **9**(6): 431-5.
- Huttenhofer, A., J. Cavaille, et al. (2004). "Experimental RNomics: a global approach to identifying small nuclear RNAs and their targets in different model organisms." Methods Mol Biol **265**: 409-28.
- Huttenhofer, A. and J. Vogel (2006). "Experimental approaches to identify non-coding RNAs." Nucleic Acids Res **34**(2): 635-46.
- Ikawa, Y., H. Shiraishi, et al. (2000). "Minimal catalytic domain of a group I self-splicing intron RNA." Nat Struct Biol **7**(11): 1032-5.
- Imanishi, T., T. Itoh, et al. (2004). "Integrative annotation of 21,037 human genes validated by full-length cDNA clones." PLoS Biol **2**(6): e162.
- Inada, D. C., A. Bashir, et al. (2003). "Conserved noncoding sequences in the grasses." Genome Res **13**(9): 2030-41.
- Inada, M. and C. Guthrie (2004). "Identification of Lhp1p-associated RNAs by microarray analysis in *Saccharomyces cerevisiae* reveals association with coding and noncoding RNAs." Proc Natl Acad Sci U S A **101**(2): 434-9.
- Inagaki, S., K. Numata, et al. (2005). "Identification and expression analysis of putative mRNA-like non-coding RNA in *Drosophila*." Genes Cells **10**(12): 1163-73.
- Inagaki, Y., C. Blouin, et al. (2003). "Assessing functional divergence in EF-1alpha and its paralogs in eukaryotes and archaeobacteria." Nucleic Acids Res **31**(14): 4227-37.
- James, S. A., W. Turner, et al. (2002). "How Slu7 and Prp18 cooperate in the second step of yeast pre-mRNA splicing." Rna **8**(8): 1068-77.
- Januschka, M. M., S. L. Erlandsen, et al. (1988). "A comparison of *Giardia microti* and *Spironucleus muris* cysts in the vole: an immunocytochemical, light, and electron microscopic study." J Parasitol **74**(3): 452-8.
- Jeffares, D. C., A. M. Poole, et al. (1998). "Relics from the RNA world." J Mol Evol **46**(1): 18-36.
- Jensen, R. C., Y. Wang, et al. (1998). "The proximal sequence element (PSE) plays a major role in establishing the RNA polymerase specificity of *Drosophila* U-snoRNA genes." Nucleic Acids Res **26**(2): 616-22.
- Joss, J. M. (2006). "Lungfish evolution and development." Gen Comp Endocrinol **148**(3): 285-9.

- Juneau, K., M. Miranda, et al. (2006). "Introns regulate RNA and protein abundance in yeast." Genetics **174**(1): 511-8.
- Jurica, M. S. and M. J. Moore (2003). "Pre-mRNA splicing: awash in a sea of proteins." Mol Cell **12**(1): 5-14.
- Kassavetis, G. A. and E. P. Geiduschek (2006). "Transcription factor TFIIB and transcription by RNA polymerase III." Biochem Soc Trans **34**(Pt 6): 1082-7.
- Kassavetis, G. A., S. Han, et al. (2003). "The role of transcription initiation factor IIB subunits in promoter opening probed by photochemical cross-linking." J Biol Chem **278**(20): 17912-7.
- Kassavetis, G. A., G. A. Letts, et al. (1999). "A minimal RNA polymerase III transcription system." Embo J **18**(18): 5042-51.
- Katayama, S., Y. Tomaru, et al. (2005). "Antisense transcription in the mammalian transcriptome." Science **309**(5740): 1564-6.
- Katinka, M. D., S. Duprat, et al. (2001). "Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*." Nature **414**(6862): 450-3.
- Ke, A., F. Ding, et al. (2007). "Structural roles of monovalent cations in the HDV ribozyme." Structure **15**(3): 281-7.
- Keeling, P. J., G. Burger, et al. (2005). "The tree of eukaryotes." Trends Ecol Evol **20**(12): 670-6.
- Keeling, P. J. and W. F. Doolittle (1996). "Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family." Mol Biol Evol **13**(10): 1297-305.
- Kent, O., S. G. Chaulk, et al. (2000). "Kinetic analysis of the M1 RNA folding pathway." J Mol Biol **304**(5): 699-705.
- Khvorova, A., A. Lescoute, et al. (2003). "Sequence elements outside the hammerhead ribozyme catalytic core enable intracellular activity." Nat Struct Biol **10**(9): 708-12.
- Kidwell, M. G. (2002). "Transposable elements and the evolution of genome size in eukaryotes." Genetica **115**(1): 49-63.
- Kikovska, E., S. G. Svard, et al. (2007). "From the Cover: Eukaryotic RNase P RNA mediates cleavage in the absence of protein." Proc Natl Acad Sci U S A **104**(7): 2062-7.
- Kim, D. S., V. Gusti, et al. (2005). "An artificial riboswitch for controlling pre-mRNA splicing." Rna **11**(11): 1667-77.
- Kim, S., H. Shi, et al. (2003). "Specific SR protein-dependent splicing substrates identified through genomic SELEX." Nucleic Acids Res **31**(7): 1955-61.
- Kin, T., K. Yamada, et al. (2007). "fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences." Nucleic Acids Res **35**(Database issue): D145-8.
- Kirsebom, L. A. (2007). "RNase P RNA mediated cleavage: Substrate recognition and catalysis." Biochimie.

- Kiryu, H., Y. Tabei, et al. (2007). "Murlet: A practical multiple alignment tool for structural RNA sequences." Bioinformatics.
- Kiss, T. (2006). "SnoRNP biogenesis meets Pre-mRNA splicing." Mol Cell **23**(6): 775-6.
- Kisseleva, N., A. Khvorova, et al. (2005). "Binding of manganese(II) to a tertiary stabilized hammerhead ribozyme as studied by electron paramagnetic resonance spectroscopy." Rna **11**(1): 1-6.
- Klein, R. J. and S. R. Eddy (2003). "RSEARCH: finding homologs of single structured RNA sequences." BMC Bioinformatics **4**: 44.
- Klemm, J. D., M. A. Rould, et al. (1994). "Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules." Cell **77**(1): 21-32.
- Kramer, A. (1996). "The structure and function of proteins involved in mammalian pre-mRNA splicing." Annu Rev Biochem **65**: 367-409.
- Krek, A., D. Grun, et al. (2005). "Combinatorial microRNA target predictions." Nat Genet **37**(5): 495-500.
- Kruger, K., P. J. Grabowski, et al. (1982). "Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena." Cell **31**(1): 147-57.
- Krull, M., M. Petrusma, et al. (2007). "Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs)." Genome Res **17**(8): 1139-45.
- Kubodera, T., M. Watanabe, et al. (2003). "Thiamine-regulated gene expression of *Aspergillus oryzae* thiA requires splicing of the intron containing a riboswitch-like domain in the 5'-UTR." FEBS Lett **555**(3): 516-20.
- Kulakova, L., S. M. Singer, et al. (2006). "Epigenetic mechanisms are involved in the control of *Giardia lamblia* antigenic variation." Mol Microbiol **61**(6): 1533-42.
- Kunkel, G. R. and T. Pederson (1988). "Upstream elements required for efficient transcription of a human U6 RNA gene resemble those of U1 and U2 genes even though a different polymerase is used." Genes Dev **2**(2): 196-204.
- Lagos-Quintana, M., R. Rauhut, et al. (2003). "New microRNAs from mouse and human." Rna **9**(2): 175-9.
- Lai, C. J., J. E. Dahlberg, et al. (1973). "Structure of an inducibly methylatable nucleotide sequence in 23S ribosomal ribonucleic acid from erythromycin-resistant *Staphylococcus aureus*." Biochemistry **12**(3): 457-60.
- Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.
- Lanfredi-Rangel, A., M. Attias, et al. (1998). "The peripheral vesicles of trophozoites of the primitive protozoan *Giardia lamblia* may correspond

- to early and late endosomes and to lysosomes." J Struct Biol **123**(3): 225-35.
- Lau, N. C., A. G. Seto, et al. (2006). "Characterization of the piRNA complex from rat testes." Science **313**(5785): 363-7.
- Le Blancq, S. M., R. S. Kase, et al. (1991). "Analysis of a Giardia lamblia rRNA encoding telomere with [TAGGG]n as the telomere repeat." Nucleic Acids Res **19**(20): 5790.
- Leader, D. J., G. P. Clark, et al. (1997). "Clusters of multiple different small nucleolar RNA genes in plants are expressed as and processed from polycistronic pre-snoRNAs." Embo J **16**(18): 5742-51.
- Lee, B. M., J. Xu, et al. (2006). "Induced fit and "lock and key" recognition of 5S RNA by zinc fingers of transcription factor IIIA." J Mol Biol **357**(1): 275-91.
- Lee, Y., C. Ahn, et al. (2003). "The nuclear RNase III Drosha initiates microRNA processing." Nature **425**(6956): 415-9.
- Lee, Y. S., K. Nakahara, et al. (2004). "Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways." Cell **117**(1): 69-81.
- Lehmann, J. (2002). "Amplification of the sequences displaying the pattern RNY in the RNA world: the translation --> translation/replication hypothesis." J Theor Biol **219**(4): 521-37.
- Lehmann, J., B. Riedo, et al. (2004). "Folding of small RNAs displaying the GNC base-pattern: implications for the self-organization of the genetic system." J Theor Biol **227**(3): 381-95.
- Lehmann, K. and U. Schmidt (2003). "Group II introns: structure and catalytic versatility of large natural ribozymes." Crit Rev Biochem Mol Biol **38**(3): 249-303.
- Levine, N. D., J. O. Corliss, et al. (1980). "A newly revised classification of the protozoa." J Protozool **27**(1): 37-58.
- Lewis, B. P., C. B. Burge, et al. (2005). "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." Cell **120**(1): 15-20.
- Li, M. and P. M. B. Vitanyi (1997). An introduction to Kolmogorov complexity and its applications. New York, Springer-Verlag.
- Li, S. C., P. Tang, et al. (2007). "Intronic microRNA: discovery and biological implications." DNA Cell Biol **26**(4): 195-207.
- Liang, D., H. Zhou, et al. (2002). "A novel gene organization: intronic snoRNA gene clusters from Oryza sativa." Nucleic Acids Res **30**(14): 3262-72.
- Liang, X. H., Q. Liu, et al. (2003). "Small nucleolar RNA interference induced by antisense or double-stranded RNA in trypanosomatids." Proc Natl Acad Sci U S A **100**(13): 7521-6.
- Lim, L. P., N. C. Lau, et al. (2005). "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs." Nature **433**(7027): 769-73.

- Linding, R., L. J. Jensen, et al. (2003). "Protein disorder prediction: implications for structural proteomics." Structure **11**(11): 1453-9.
- Linding, R., R. B. Russell, et al. (2003). "GlobPlot: Exploring protein sequences for globularity and disorder." Nucleic Acids Res **31**(13): 3701-8.
- Lindmark, D. G. and M. Muller (1973). "Hydrogenosome, a cytoplasmic organelle of the anaerobic flagellate *Tritrichomonas foetus*, and its role in pyruvate metabolism." J Biol Chem **248**(22): 7724-8.
- Little, P. F. (2005). "Structure and function of the human genome." Genome Res **15**(12): 1759-66.
- Liu, C., B. Bai, et al. (2005). "NONCODE: an integrated knowledge database of non-coding RNAs." Nucleic Acids Res **33**(Database issue): D112-5.
- Lloyd, D., J. R. Ralphs, et al. (2002). "*Giardia intestinalis*, a eukaryote without hydrogenosomes, produces hydrogen." Microbiology **148**(Pt 3): 727-33.
- Lossky, M., G. J. Anderson, et al. (1987). "Identification of a yeast snRNP protein and detection of snRNP-snRNP interactions." Cell **51**(6): 1019-26.
- Lowe, T. M. and S. R. Eddy (1997). "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." Nucleic Acids Res **25**(5): 955-64.
- Lowe, T. M. and S. R. Eddy (1999). "A computational screen for methylation guide snoRNAs in yeast." Science **283**(5405): 1168-71.
- Lu, D., M. A. Searles, et al. (2003). "Crystal structure of a zinc-finger-RNA complex reveals two modes of molecular recognition." Nature **426**(6962): 96-100.
- Lucke, S., T. Klockner, et al. (1997). "Trans mRNA splicing in trypanosomes: cloning and analysis of a PRP8-homologous gene from *Trypanosoma brucei* provides evidence for a U5-analogous RNP." Embo J **16**(14): 4433-40.
- Luo, H. R., G. A. Moreau, et al. (1999). "The human Prp8 protein is a component of both U2- and U12-dependent spliceosomes." Rna **5**(7): 893-908.
- Luo, J., H. Zhou, et al. (2006). "Identification and evolutionary implication of four novel box H/ACA snoRNAs from *Giardia lamblia*." Chinese Science Bulletin **51**(20): 2451-2456.
- Lynch, M. (2006). "The origins of eukaryotic gene structure." Mol Biol Evol **23**(2): 450-68.
- Ma, J. B., K. Ye, et al. (2004). "Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain." Nature **429**(6989): 318-22.
- MacMillan, A. M., C. C. Query, et al. (1994). "Dynamic association of proteins with the pre-mRNA branch region." Genes Dev **8**(24): 3008-20.

- Macrae, I. J., K. Zhou, et al. (2007). "Structural determinants of RNA recognition and cleavage by Dicer." Nat Struct Mol Biol.
- Macrae, I. J., K. Zhou, et al. (2006). "Structural basis for double-stranded RNA processing by Dicer." Science **311**(5758): 195-8.
- Maden, B. E. (1990). "The numerous modified nucleotides in eukaryotic ribosomal RNA." Prog Nucleic Acid Res Mol Biol **39**: 241-303.
- Madhani, H. D. and C. Guthrie (1992). "A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome." Cell **71**(5): 803-17.
- Maizels, N. and A. M. Weiner (1994). "Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation." Proc Natl Acad Sci U S A **91**(15): 6729-34.
- Malhotra, P., P. V. Dasaradhi, et al. (2002). "Double-stranded RNA-mediated gene silencing of cysteine proteases (falcipain-1 and -2) of *Plasmodium falciparum*." Mol Microbiol **45**(5): 1245-54.
- Manak, J. R., S. Dike, et al. (2006). "Biological function of unannotated transcription during the early development of *Drosophila melanogaster*." Nat Genet **38**(10): 1151-8.
- Mandal, M. and R. R. Breaker (2004). "Adenine riboswitches and gene activation by disruption of a transcription terminator." Nat Struct Mol Biol **11**(1): 29-35.
- Marck, C., O. Lefebvre, et al. (1993). "The TFIIB-assembling subunit of yeast transcription factor TFIIC has both tetratricopeptide repeats and basic helix-loop-helix motifs." Proc Natl Acad Sci U S A **90**(9): 4027-31.
- Maroney, P. A., C. M. Romfo, et al. (2000). "Functional recognition of 5' splice site by U4/U6.U5 tri-snRNP defines a novel ATP-dependent step in early spliceosome assembly." Mol Cell **6**(2): 317-28.
- Marquez, S. M., J. K. Harris, et al. (2005). "Structural implications of novel diversity in eucaryal RNase P RNA." Rna **11**(5): 739-51.
- Martens, J. A., L. Laprade, et al. (2004). "Intergenic transcription is required to repress the *Saccharomyces cerevisiae* SER3 gene." Nature **429**(6991): 571-4.
- Martianov, I., A. Ramadass, et al. (2007). "Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript." Nature **445**(7128): 666-70.
- Martick, M. and W. G. Scott (2006). "Tertiary contacts distant from the active site prime a ribozyme for catalysis." Cell **126**(2): 309-20.
- Martienssen, R. A., M. Zaratiegui, et al. (2005). "RNA interference and heterochromatin in the fission yeast *Schizosaccharomyces pombe*." Trends Genet **21**(8): 450-6.
- Martin, G. and W. Keller (1998). "Tailing and 3'-end labeling of RNA with yeast poly(A) polymerase and various nucleotides." Rna **4**(2): 226-30.
- Martinez, J. and T. Tuschl (2004). "RISC is a 5' phosphomonoester-producing RNA endonuclease." Genes Dev **18**(9): 975-80.

- Masquida, B. and E. Westhof (2000). "On the wobble GoU and related pairs." Rna **6**(1): 9-15.
- Mathews, D. H., M. D. Disney, et al. (2004). "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure." Proc Natl Acad Sci U S A **101**(19): 7287-92.
- Mathews, D. H. and D. H. Turner (2002). "Dyalign: an algorithm for finding the secondary structure common to two RNA sequences." J Mol Biol **317**(2): 191-203.
- Mattick, J. S. (2004). "RNA regulation: a new genetics?" Nat Rev Genet **5**(4): 316-23.
- Mattick, J. S. and M. J. Gagen (2005). "Mathematics/computation. Accelerating networks." Science **307**(5711): 856-8.
- Mattick, J. S. and I. V. Makunin (2006). "Non-coding RNA." Hum Mol Genet **15 Spec No 1**: R17-29.
- Maytal-Kivity, V., N. Reis, et al. (2002). "MPN+, a putative catalytic motif found in a subset of MPN domain proteins from eukaryotes and prokaryotes, is critical for Rpn11 function." BMC Biochem **3**: 28.
- McArthur, A. G., H. G. Morrison, et al. (2000). "The Giardia genome project database." FEMS Microbiol Lett **189**(2): 271-3.
- McKay, D. B. (1996). "Structure and function of the hammerhead ribozyme: an unfinished story." Rna **2**(5): 395-403.
- McManus, C. J., M. L. Schwartz, et al. (2007). "A dynamic bulge in the U6 RNA internal stem loop functions in spliceosome assembly and activation." Rna **13**(12): 2252-2265.
- McPheeters, D. S. and P. Muhlenkamp (2003). "Spatial organization of protein-RNA interactions in the branch site-3' splice site region during pre-mRNA splicing in yeast." Mol Cell Biol **23**(12): 4174-86.
- Meister, G. and T. Tuschl (2004). "Mechanisms of gene silencing by double-stranded RNA." Nature **431**(7006): 343-9.
- Miller, J., A. D. McLachlan, et al. (1985). "Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus oocytes*." Embo J **4**(6): 1609-14.
- Miranda, R., L. M. Salgado, et al. (1996). "Identification and analysis of the u6 small nuclear RNA gene from *Entamoeba histolytica*." Gene **180**(1-2): 37-42.
- Mittal, V., M. A. Cleary, et al. (1996). "The Oct-1 POU-specific domain can stimulate small nuclear RNA gene transcription by stabilizing the basal transcription complex SNAPc." Mol Cell Biol **16**(5): 1955-65.
- Mockler, T. C., S. Chan, et al. (2005). "Applications of DNA tiling arrays for whole-genome analysis." Genomics **85**(1): 1-15.
- Mokdad, A., M. V. Krasovska, et al. (2006). "Structural and evolutionary classification of G/U wobble basepairs in the ribosome." Nucleic Acids Res **34**(5): 1326-41.

- Moreira, D., S. von der Heyden, et al. (2006). "Global eukaryote phylogeny: Combined small- and large-subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata." Mol Phylogenet Evol.
- Moroy, T. and F. Heyd (2007). "The impact of alternative splicing in vivo: mouse models show the way." Rna **13**(8): 1155-71.
- Morris, K. V. (2005). "siRNA-mediated transcriptional gene silencing: the potential mechanism and a possible role in the histone code." Cell Mol Life Sci **62**(24): 3057-66.
- Morris, K. V., S. W. Chan, et al. (2004). "Small interfering RNA-induced transcriptional gene silencing in human cells." Science **305**(5688): 1289-92.
- Morrison, H. G., A. G. McArthur, et al. (2007). "Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*." Science **317**(5846): 1921-6.
- Morrissey, J. P. and D. Tollervey (1995). "Birth of the snoRNPs: the evolution of RNase MRP and the eukaryotic pre-rRNA-processing system." Trends Biochem Sci **20**(2): 78-82.
- Murphy, S., J. B. Yoon, et al. (1992). "Oct-1 and Oct-2 potentiate functional interactions of a transcription factor with the proximal sequence element of small nuclear RNA genes." Mol Cell Biol **12**(7): 3247-61.
- Murray, J. B., A. A. Seyhan, et al. (1998). "The hammerhead, hairpin and VS ribozymes are catalytically proficient in monovalent cations alone." Chem Biol **5**(10): 587-95.
- Musier-Forsyth, K., N. Usman, et al. (1991). "Specificity for aminoacylation of an RNA helix: an unpaired, exocyclic amino group in the minor groove." Science **253**(5021): 784-6.
- Myslinski, E., J. C. Ame, et al. (2001). "An unusually compact external promoter for RNA polymerase III transcription of the human H1RNA gene." Nucleic Acids Res **29**(12): 2502-9.
- Myslinski, E., A. Krol, et al. (1998). "ZNF76 and ZNF143 are two human homologs of the transcriptional activator Staf." J Biol Chem **273**(34): 21998-2006.
- Myslinski, E., C. Schuster, et al. (1993). "Promoter strength and structure dictate module composition in RNA polymerase III transcriptional activator elements." J Mol Biol **234**(2): 311-8.
- Nagasaki, H., M. Arita, et al. (2005). "Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes." Gene **364**: 53-62.
- Nagel, R. and M. Ares, Jr. (2000). "Substrate recognition by a eukaryotic RNase III: the double-stranded RNA-binding domain of Rnt1p selectively binds RNA containing a 5'-AGNN-3' tetraloop." Rna **6**(8): 1142-56.

- Narcisi, E. M., C. V. Glover, et al. (1998). "Fibrillarin, a conserved pre-ribosomal RNA processing protein of Giardia." J Eukaryot Microbiol **45**(1): 105-11.
- Nash, T. E., A. Aggarwal, et al. (1988). "Antigenic variation in Giardia lamblia." J Immunol **141**(2): 636-41.
- Nash, T. E., H. T. Lujan, et al. (2001). "Variant-specific surface protein switching in Giardia lamblia." Infect Immun **69**(3): 1922-3.
- Neuwald, A. F., J. S. Liu, et al. (1995). "Gibbs motif sampling: detection of bacterial outer membrane protein repeats." Protein Sci **4**(8): 1618-32.
- Newby, M. I. and N. L. Greenbaum (2001). "A conserved pseudouridine modification in eukaryotic U2 snRNA induces a change in branch-site architecture." Rna **7**(6): 833-45.
- Ngo, H., C. Tschudi, et al. (1998). "Double-stranded RNA induces mRNA degradation in Trypanosoma brucei." Proc Natl Acad Sci U S A **95**(25): 14687-92.
- Ni, J., A. L. Tien, et al. (1997). "Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA." Cell **89**(4): 565-73.
- Nilsen, T. W. (2003). "The spliceosome: the most complex macromolecular machine in the cell?" Bioessays **25**(12): 1147-9.
- Nishida, K., S. Suzuki, et al. (1998). "Group I introns found in Chlorella viruses: biological implications." Virology **242**(2): 319-26.
- Nishihara, H., A. F. Smit, et al. (2006). "Functional noncoding sequences derived from SINEs in the mammalian genome." Genome Res **16**(7): 864-74.
- Nissen, P., J. Hansen, et al. (2000). "The structural basis of ribosome activity in peptide bond synthesis." Science **289**(5481): 920-30.
- Niu, X. H., T. Hartshorne, et al. (1994). "Characterization of putative small nuclear RNAs from Giardia lamblia." Mol Biochem Parasitol **66**(1): 49-57.
- Nixon, J. E., A. Wang, et al. (2002). "A spliceosomal intron in Giardia lamblia." Proc Natl Acad Sci U S A **99**(6): 3701-5.
- Noeske, J., C. Richter, et al. (2005). "An intermolecular base triple as the basis of ligand specificity and affinity in the guanine- and adenine-sensing riboswitch RNAs." Proc Natl Acad Sci U S A **102**(5): 1372-7.
- Noller, H. F. and J. B. Chaires (1972). "Functional modification of 16S ribosomal RNA by kethoxal." Proc Natl Acad Sci U S A **69**(11): 3115-8.
- Nottrott, S., H. Urlaub, et al. (2002). "Hierarchical, clustered protein interactions with U4/U6 snRNA: a biochemical role for U4/U6 proteins." Embo J **21**(20): 5527-38.
- Oh, H., Y. Park, et al. (2003). "Local spreading of MSL complexes from roX genes on the Drosophila X chromosome." Genes Dev **17**(11): 1334-9.
- Ohno, S., U. Wolf, et al. (1968). "Evolution from fish to mammals by gene duplication." Hereditas **59**(1): 169-87.

- Omer, A. D., S. Ziesche, et al. (2002). "In vitro reconstitution and activity of a C/D box methylation guide ribonucleoprotein complex." Proc Natl Acad Sci U S A **99**(8): 5289-94.
- Otto, S. P. and J. Whitton (2000). "Polyploid incidence and evolution." Annu Rev Genet **34**: 401-437.
- Ouyang, C., M. J. Martinez, et al. (2000). "TATA-Binding protein-TATA interaction is a key determinant of differential transcription of silkworm constitutive and silk gland-specific tRNA(Ala) genes." Mol Cell Biol **20**(4): 1329-43.
- Palfi, Z., B. Schimanski, et al. (2005). "U1 small nuclear RNP from *Trypanosoma brucei*: a minimal U1 snRNA with unusual protein components." Nucleic Acids Res **33**(8): 2493-503.
- Pang, K. C., S. Stephen, et al. (2005). "RNADB--a comprehensive mammalian noncoding RNA database." Nucleic Acids Res **33**(Database issue): D125-30.
- Pannucci, J. A., E. S. Haas, et al. (1999). "RNase P RNAs from some Archaea are catalytically active." Proc Natl Acad Sci U S A **96**(14): 7803-8.
- Parker, J. S., S. M. Roe, et al. (2004). "Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity." Embo J **23**(24): 4727-37.
- Pasquinelli, A. E., S. Hunter, et al. (2005). "MicroRNAs: a developing story." Curr Opin Genet Dev **15**(2): 200-5.
- Peaston, A. E., A. V. Evsikov, et al. (2004). "Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos." Dev Cell **7**(4): 597-606.
- Peattie, D. A. (1979). "Direct chemical method for sequencing RNA." Proc Natl Acad Sci U S A **76**(4): 1760-4.
- Penedo, J. C., T. J. Wilson, et al. (2004). "Folding of the natural hammerhead ribozyme is enhanced by interaction of auxiliary elements." Rna **10**(5): 880-8.
- Peng, W. T., M. D. Robinson, et al. (2003). "A panoramic view of yeast noncoding RNA processing." Cell **113**(7): 919-33.
- Petryniak, B., L. M. Staudt, et al. (1990). "Characterization of chicken octamer-binding proteins demonstrates that POU domain-containing homeobox transcription factors have been highly conserved during vertebrate evolution." Proc Natl Acad Sci U S A **87**(3): 1099-103.
- Pham, J. W., J. L. Pellino, et al. (2004). "A Dicer-2-dependent 80s complex cleaves targeted mRNAs during RNAi in *Drosophila*." Cell **117**(1): 83-94.
- Phizicky, E. M. and C. L. Greer (1993). "Pre-tRNA splicing: variation on a theme or exception to the rule?" Trends Biochem Sci **18**(1): 31-4.
- Pichon, C. and B. Felden (2005). "Small RNA genes expressed from *Staphylococcus aureus* genomic and pathogenicity islands with specific

- expression among pathogenic strains." Proc Natl Acad Sci U S A **102**(40): 14249-54.
- Pley, H. W., K. M. Flaherty, et al. (1994). "Three-dimensional structure of a hammerhead ribozyme." Nature **372**(6501): 68-74.
- Pley, H. W., D. S. Lindes, et al. (1994). "Crystals of a hammerhead ribozyme." J Biol Chem **269**(6): 4692.
- Pollard, K. S., S. R. Salama, et al. (2006). "An RNA gene expressed during cortical development evolved rapidly in humans." Nature **443**(7108): 167-72.
- Poole, A. M., D. C. Jeffares, et al. (1998). "The path from the RNA world." J Mol Evol **46**(1): 1-17.
- Poole, A. M. and D. Penny (2007). "Evaluating hypotheses for the origin of eukaryotes." Bioessays **29**(1): 74-84.
- Pozzoli, U., G. Menozzi, et al. (2007). "Intron size in mammals: complexity comes to terms with economy." Trends Genet **23**(1): 20-4.
- Prasanth, K. V., S. G. Prasanth, et al. (2005). "Regulating gene expression through RNA nuclear retention." Cell **123**(2): 249-63.
- Prilusky, J., C. E. Felder, et al. (2005). "FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded." Bioinformatics **21**(16): 3435-8.
- Pyle, A. M. (1993). "Ribozymes: a distinct class of metalloenzymes." Science **261**(5122): 709-14.
- Qu, L. H., A. Henras, et al. (1999). "Seven novel methylation guide small nucleolar RNAs are processed from a common polycistronic transcript by Rat1p and RNase III in yeast." Mol Cell Biol **19**(2): 1144-58.
- Rameau, G., K. Puglia, et al. (1994). "A mutation in the second largest subunit of TFIIC increases a rate-limiting step in transcription by RNA polymerase III." Mol Cell Biol **14**(1): 822-30.
- Randau, L., R. Munch, et al. (2005). "Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5'- and 3'-halves." Nature **433**(7025): 537-41.
- Ranz, J. M. and C. A. Machado (2006). "Uncovering evolutionary patterns of gene expression using microarrays." Trends Ecol Evol **21**(1): 29-37.
- Ravasi, T., H. Suzuki, et al. (2006). "Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome." Genome Res **16**(1): 11-9.
- Reinhold-Hurek, B. and D. A. Shub (1992). "Self-splicing introns in tRNA genes of widely divergent bacteria." Nature **357**(6374): 173-6.
- Rodionov, D. A., A. G. Vitreschak, et al. (2002). "Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms." J Biol Chem **277**(50): 48949-59.
- Roger, A. J., S. G. Svard, et al. (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): 229-34.
- Rosenfeld, R. and H. Margalit (1993). "Zinc fingers: conserved properties that can distinguish between spurious and actual DNA-binding motifs." J Biomol Struct Dyn **11**(3): 557-70.
- Rousset, F., M. Pelandakis, et al. (1991). "Evolution of compensatory substitutions through G.U intermediate state in Drosophila rRNA." Proc Natl Acad Sci U S A **88**(22): 10032-6.
- Roy, S. W. (2003). "Recent evidence for the exon theory of genes." Genetica **118**(2-3): 251-66.
- Roy, S. W. and W. Gilbert (2006). "The evolution of spliceosomal introns: patterns, puzzles and progress." Nat Rev Genet **7**(3): 211-21.
- Roychowdhury-Saha, M. and D. H. Burke (2007). "Distinct reaction pathway promoted by non-divalent-metal cations in a tertiary stabilized hammerhead ribozyme." Rna **13**(6): 841-8.
- Royo, H., E. Basyuk, et al. (2007). "Bsr, a nuclear-retained RNA with monoallelic expression." Mol Biol Cell **18**(8): 2817-27.
- Russell, A. G., T. E. Shutt, et al. (2005). "An ancient spliceosomal intron in the ribosomal protein L7a gene (Rpl7a) of Giardia lamblia." BMC Evol Biol **5**: 45.
- Sakonju, S., D. F. Bogenhagen, et al. (1980). "A control region in the center of the 5S RNA gene directs specific initiation of transcription: I. The 5' border of the region." Cell **19**(1): 13-25.
- Salehi-Ashtiani, K. and J. W. Szostak (2001). "In vitro evolution suggests multiple origins for the hammerhead ribozyme." Nature **414**(6859): 82-4.
- Sandegren, L. and B. M. Sjoberg (2004). "Distribution, sequence homology, and homing of group I introns among T-even-like bacteriophages: evidence for recent transfer of old introns." J Biol Chem **279**(21): 22218-27.
- Sanges, R., E. Kalmar, et al. (2006). "Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage." Genome Biol **7**(7): R56.
- Sankoff, D. (1985). "Simultaneous solution of the RNA folding: alignment and protosequence problems." SIAM. J. Appl. Math **45**: 810-825.
- Sashital, D. G., A. M. Allmann, et al. (2003). "Structural basis for a lethal mutation in U6 RNA." Biochemistry **42**(6): 1470-7.
- Sashital, D. G., G. Cornilescu, et al. (2004). "U2-U6 RNA folding reveals a group II intron-like domain and a four-helix junction." Nat Struct Mol Biol **11**(12): 1237-42.
- Saville, B. J. and R. A. Collins (1991). "RNA-mediated ligation of self-cleavage products of a Neurospora mitochondrial plasmid transcript." Proc Natl Acad Sci U S A **88**(19): 8826-30.

- Scannell, D. R., K. P. Byrne, et al. (2006). "Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts." Nature **440**(7082): 341-5.
- Schaub, M., A. Krol, et al. (1999). "Flexible zinc finger requirement for binding of the transcriptional activator staf to U6 small nuclear RNA and tRNA(Sec) promoters." J Biol Chem **274**(34): 24241-9.
- Schaub, M., E. Myslinski, et al. (1997). "Staf, a promiscuous activator for enhanced transcription by RNA polymerases II and III." Embo J **16**(1): 173-81.
- Schmitt, M. E., J. L. Bennett, et al. (1993). "Secondary structure of RNase MRP RNA as predicted by phylogenetic comparison." Faseb J **7**(1): 208-13.
- Schwarz, D. S., G. Hutvagner, et al. (2003). "Asymmetry in the assembly of the RNAi enzyme complex." Cell **115**(2): 199-208.
- Scott, W. G., J. T. Finch, et al. (1995). "The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage." Cell **81**(7): 991-1002.
- Seetharaman, M., N. V. Eldho, et al. (2006). "Structure of a self-splicing group II intron catalytic effector domain 5: parallels with spliceosomal U6 RNA." Rna **12**(2): 235-47.
- Selinger, D. W., K. J. Cheung, et al. (2000). "RNA expression analysis using a 30 base pair resolution Escherichia coli genome array." Nat Biotechnol **18**(12): 1262-8.
- Selvamurugan, N., M. K. Nag, et al. (1995). "Intron-encoded small nucleolar RNAs: new RNA sequence variants and genomic loci." Biochim Biophys Acta **1260**(2): 230-4.
- Sen, C. K. and S. Roy (2007). "miRNA: licensed to kill the messenger." DNA Cell Biol **26**(4): 193-4.
- Senior, B. W. and I. B. Holland (1971). "Effect of colicin E3 upon the 30S ribosomal subunit of Escherichia coli." Proc Natl Acad Sci U S A **68**(5): 959-63.
- Seto, A. G., R. E. Kingston, et al. (2007). "The coming of age for Piwi proteins." Mol Cell **26**(5): 603-9.
- Shan, S., A. V. Kravchuk, et al. (2001). "Defining the catalytic metal ion interactions in the Tetrahymena ribozyme reaction." Biochemistry **40**(17): 5161-71.
- Shan, S., A. Yoshida, et al. (1999). "Three metal ions at the active site of the Tetrahymena group I ribozyme." Proc Natl Acad Sci U S A **96**(22): 12299-304.
- Shi, H., A. Djikeng, et al. (2004). "Argonaute protein in the early divergent eukaryote Trypanosoma brucei: control of small interfering RNA accumulation and retroposon transcript abundance." Mol Cell Biol **24**(1): 420-7.

- Shi, H., C. Tschudi, et al. (2006). "An unusual Dicer-like1 protein fuels the RNA interference pathway in *Trypanosoma brucei*." Rna **12**(12): 2063-72.
- Shtatland, T., S. C. Gill, et al. (2000). "Interactions of *Escherichia coli* RNA with bacteriophage MS2 coat protein: genomic SELEX." Nucleic Acids Res **28**(21): E93.
- Shukla, G. C. and R. A. Padgett (2002). "A catalytically active group II intron domain 5 can function in the U12-dependent spliceosome." Mol Cell **9**(5): 1145-50.
- Simpson, A. G. (2003). "Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota)." Int J Syst Evol Microbiol **53**(Pt 6): 1759-77.
- Simpson, R. J., E. D. Cram, et al. (2003). "CCHX zinc finger derivatives retain the ability to bind Zn(II) and mediate protein-DNA interactions." J Biol Chem **278**(30): 28011-8.
- Singer, B. S., T. Shtatland, et al. (1997). "Libraries for genomic SELEX." Nucleic Acids Res **25**(4): 781-6.
- Slamovits, C. H. and P. J. Keeling (2006). "A high density of ancient spliceosomal introns in oxymonad excavates." BMC Evol Biol **6**: 34.
- Smardon, A., J. M. Spoerke, et al. (2000). "EGO-1 is related to RNA-directed RNA polymerase and functions in germ-line development and RNA interference in *C. elegans*." Curr Biol **10**(4): 169-78.
- Sogin, M. L. (1991). "Early evolution and the origin of eukaryotes." Curr Opin Genet Dev **1**(4): 457-63.
- Sogin, M. L., J. H. Gunderson, et al. (1989). "Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia*." Science **243**(4887): 75-7.
- Song, J. J., S. K. Smith, et al. (2004). "Crystal structure of Argonaute and its implications for RISC slicer activity." Science **305**(5689): 1434-7.
- Sorek, R. (2007). "The birth of new exons: Mechanisms and evolutionary consequences." Rna.
- Stahley, M. R., P. L. Adams, et al. (2007). "Structural metals in the group I intron: a ribozyme with a multiple metal ion core." J Mol Biol **372**(1): 89-102.
- Staley, J. P. and C. Guthrie (1998). "Mechanical devices of the spliceosome: motors, clocks, springs, and things." Cell **92**(3): 315-26.
- Stefan, L. R., R. Zhang, et al. (2006). "MeRNA: a database of metal ion binding sites in RNA structures." Nucleic Acids Res **34**(Database issue): D131-4.
- Stein, L. D., Z. Bao, et al. (2003). "The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics." PLoS Biol **1**(2): E45.
- Stevens, S. W. and J. Abelson (1999). "Purification of the yeast U4/U6.U5 small nuclear ribonucleoprotein particle and identification of its proteins." Proc Natl Acad Sci U S A **96**(13): 7226-31.

- Stoughton, R. B. (2005). "Applications of DNA microarrays in biology." Annu Rev Biochem **74**: 53-82.
- Sturm, R. A. and W. Herr (1988). "The POU domain is a bipartite DNA-binding structure." Nature **336**(6199): 601-4.
- Sudarsan, N., J. E. Barrick, et al. (2003). "Metabolite-binding RNA domains are present in the genes of eukaryotes." Rna **9**(6): 644-7.
- Sun, J. S. and J. L. Manley (1995). "A novel U2-U6 snRNA structure is necessary for mammalian mRNA splicing." Genes Dev **9**(7): 843-54.
- Swift, H. (1950). "The constancy of desoxyribose nucleic acid in plant nuclei." Proc Natl Acad Sci U S A **36**(11): 643-54.
- Szewczak, A. A., A. B. Kosek, et al. (2002). "Identification of an active site ligand for a group I ribozyme catalytic metal ion." Biochemistry **41**(8): 2516-25.
- Tachezy, J., L. B. Sanchez, et al. (2001). "Mitochondrial type iron-sulfur cluster assembly in the amitochondriate eukaryotes *Trichomonas vaginalis* and *Giardia intestinalis*, as indicated by the phylogeny of IscS." Mol Biol Evol **18**(10): 1919-28.
- Taft, R. J., M. Pheasant, et al. (2007). "The relationship between non-protein-coding DNA and eukaryotic complexity." Bioessays **29**(3): 288-99.
- Tahbaz, N., F. A. Kolb, et al. (2004). "Characterization of the interactions between mammalian PAZ PIWI domain proteins and Dicer." EMBO Rep **5**(2): 189-94.
- Tanaka, T. and Y. Kikuchi (2000). "Origin of the cloverleaf shape of transfer RNA -- the double-hairpin model: implication for the role of tRNA intron and the long extra loop." Viva Origino **29**: 134-142.
- Tang, T. H., N. Polacek, et al. (2005). "Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*." Mol Microbiol **55**(2): 469-81.
- Tani, T. and Y. Ohshima (1991). "mRNA-type introns in U6 small nuclear RNA genes: implications for the catalysis in pre-mRNA splicing." Genes Dev **5**(6): 1022-31.
- Tanzer, A. and P. F. Stadler (2004). "Molecular evolution of a microRNA cluster." J Mol Biol **339**(2): 327-35.
- Teigelkamp, S., A. J. Newman, et al. (1995). "Extensive interactions of PRP8 protein with the 5' and 3' splice sites during splicing suggest a role in stabilization of exon alignment by U5 snRNA." Embo J **14**(11): 2602-12.
- Thomas, B. C., B. Pedersen, et al. (2006). "Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes." Genome Res **16**(7): 934-46.

- Thompson, W., E. C. Rouchka, et al. (2003). "Gibbs Recursive Sampler: finding transcription factor binding sites." Nucleic Acids Res **31**(13): 3580-5.
- Thore, S., M. Leibundgut, et al. (2006). "Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand." Science **312**(5777): 1208-11.
- Tinsley, R. A. and N. G. Walter (2007). "Long-range impact of peripheral joining elements on structure and function of the hepatitis delta virus ribozyme." Biol Chem **388**(7): 705-15.
- Tjaden, B., R. M. Saxena, et al. (2002). "Transcriptome analysis of Escherichia coli using high-density oligonucleotide probe arrays." Nucleic Acids Res **30**(17): 3732-8.
- Tomari, Y., T. Du, et al. (2004). "RISC assembly defects in the Drosophila RNAi mutant armitage." Cell **116**(6): 831-41.
- Torarinsson, E., J. H. Havgaard, et al. (2007). "Multiple structural alignment and clustering of RNA sequences." Bioinformatics **23**(8): 926-32.
- Torarinsson, E., M. Sawera, et al. (2006). "Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure." Genome Res **16**(7): 885-9.
- Toro, N. (2003). "Bacteria and Archaea Group II introns: additional mobile genetic elements in the environment." Environ Microbiol **5**(3): 143-51.
- Tovar, J., A. Fischer, et al. (1999). "The mitosome, a novel organelle related to mitochondria in the amitochondrial parasite Entamoeba histolytica." Mol Microbiol **32**(5): 1013-21.
- Tovar, J., G. Leon-Avila, et al. (2003). "Mitochondrial remnant organelles of Giardia function in iron-sulphur protein maturation." Nature **426**(6963): 172-6.
- Trivedi, A., L. S. Young, et al. (1999). "A TATA element is required for tRNA promoter activity and confers TATA-binding protein responsiveness in Drosophila Schneider-2 cells." J Biol Chem **274**(16): 11369-75.
- Tschudi, C., A. Djikeng, et al. (2003). "In vivo analysis of the RNA interference mechanism in Trypanosoma brucei." Methods **30**(4): 304-12.
- Turner, I. A., C. M. Norman, et al. (2004). "Roles of the U5 snRNP in spliceosome dynamics and catalysis." Biochem Soc Trans **32**(Pt 6): 928-31.
- Turner, I. A., C. M. Norman, et al. (2006). "Dissection of Prp8 protein defines multiple interactions with crucial RNA sequences in the catalytic core of the spliceosome." Rna **12**(3): 375-86.
- Ullu, E., H. D. Lujan, et al. (2005). "Small sense and antisense RNAs derived from a telomeric retroposon family in Giardia intestinalis." Eukaryot Cell **4**(6): 1155-7.
- Ullu, E., C. Tschudi, et al. (2004). "RNA interference in protozoan parasites." Cell Microbiol **6**(6): 509-19.

- Uversky, V. N. (2002). "Natively unfolded proteins: a point where biology waits for physics." Protein Sci **11**(4): 739-56.
- Valadkhan, S. (2005). "snRNAs as the catalysts of pre-mRNA splicing." Curr Opin Chem Biol **9**(6): 603-8.
- Valadkhan, S. and J. L. Manley (2001). "Splicing-related catalysis by protein-free snRNAs." Nature **413**(6857): 701-7.
- Valadkhan, S., A. Mohammadi, et al. (2007). "Protein-free spliceosomal snRNAs catalyze a reaction that resembles the first step of splicing." Rna **13**(12): 2300-2311.
- van der Giezen, M. and J. Tovar (2005). "Degenerate mitochondria." EMBO Rep **6**(6): 525-30.
- Vanacova, S., D. R. Liston, et al. (2003). "Molecular biology of the amitochondriate parasites, *Giardia intestinalis*, *Entamoeba histolytica* and *Trichomonas vaginalis*." Int J Parasitol **33**(3): 235-55.
- Vanacova, S., W. Yan, et al. (2005). "Spliceosomal introns in the deep-branching eukaryote *Trichomonas vaginalis*." Proc Natl Acad Sci U S A **102**(12): 4430-5.
- Vankan, P., D. Edoh, et al. (1988). "Structure and expression of the U5 snRNA gene of *Arabidopsis thaliana*. Conserved upstream sequence elements in plant U-RNA genes." Nucleic Acids Res **16**(22): 10425-40.
- Vasudevan, S., Y. Tong, et al. (2007). "Switching from repression to activation: microRNAs can up-regulate translation." Science **318**(5858): 1931-4.
- Vermeulen, A., L. Behlen, et al. (2005). "The contributions of dsRNA structure to Dicer specificity and efficiency." Rna **11**(5): 674-82.
- Verrijzer, C. P. and P. C. Van der Vliet (1993). "POU domain transcription factors." Biochim Biophys Acta **1173**(1): 1-21.
- Vervoort, A. (1980). "Tetraploidy in *Protopterus* (Dipnoi)." Cellular and Molecular Life Sciences **36**: 294-296.
- Vidal, V. P., L. Verdone, et al. (1999). "Characterization of U6 snRNA-protein interactions." Rna **5**(11): 1470-81.
- Vincenti, S., V. De Chiara, et al. (2007). "The position of yeast snoRNA-coding regions within host introns is essential for their biosynthesis and for efficient splicing of the host pre-mRNA." Rna **13**(1): 138-50.
- Vinogradov, A. E. (2006). "'Genome design' model: evidence from conserved intronic sequence in human-mouse comparison." Genome Res **16**(3): 347-54.
- Vitali, P., E. Basyuk, et al. (2005). "ADAR2-mediated editing of RNA substrates in the nucleolus is inhibited by C/D small nucleolar RNAs." J Cell Biol **169**(5): 745-53.
- Vitali, P., H. Royo, et al. (2003). "Identification of 13 novel human modification guide RNAs." Nucleic Acids Res **31**(22): 6543-51.

- Vitreschak, A. G., D. A. Rodionov, et al. (2004). "Riboswitches: the oldest mechanism for the regulation of gene expression?" Trends Genet **20**(1): 44-50.
- Vogel, C. and C. Chothia (2006). "Protein family expansions and biological complexity." PLoS Comput Biol **2**(5): e48.
- Vogt, M., S. Lahiri, et al. (2006). "Coordination environment of a site-bound metal ion in the hammerhead ribozyme determined by ¹⁵N and ²H ESEEM spectroscopy." J Am Chem Soc **128**(51): 16764-70.
- Volpe, T. A., C. Kidner, et al. (2002). "Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi." Science **297**(5588): 1833-7.
- Wallis, J. W., J. Aerts, et al. (2004). "A physical map of the chicken genome." Nature **432**(7018): 761-4.
- Wang, H., N. H. Chua, et al. (2006). "Prediction of trans-antisense transcripts in *Arabidopsis thaliana*." Genome Biol **7**(10): R92.
- Wang, W. and E. F. Kirkness (2005). "Short interspersed elements (SINEs) are a major source of canine genomic diversity." Genome Res **15**(12): 1798-808.
- Wang, W., H. Zheng, et al. (2005). "Origin and evolution of new exons in rodents." Genome Res **15**(9): 1258-64.
- Waterhouse, P. M., M. B. Wang, et al. (2001). "Gene silencing as an adaptive defence against viruses." Nature **411**(6839): 834-42.
- Waterston, R. H., K. Lindblad-Toh, et al. (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-62.
- Watkins, N. J., V. Segault, et al. (2000). "A common core RNP structure shared between the small nucleolar box C/D RNPs and the spliceosomal U4 snRNP." Cell **103**(3): 457-66.
- Williams, B. A., R. P. Hirt, et al. (2002). "A mitochondrial remnant in the microsporidian *Trachipleistophora hominis*." Nature **418**(6900): 865-9.
- Williamson, B. (1977). "DNA insertions and gene structure." Nature **270**: 295-297.
- Winkler, W., A. Nahvi, et al. (2002). "Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression." Nature **419**(6910): 952-6.
- Winkler, W. C. and R. R. Breaker (2005). "Regulation of bacterial gene expression by riboswitches." Annu Rev Microbiol **59**: 487-517.
- Woese, C. R., O. Kandler, et al. (1990). "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya." Proc Natl Acad Sci U S A **87**(12): 4576-9.
- Woodhams, M. D., P. F. Stadler, et al. (2007). "RNase MRP and the RNA processing cascade in the eukaryotic ancestor." BMC Evol Biol **7** **Suppl 1**: S13.

- Wright, P. E. and H. J. Dyson (1999). "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm." J Mol Biol **293**(2): 321-31.
- Wu, G., A. G. McArthur, et al. (2000). "Core histones of the amitochondriate protist, *Giardia lamblia*." Mol Biol Evol **17**(8): 1156-63.
- Wyatt, J. R., E. J. Sontheimer, et al. (1992). "Site-specific cross-linking of mammalian U5 snRNP to the 5' splice site before the first step of pre-mRNA splicing." Genes Dev **6**(12B): 2542-53.
- Xie, X., M. Kamal, et al. (2006). "A family of conserved noncoding elements derived from an ancient transposable element." Proc Natl Acad Sci U S A **103**(31): 11659-64.
- Xin, D. D., J. F. Wen, et al. (2005). "Identification of a *Giardia krr1* homolog gene and the secondarily anucleolate condition of *Giardia lamblia*." Mol Biol Evol **22**(3): 391-4.
- Xing, Y. and C. Lee (2006). "Alternative splicing and RNA selection pressure--evolutionary consequences for eukaryotic genomes." Nat Rev Genet **7**(7): 499-509.
- Yang, C. Y., H. Zhou, et al. (2005). "Identification of 20 snoRNA-like RNAs from the primitive eukaryote, *Giardia lamblia*." Biochem Biophys Res Commun **328**(4): 1224-31.
- Yarus, M. and B. G. Barrell (1971). "The sequence of nucleotides in tRNA Ile from *E. coli* B." Biochem Biophys Res Commun **43**(4): 729-34.
- Yazgan, O. and J. E. Krebs (2007). "Noncoding but nonexpendable: transcriptional regulation by large noncoding RNA in eukaryotes." Biochem Cell Biol **85**(4): 484-96.
- Ye, J., S. McGinnis, et al. (2006). "BLAST: improvements for better sequence analysis." Nucleic Acids Res **34**(Web Server issue): W6-9.
- Yean, S. L. and R. J. Lin (1991). "U4 small nuclear RNA dissociates from a yeast spliceosome and does not participate in the subsequent splicing reaction." Mol Cell Biol **11**(11): 5571-7.
- Yean, S. L., G. Wuenschell, et al. (2000). "Metal-ion coordination by U6 small nuclear RNA contributes to catalysis in the spliceosome." Nature **408**(6814): 881-4.
- Yee, J., M. R. Mowatt, et al. (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): 11432-9.
- Young, K. J., F. Gill, et al. (1997). "Metal ions play a passive role in the hairpin ribozyme catalysed reaction." Nucleic Acids Res **25**(19): 3760-6.
- Yuan, F., L. Griffin, et al. (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): 2833-45.

- Zhang, A., K. M. Wassarman, et al. (2003). "Global analysis of small RNA and mRNA targets of Hfq." Mol Microbiol **50**(4): 1111-24.
- Zhang, B., B. Kraemer, et al. (1999). "Yeast three-hybrid system to detect and analyze interactions between RNA and protein." Methods Enzymol **306**: 93-113.
- Zhang, H., F. A. Kolb, et al. (2004). "Single processing center models for human Dicer and bacterial RNase III." Cell **118**(1): 57-68.